Bioinformatics, 31(11), 2015, 1771–1779
doi: 10.1093/bioinformatics/btv036
Advance Access Publication Date: 23 January 2015
Original paper

Systems biology

Context-specific metabolic network reconstruction of a naphthalene-degrading bacterial community guided by metaproteomic data

Luis Tobalina[1], Rafael Bargiela[2], Jon Pey[1], Florian-Alexander Herbst[3,†], Iván Lores[4], David Rojo[5], Coral Barbas[5], Ana I. Peláez[4], Jesús Sánchez[4], Martin von Bergen[3,6,7], Jana Seifert[3,‡], Manuel Ferrer[2,*] and Francisco J. Planes[1,*]

[1]CEIT and Tecnun (University of Navarra), San Sebastián, Spain, [2]CSIC, Institute of Catalysis, Madrid, Spain, [3]Helmholtz Centre for Environmental Research, Department of Proteomics, Leipzig, Germany, [4]Área de Microbiología, IUBA, Universidad de Oviedo, Oviedo, Spain, [5]Centro de Metabolómica y Bioanálisis (CEMBIO), Facultad de Farmacia, Universidad CEU San Pablo, Campus Monteprincipe, Boadilla del Monte, Madrid, Spain, [6]Department of Metabolomics, UFZ-Helmholtz-Zentrum für Umweltforschung GmbH, Leipzig, Germany and [7]Department of Chemistry and Bioscience, Aalborg University, Aalborg, Denmark

*To whom correspondence should be addressed.

[†]Present address: Center for Microbial Communities, Department of Chemistry and Bioscience, Aalborg University, Aalborg, Denmark.

[‡]Present address: Institute of Animal Science, Universität Hohenheim, Stuttgart, Germany.

Associate Editor: Igor Jurisica

Received on July 14, 2014; revised on January 9, 2015; accepted on January 18, 2015

Abstract

Motivation: With the advent of meta-'omics' data, the use of metabolic networks for the functional analysis of microbial communities became possible. However, while network-based methods are widely developed for single organisms, their application to bacterial communities is currently limited.

Results: Herein, we provide a novel, context-specific reconstruction procedure based on metaproteomic and taxonomic data. Without previous knowledge of a high-quality, genome-scale metabolic networks for each different member in a bacterial community, we propose a meta-network approach, where the expression levels and taxonomic assignments of proteins are used as the most relevant clues for inferring an active set of reactions. Our approach was applied to draft the context-specific metabolic networks of two different naphthalene-enriched communities derived from an anthropogenically influenced, polyaromatic hydrocarbon contaminated soil, with (CN2) or without (CN1) bio-stimulation. We were able to capture the overall functional differences between the two conditions at the metabolic level and predict an important activity for the fluorobenzoate degradation pathway in CN1 and for geraniol metabolism in CN2. Experimental validation was conducted, and good agreement with our computational predictions was observed. We also hypothesize different pathway organizations at the organismal level, which is relevant to disentangle the role of each member in the communities. The approach presented here can be easily transferred to the analysis of genomic, transcriptomic and metabolomic data.

Contact: fplanes@ceit.es or mferrer@icp.csic.es
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Microbes are shaping the world and, by forming communities, are causal of geochemical cycles (Mascarelli, 2009), human health (Kinross *et al.*, 2011), and biotechnological processes (Beloqui *et al.*, 2008). Thus, it is not surprising to find increasing interest in studying how these consortia lead to function (Carter *et al.*, 2012).

The analysis of microbial communities begins by assessing the structure of the population, which is currently often achieved using metagenomic data (Röling *et al.*, 2010). The next step consists of characterizing the metabolic capacity of the microbial community, but this has proven to be a considerable challenge, even when using metatranscriptomic data (Moran *et al.*, 2013). This need has led to the development of metaproteomics, by which at least the abundance of metabolically active molecules can be detected (Seifert *et al.*, 2013). In parallel, methods for analysing these data have arisen and evolved.

From the outset, computational methods have been essential for capitalizing on data to obtain clear and novel insights (Guazzaroni and Ferrer, 2011). Traditional approaches described in the literature typically map data for genes, proteins or metabolites onto well-known pathways or Gene Ontology (GO) terms (Yamada *et al.*, 2011). This enables identifying molecular functions of identified proteins in light of metabolic pathways. However, the high connectivity among biological pathways has shifted the focus to networks (Letunic *et al.*, 2008; Palsson, 2009), which allows us to capture more global properties (McCloskey *et al.*, 2013). Molecular networks integrate different pathways and constitute a more general framework for interpreting 'omics' data (Bordbar and Palsson, 2012).

On the single-species level, different computational methods have been developed to analyse 'omics' data using genome-scale metabolic networks. In particular, a number of approaches have been specifically designed to incorporate gene and protein expression data (Jerby *et al.*, 2010; Rezola *et al.*, 2013) as well as metabolite and flux levels (Pey *et al.*, 2013; Zamboni *et al.*, 2008) to characterize context-specific metabolic phenotypes. Based on these methods, novel insights into the metabolic organization of organisms have been obtained (Vitkin and Shlomi, 2012) and more practical questions have arisen, such as how to improve the efficiency of industrial biosynthetic processes (Agren *et al.*, 2013; Boyle and Silver, 2012; Curran and Alper, 2012; Poblete-Castro *et al.*, 2013).

These approaches start from genome-scale metabolic networks, which are reconstructed from the annotated genome of an organism (Bachmann *et al.*, 2013; Zomorrodi *et al.*, 2012) and a reference metabolic database as input information. Thiele and Palsson (2010) described the steps for accurately building a metabolic network, which can be time-consuming, easily taking from 6 months to 2 years. With the aim of easing this process, the Model SEED provides an integrative and automatic approach that substantially speeds up the time required to obtain a first network draft (Henry *et al.*, 2010).

For microbial communities, the reconstruction of metabolic networks is more complicated and faces new challenges. Ideally, each organism can be represented by its own metabolic network and its input/output metabolites define its possible interaction with other members of the consortia. Should this information be available, recently developed constraint-based modelling approaches could be applied. In this situation, methods mentioned above to incorporate 'omics' data for single organisms could be extended to deal with bacterial communities.

However, in complex bacterial communities the number of organisms could be extremely high, most typically lacking a high-quality, genome-scale metabolic network, which makes the identification of shared components between organisms even more complicated. For this reason, current approaches have been applied to well-known microbial consortia, including only a limited number of organisms, typically 2 or 3 (dos Santos *et al.*, 2013; Khandelwal *et al.*, 2013; Zomorrodi and Maranas, 2012).

To overcome this issue, the use of a supraorganism or metanetwork has been proposed (Borenstein, 2012), which ignores boundaries for each organism, but models community-level metabolism. In an early work (Greenblum *et al.*, 2012), using a graph theoretical approach, metagenomic data were used to reconstruct the human gut microbiome metanetwork in different conditions, finding key variations in patients with obesity and inflammatory bowel disease.

In this study, we exploit this metanetwork strategy and present a novel computational procedure for obtaining a context-specific metabolic network for a bacterial community using metaproteomic data. In contrast with the work presented in Greenblum *et al.* (2012), we did not use a graph-theoretical approach, but a constraint-based one, which takes into account stoichiometric relationships. In particular, our approach takes some ingredients from the mathematical optimization model presented in the Model SEED (Henry *et al.*, 2010). However, our approach is fundamentally different: it is designed for bacterial communities, not for a single organism, and focuses on the usage of metaproteomic data, which directly leads to a contextualized network that gives cohesion to identified proteins. We also use the taxonomic assignment of the identified proteins to favour the inclusion of enzymes annotated in the genomes of those organisms in cases where such information is available.

Our approach is applied to draft the metabolic networks of two different, naphthalene-enriched communities (Guazzaroni *et al.*, 2013) derived from an anthropogenically influenced, polyaromatic hydrocarbon (PAH)-contaminated soil with (CN2) or without (CN1) bio-stimulation with calcium ammonia nitrate, NH_4NO_3 and KH_2PO_4 and the commercial surfactant Iveysol®. Naphthalene, a model PAH compound, is a common, persistent pollutant in crude oil and industrial chemical manufactures that can be released into the environment (i.e. soils) through anthropogenic activities (Kästner, 2000). Current treatments for naphthalene- and other PAH-contaminated sites involve the use of bio-surfactants and additional electron acceptors as well as nitrogen sources (nitrate and ammonia) to improve the bioavailability and bioremediation of these compounds. It has also been observed that many bacteria are capable of degrading and growing on naphthalene (Guazzaroni *et al.*, 2013; Lu *et al.*, 2011), and their activities might only be limited by environmental conditions. Thus, gaining insight into the mechanisms underlying naphthalene degradation can aid in the design of better remediation strategies.

2 Methods

Here, we present our computational procedure for determining a functional, context-specific metabolic network for bacterial communities using metaproteomic data. Based on a reference metabolic database (Henry *et al.*, 2010), we seek a functional network that includes the maximum number of measured proteins (highly likely set, *H*) in a given sample. We may have evidence that some proteins are

not expressed (lowly likely set, *L*) and, therefore, their participation is minimized. Then, we complete the network using enzymes in the database, preferably those annotated in active organisms in the community (medium likely set, *M*) (Guazzaroni *et al.*, 2013).

We denote the set of enzymes from the reference database not included in *H*, *L* and *M* as *D*, namely *D* involves the subset of non-identified enzymes that are currently annotated for organisms not present in the community. By linking proteins to reactions via Enzyme Commission (EC) numbers (Bairoch, 2000), sets *H*, *L*, *M* and *D* may also refer to reactions.

When we refer to a functional network, we mean a subset of reactions that are able to produce biomass at steady state under the specified medium conditions. We describe these conditions in detail and introduce the mathematical notation below.

We denote the sets of reactions and compounds in the reference metabolic database as *R* and *C*, respectively. The set of reactions is typically classified into reversible and irreversible reactions. For convenience, both reversible and irreversible reactions are divided into two non-negative steps: forward and backward reactions. We define the set $B = \{(f,b) |$ reaction f and reaction b are the reverse of each other, $f < b\}$. For each reaction $i \in R$ we define a flux variable, v_i, and a binary variable, z_i, where $z_i = 1$, if $v_i > 0$, 0 otherwise. We denote the stoichiometric coefficient associated with the metabolite $i \in C$ and reaction $j \in R$ as s_{ij}. This information is stored in the stoichiometric matrix, *S*.

The steady-state assumption implies mass balancing and, therefore, the accumulation/depletion of metabolites inside the system is not possible, as observed in Equation (1). The definition of the boundaries of the system is an important issue. As noted above, in complex bacterial communities, the previous knowledge of shared input/output metabolites is typically not available. For the sake of simplicity, we only include boundaries for the whole community and remove boundaries between individual organisms. Therefore, we obtain a metanetwork in which the identified proteins from various organisms in the community are coexpressed. Using exchange reactions, we then define metabolites that can be taken up from outside (the boundaries of) the system (culture medium conditions) and those that can be excreted outside (the boundaries of) the system, which may prevent the network from utilizing unavailable nutrient sources.

$$S\nu = 0 \tag{1}$$

As our aim is to obtain a metabolic network that supports growth, we must define a biomass reaction. Given that we are using a metanetwork strategy, the biomass reaction represents a consensus equation for all organisms in the community. Note that determining an appropriate biomass reaction is a challenging task, even for single organisms (Feist and Palsson, 2010). However, using an existing biomass reaction from a different organism is a common practice (Nogales *et al.*, 2008), as many constitutive compounds are shared across a wide range of organisms. Equation (2) forces a minimum flux through the biomass reaction (v_{bio}).

$$v_{bio} \geq 1 \tag{2}$$

As fluxes are non-negative, their lower bound is 0, as observed in Equation (3). We also fixed a sufficiently large value for their upper bounds.

$$0 \leq v_i \leq 1000 \quad \forall i \in R \tag{3}$$

As noted above, our approach takes some elements of the mathematical optimization model presented in the Model SEED (Henry *et al.*, 2010). However, our purpose is different, as we aim to obtain a context-specific reconstruction for bacterial communities, not for a single organism, based on metaproteomic data. Instead of a general metabolic reconstruction, our aim is to build a network as specific to the observed phenotype as possible, given the measured data. To this end, we include important technical differences. In particular, we use a 3-step iterative procedure based on linear programming and mixed integer linear programming. We describe each of these three steps below. A graphical summary of the complete workflow of our approach can be found in Fig. 1.

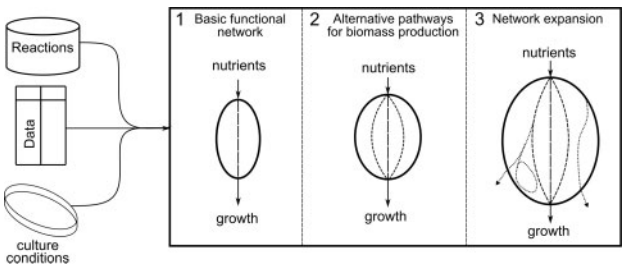

Fig. 1. Proposed reconstruction workflow. The context-specific metabolic network reconstruction algorithm starts from a database of reactions, experimental (metaproteomic) data and knowledge about growth medium as input data. It involves three steps: (i) construction of a basic network capable of using the available nutrients to produce biomass; (ii) addition of alternative pathways for biomass production and (iii) network expansion with pathways not necessarily involved in biomass production

2.1 Step 1: basic functional network

Due to regulatory effects, the experimental measurement of proteins is not sufficient to guarantee their activity (Seifert *et al.*, 2013). This is typically observed in the conflicting trade-off between enzymes in the *H* and *L* sets (Agren *et al.*, 2012; Becker and Palsson, 2008; Shlomi *et al.*, 2008). In other words, the use of all enzymes in *H* may involve a considerable number of enzymes in *L*, whose participation in the reconstruction must be in general avoided (Åkesson *et al.*, 2004). For this reason, we prefer to leave the decision of selecting enzymes in *H* and *L* to the optimization model, which incorporates evidence from metaproteomic data in the objective function. This allows metaproteomic data to influence the resulting network, without directly constraining it.

The objective function is presented in Equation (4). In particular, the flux activity is guided by its penalization, p_i, and bonus, b_i, terms, similar to what is performed in the Model SEED (Henry *et al.*, 2010).

$$\text{minimise} \quad \sum_{i \in R} (p_i - b_i) v_i \tag{4}$$

where p_i and b_i are the sums of various concepts. Weights are defined such that, by minimizing Equation (4), which is subject to Equations (1)–(3), we obtain a functional metabolic network in which fluxes in *H* will prevail, followed by those in *M*, then those in *D* and, finally, those in *L*. As in the Model SEED (Henry *et al.*, 2010), we penalized reversibility changes and favoured the completion of KEGG modules that were substantially covered with metaproteomic data. Finally, in order to avoid the flux bias introduced by the specific stoichiometric representation of each reaction (Brochado *et al.*, 2012), weights were rescaled using maximum flux values obtained from flux variability analysis (Mahadevan and Schilling, 2003). See Supplementary Material I for further details.

It should be noted that the introduction of continuous fluxes in Equation (4), in contrast to the Model SEED (Henry *et al.*, 2010), which includes binary variables (*z*), does not guarantee the optimal use of metaproteomic data, i.e. the reactions in *H*. The removal of binary variables converts a highly expensive, mixed integer linear

program into a linear program, which can be easily solved; however, optimal solutions obtained from linear programming are extreme points, which, in conjunction with our minimization objective function, generate networks involving a limited number of degrees of freedom. Through Steps 2 and 3 described below, we aim to further exploit experimental evidence from protein expression data.

2.2 Step 2: alternative pathways for biomass production

Once Step 1 is solved, we obtain a list of active reactions, N_1. In this second step, we aim to capture alternative pathways for biomass production that are not included in N_1 using the reactions in H, but not in L. To this end, we block each of the reactions in N_1, one-by-one, and resolve the linear program posed in Step 1, i.e. Equation (4), which is subject to Equations (1)–(3). As a result, we obtain a number, card(N_1), of functional networks. The rule here is to merge N_1 with those networks that include additional reactions in H, but not in L. Therefore, we obtain a functional network (N_2) that makes better use of the metaproteomic data for biomass production.

Note that the solution from Step 1 is the result of solving a linear program, whose optimal solution is an extreme point, a solution with zero degrees of freedom. For that reason, all the reactions involved in N_1 are necessary to produce biomass. As relevant alternative routes for biomass production are added into the model using Step 2, the number of degrees of freedom of the resulting solution is increased, while reducing the number of essential reactions/enzymes.

2.3 Step 3: network expansion

Once Step 2 is concluded, we may not have included all measured enzymes in N_2. We denote K as a particular set of these enzymes that we aim to include in the reconstruction. Note that K will typically involve all enzymes from H that are not included in N_2, possibly obtaining a maximum use of the metaproteomic data. However, if our purpose is to emphasize the metabolic differences between two conditions under study, K could involve a subset of them, namely, those that are differentially expressed. To achieve this goal, we address one further optimization problem.

We begin from Equations (1) and (3). The constraint on biomass production, Equation (2), is removed, as it is currently satisfied in N_2. We now make use of binary variables, z_i. In particular, Equation (5) relates v and z variables, where M is the maximum flux, and the minimum (non-zero) flux is 1. Equation (6) prevents reaction f and its reverse b from being active simultaneously.

$$z_i \leq v_i \leq M z_i \quad \forall i \in R \tag{5}$$

$$z_f + z_b \leq 1 \quad \forall (f,b) \in B \tag{6}$$

Then, for each enzyme, $j \in K$, we introduce a continuous variable, e_j, with a value between 0 and 1, as observed in Equation (7). In Equation (8), if any of the set of reactions, R^j, that are associated with enzyme $j \in K$ cannot be activated, then e_j is necessarily 1; therefore, to maximize the use of the enzymes in K, we must minimize the e_j variables. This is achieved by amending the objective function as in Equation (9). In particular, for the e_j variables, we assign the maximum overall penalty, w_j.

$$0 \leq e_j \leq 1 \quad \forall j \in K \tag{7}$$

$$\sum_{i \in R^j} z_i + e_j \geq 1 \quad \forall j \in K \tag{8}$$

$$\text{Minimise} \quad \sum_{i \in R} (p_i - b_i) v_i + \sum_{j \in K} w_j e_j \tag{9}$$

Equation (9), which is subject to Equations (1), (3), and (5)–(9), is a mixed linear integer program and empirical evidence shows that it is not an expensive problem (<100 s in the instances considered in Section 3). Active reactions from this optimization problem are added to N_2 and define the final resulting metabolic network, N_3.

3. Results

3.1 Reconstruction of the CN1 and CN2 functional networks

The approach presented above was applied to draft the context-specific metabolic networks of two different naphthalene-enriched communities, CN1 and CN2 (Guazzaroni *et al.*, 2013). These communities were obtained from an anthropogenically influenced, PAH-contaminated soil with (CN2) or without (CN1) bio-stimulation. Metaproteomic data from CN1 and CN2 constituted the input information in our study (Guazzaroni *et al.*, 2013).

In our analysis, we only considered proteins with an annotated metabolic function, i.e. with an EC number, namely 570 out of 1234 measured proteins, collectively involving 327 unique EC identifiers. Based on the relative protein concentrations, we classified enzymes found in CN1 and CN2 (Guazzaroni *et al.*, 2013) into the H, L, M and K sets as follows (see Section 2). For one scenario, enzymes listed in that sample were included in the H set, while enzymes that did not appear in that sample, but did appear in the other scenario, were included in the L set. As we were interested in obtaining networks that emphasized the differences between both scenarios, the K set involved up-regulated enzymes in each scenario. In particular, enzymes showing a 1.5-fold change in their relative protein concentrations in one sample compared with the other were considered up-regulated.

Note that the relative protein concentrations were calculated as previously described (Guazzaroni *et al.*, 2013). Briefly, similar to IBAQ values (Arike *et al.*, 2012), protein intensities per sample (CN1 or CN2) were calculated as the average peptide intensities in the replicas. Obtained quantitative values were normalized by median division. Then, these protein intensities (per sample) were summed and the relative concentration of an individual protein (per sample) was calculated by dividing the protein intensities by the summed protein intensities. Replicate values were finally averaged. This approach should reflect the individual protein abundance in each community proteome.

Using full-length and partial 16S rRNA gene sequences obtained through a metagenomic approach (Guazzaroni *et al.*, 2013), it was found that 13 and 12 distinct species constituted the CN1 and CN2 communities, respectively, with only two species (*Achromobacter* and *Azospirillum*) conforming to the common set (see Supplementary Material I). While *Azospirillum*, *Comamonas*, *Achromobacter* and *Pseudoxanthomonas* species dominated CN1, *Pseudomonas* and *Achromobacter* species dominated CN2. This information was used to aid in the context-specific network reconstruction process. In particular, the set of related genome annotations for CN1 and CN2, which was established on the basis of phylogenetic affiliations (Guazzaroni *et al.*, 2013), was obtained from the KEGG website. The enzymes (ECs) from these genome annotations, which were neither included in H nor L, were included in the M set. Full details regarding these organisms and genome annotations can be found in the Supplementary Material I.

The list of reactions and metabolites was downloaded from the Model SEED database (Henry *et al.*, 2010). The above enzyme lists were translated into reaction lists using their EC numbers annotated in this database. The D set comprised ECs (enzymes) from the Model SEED database not included in H, M and L. When a reaction

was associated with more than one EC belonging to different sets, the reaction was assigned to the most favourable set. For example, if a reaction could be catalyzed by one enzyme from *H* and one from *L*, then the reaction was assigned to *H*.

A minimal medium based on naphthalene as the only carbon source (as was used in the enrichment cultures; Guazzaroni *et al.*, 2013) was defined for the reconstruction process (see Supplementary Material I). The biomass reaction was taken from a *Pseudomonas* reconstruction that was provided by SEED (rxn12834), as this specie plays a major role in CN2. We also used annotated modules from KEGG.

The computation time for both CN1 and CN2 reconstruction was <200 s. All computations were performed on a 64-bit Windows XP machine with an Intel Core 2 CPU at 2.4 GHz and 8 GB of RAM. The code was written in MATLAB and CPLEX was used to solve the linear optimization problems.

The complete details of the reactions, metabolites, and enzymes involved in the contextualized reconstruction of CN1 and CN2 can be found in the Supplementary Material II. Note that randomly perturbing the selected weights with a 10% uniform noise only changed a few reactions, giving rise to very similar networks (see Supplementary Material I). For CN1, we used 148 of the 206 enzymes from *H* and 21 enzymes from *L*, and we completed the network with 274 and 165 enzymes from *M* and *D*, respectively. Similarly, for CN2, we employed 259 of the 311 enzymes from *H* and 1 enzyme from *L*; and we completed the network with 267 and 282 enzymes from *M* and *D*, respectively. In both cases, the use of enzymes from *H* was remarkable, corresponding to >70% of the measured data, which was increased to ∼90% for up-regulated enzymes (*K* set). In contrast, the number of enzymes in *L* required a more careful reading.

As noted above, there are reactions in the Model SEED database that involve more than one EC number and are, therefore, catalyzed by different enzymes. For example, if a reaction is catalyzed by one enzyme from *H* and one from *L*, we assume that the flux through this reaction is supported by the enzyme from *H*, which is consistent with the experimental data. An inconsistency arises when reactions that are exclusively catalyzed by enzymes in *L* are included in the reconstruction. We found four reactions of this type in CN1, which collectively involved three enzymes of 21. In particular, two of these inconsistent reactions in CN1 were associated with EC 2.5.1.9 (*riboflavin synthase*) and are required to produce FAD (*flavin adenine dinucleotide*), an essential metabolite for biomass production. The third reaction was linked to EC 2.4.1.227 and is required to produce the *peptidoglycan* subunit of *P. putida KT2440*, which is involved in biomass production. As this metabolite is specific for *Pseudomonas*, which is not involved in CN1, the need for this enzyme is unlikely. The fourth reaction is associated with EC 3.5.1.18 and is activated to support up-regulated enzymes. There is only one inconsistent reaction in CN2, which is associated with EC 2.7.7.38 and their activation is due to the same reason as for EC 3.5.1.18. The inclusion of these enzymes in CN1 is not in accord with the evidence from metaproteomic data, which may be attributed to three possible causes: (i) incompleteness of the Model SEED database; (ii) inaccuracy of the biomass reaction or (iii) a lack of resolution in the metaproteomic data. To address this issue, further experimental evidence is required.

With the resulting context-specific networks for CN1 and CN2, we decided to evaluate how single-reaction deletions could hamper their ability to produce biomass. CN2 turned out to be more resilient, as only 22 single-reaction deletions prevented its biomass production capacity, in contrast with 42 single-reaction deletions in CN1. However, the overlap was significant, as 19 of those reactions affected both networks. When deleting enzymes related to a given EC number, 31 instances affected growth in CN2 and 42 in CN1, with 23 of them being the same in both cases. In addition, we substituted naphthalene as the only carbon source with each of the compounds present in the reconstructed networks. CN1 was able to take advantage of 26 compounds to produce biomass, while CN2 exhibited theoretical ability to use 446 compounds. We conducted the same analysis with the nitrogen, phosphorus and sulphur sources, finding that CN1 could make use of 166, 114 and 34 compounds, respectively, while CN2 could make use of 270, 212 and 104 compounds, respectively. Although these results should be taken with caution, they suggest that the metabolism of CN2 is more robust and varied than CN1. The fact that the availability of substrates is promoted during the bio-stimulation process used for obtaining a CN2 community (Guazzaroni *et al.*, 2013) might agree with this hypothesis.

3.2 CN1 and CN2 pathway analysis

To obtain a global picture of the pathways characterized in the CN1 and CN2 contextualized networks, we resorted to the use of KEGG maps (see Supplementary Material I). In particular, to extract the functional differences between CN1 and CN2, we compared the KEGG maps using a score, J_p, derived from the Jaccard distance. In particular, for each KEGG map, we first calculated the Jaccard index, J, between CN1 and CN2, with $J = |A \cap B|/|A \cup B|$, where A and B represent the set of reactions involved in CN1 and CN2, respectively, in a given KEGG map. Then, we determined the Jaccard distance, $J_\delta = 1 - J$, which measures the dissimilarity between CN1 and CN2 for a particular pathway. Finally, we multiplied the Jaccard distance by the maximum between the number of reactions that belonged to CN1, but not to CN2, and vice versa, i.e. $J_p = J_\delta \cdot \max(|A \cap \overline{B}|, |\overline{A} \cap B|)$. This score gives more importance to pathways where the CN1 and/or CN2 networks show high coverage and share few reactions. An illustration of this process can be found in Supplementary Material I for the 'Histidine metabolism' KEGG map (Supplementary Fig. S1). Functional differences between CN1 and CN2 can be analysed via J_p, where the higher the value of J_p, the greater the difference between CN1 and CN2 and, hence, the more relevant the pathway.

We ranked the KEGG pathways according to this measure for the CN1 and CN2 metabolic networks. Table 1 shows some of the top most different KEGG pathways between CN1 and CN2. We repeated the same analysis in two additional cases: (1) direct use of metaproteomic data from CN1 and CN2, neglecting our network reconstruction approach ('Rank only metaproteomics'); (2) removal of metaproteomic data, only considering CN1 and CN2 taxonomic data and their annotated genomes in our network reconstruction approach ('Rank taxonomics'). As observed in Table 1, substantial differences can be found among them, which emphasize the effect of our reconstruction approach, showing a clear contribution of proteomics to genomics data. The full details can be found in the Supplementary Material I and III.

Table 1 shows clear differences between CN1 and CN2. The geraniol degradation pathway (map00281) was predicted to be completely functional in CN2, but inactive in CN1. In CN2, enzymes from *H* in this pathway were complemented with enzymes from *M* and *D*. In contrast, in CN1, enzymes from *H* were discarded from the reconstruction. On the other hand, the fluorobenzoate degradation pathway (map00364) was filled in to some extent in the CN1 reconstruction, whereas it was inactive in CN2. Note that these differences between CN1 and CN2 cannot be easily obtained from the other two cases considered (see 'Rank only metaproteomics' and

Table 1. Ranking of KEGG pathways after reconstruction using functional network data for CN1 and CN2

KEGGID	Name	CN1	CN2	Score	Rank	Rank (metaproteomics)	Rank (taxonomics)
map00071	Fatty acid metabolism	4	27	19.5926	1	22	1
map00062	Fatty acid elongation	0	15	15	2	42	12
map00330	Arginine and proline metabolism	17	31	14.0541	3	19	10
map00540	Lipopolysaccharide biosynthesis	3	18	12.5	4	54	39
map00760	Nicotinate and nicotinamide metabolism	13	25	12.4667	5	24	31
map00230	Purine metabolism	42	57	12	6	5	23
map00281	Geraniol degradation	0	12	12	7	37	–
map00260	Glycine, serine and threonine metabolism	14	26	9.31034	8	33	21
map00400	Phenylalanine, tyrosine and tryptophan biosynthesis	12	24	8.61538	9	14	79
map00500	Starch and sucrose metabolism	5	10	8.35714	10	62	8
map00523	Polyketide sugar unit biosynthesis	0	8	8	11	20	6
map00620	Pyruvate metabolism	12	23	7.8	12	36	28
map00130	Ubiquinone and other terpenoid-quinone biosynthesis	2	9	7.2	13	2	7
map00364	Fluorobenzoate degradation	7	0	7	14	102	–
map00650	Butanoate metabolism	14	12	6.85714	15	43	33

The columns 'CN1' and 'CN2' indicate the number of reactions involved in CN1 and CN2 reconstructions active in the KEGG pathway under consideration. The 'Rank' column indicates the position of KEGG pathways according to descending order of the obtained score. The 'Rank (metaproteomics)' column indicates the rank obtained for KEGG pathways before the reconstruction process, namely with the score exclusively calculated from metaproteomic data of CN1 and CN2. The 'Rank (taxonomics)' indicates the rank obtained after the reconstruction only with taxonomic data, i.e. with empty *H*, *L* and *K* sets

'Rank taxonomics'). This is particularly relevant since these differences between CN1 and CN2 are experimentally validated below, which shows the predictive power and need of the approach presented here.

3.3 Experimental analysis of fluorobenzoate and geraniol metabolism in CN1 and CN2

Given the high rank obtained by the fluorobenzoate and geraniol degradation pathways and its specificity for CN1 and CN2, respectively, we evaluated the correctness of these hypotheses. First, *in silico* stoichiometric analysis showed that CN1 was capable of growing with fluorobenzoate as the sole carbon source, and the same was observed for CN2 and geraniol, after some minor corrections to the obtained networks. In particular, for CN1 to produce biomass from fluorobenzoate, we needed to allow for the net production of fluoride (F-), which is found in abundance in soils (McQuaker and Gurney, 1977) and in bacterial cultures as cellular degradation by-product (Hidde Boersma *et al.*, 2004). In the case of CN2 and geraniol, we needed to change the direction of reaction rxn07886 (*geranic acid CoA-transferase*) in the SEED Database, which converts geranic acid into *trans*-geranyl-CoA and was originally defined to act in the opposite direction. Based on KEGG (map00281) and existing literature (Clemente-Soto *et al.*, 2014), we found that this reaction is commonly annotated in the direction proposed. Note here that the opposite is not possible, i.e. the growth of CN2 and CN1 on fluorobenzote and geraniol, respectively, as they are not active in CN2 and CN1, respectively. It is important to clarify that these issues come from inaccuracies in The SEED Model database and not from the algorithm presented here. When doing this modification prior to the reconstruction process, the resulting networks directly grow on fluorobenzoate in CN1 and geraniol in CN2.

Secondly, in order to discard that the relevance of fluorobenzoate in CN1 and geraniol in CN2 is an artefact derived from an inaccurate biomass equation, we conduct a sensitivity analysis with different existing biomass equations, finding that the major conclusions achieved are conserved in most cases (see Supplementary Material I for details).

Experimental validation assays were conducted to prove the extent of agreement with our computational predictions. For that, we set up CN1 and CN2 enrichment cultures using previously described conditions (Guazzaroni *et al.*, 2013); instead of naphthalene as the carbon source, geraniol and 3/4-fluorobenzoate (0.1% w/v) were used, and samples were taken at different time points (see Supplementary Material I). Fingerprinting by Gas Chromatography-Mass Spectrometry (GC-MS) was used to confirm the presence of the initial substrates as well as the existence of degradation intermediates in both cultures. A careful inspection of the MS signatures of the initial metabolites known to participate in geraniol (map00281) and 3/4-fluorobenzoate (map00364) degradation (see Supplementary Material I) confirmed the presence of 3/4-fluorocatechol in CN1 and citral and geranic acid in CN2. These findings demonstrated that the fluorobenzoate-degradation pathway occurred or was active in CN1, while the geraniol-degradation pathway is active in CN2. This was also confirmed by measuring the OD_{600} of the enrichment cultures at different time intervals (Fig. 2). As shown, CN1 grew only in the presence of fluorobenzoate (0.1% *w/v*), whereas CN2 grew only in the presence of geraniol (0.1% *w/v*).

3.4 Contributions of bacteria to the CN1 and CN2 functional networks

We also attempted to quantify the contributions of particular sets of microbes to the entire reconstructed, context-specific metabolic network, where multiple proteins from multiple organisms are coexpressed. This is an important advance because the complement of proteins used to metabolize recalcitrant pollutants and the specific roles of different bacterial members within a consortium in pollutant (or other potential carbon/energy sources) deconstruction are not well explored.

As the population diversity and structures of the two enrichment cultures were relatively low and well characterized, the taxonomic affiliations of the proteins quantified in the shotgun metaproteomes could be unambiguously established (Guazzaroni *et al.*, 2013). Based on this, for CN1 and CN2, we knew which members of the community were actually expressing the enzymes used to catalyse each reaction in *H*.

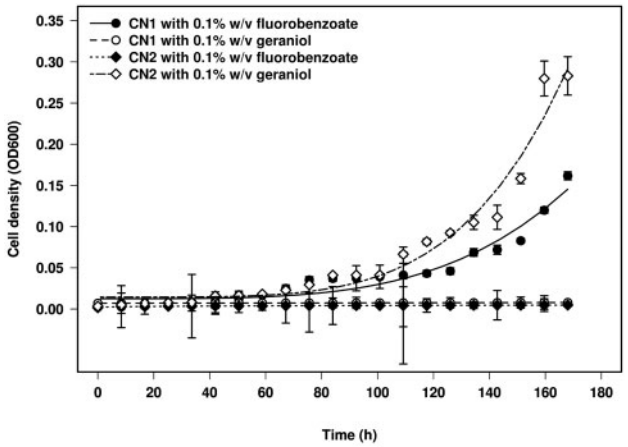

Fig. 2. Growth curve of CN1 and CN2 enrichment cultures in Bushnell Hass minimal medium in the presence of 0.1% (*w/v*) 3/4-fluorobenzoate and geraniol, respectively, at 30°C and 250 rpm. As shown, within the examined time frame, no appreciable growth was observed in cultures of the CN1 and CN2 consortia in the presence of geraniol and fluorobenzoate, respectively

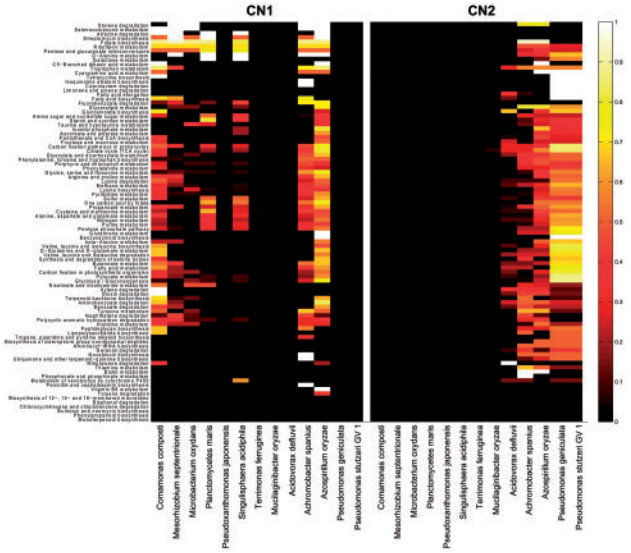

Fig. 3. Heatmap showing the contributions of the most relevant bacterial members of CN1 and CN2 to the KEGG maps. Relative contributions of each of the 13 distinct species found to constitute the CN1 and CN2 communities (Guazzaroni *et al.*, 2013) are differentiated by a colour code. A high-resolution image can be found in Supplementary Material V

To evaluate the role of each bacterial member in CN1 and CN2 at the functional level, we determined its contribution to each KEGG map. The contribution was determined as the number of times a bacterium appeared in a KEGG map divided by its total number of active reactions. For this analysis, we only took into account the reactions in *H* and *M* that were involved in the CN1 and CN2 reconstructed networks. As noted above, the taxonomic affiliation was known for the reactions in *H*. In contrast, for the reactions in *M*, different members of the community might be involved in a reaction. For simplicity, in these situations, if possible, we assigned an organism that was previously included in the KEGG map via the reactions from *H*. Full details as to the taxonomic assignment of reactions involved in the CN1 and CN2 metabolic networks can be found in Supplementary Material IV.

Figure 3 shows the contribution of each organism found in both CN1 and CN2 to each KEGG map. Pathways were reconstructed for the most abundant populations, which included composite genomes for populations closely related to sequenced strains of *Achromobacter*, *Azospirillum*, *Comamonas*, *Mesorhizobium*, *Microbacterium*, *Planctomycetes*, *Pseudoxanthomonas*, *Singulisphaera* and *Pseudomonas*.

Identification of genes for naphthalene processing (map00626) and metabolic reconstructions suggested *Achromobacter* followed by *Mesorhizobium* and *Pseudoxanthomonas* in CN1 and mainly *Achromobacter* in CN2 as key groups for naphthalene degradation. In addition, we identified *Achromobacter*, *Azospirillum*, and *Comamonas* in CN1 and *Azospirillum* as well as *Pseudomonas* in CN2 as groups that might primarily metabolize low molecular weight molecules produced from naphthalene. It could also be observed that, while metabolic reconstructions indicated a central role played by *Achromobacter* in naphthalene degradation, multiple bacteria participated in the active pathways (see Supplementary Material I for further details; Figs. S3–S5). A careful examination of the data presented in Fig. 3 clearly leads to a different pathway organization at the organismal level.

4 Conclusion

In this study, we present a novel computational procedure for obtaining a context-specific functional metabolic network for a bacterial community using metaproteomic data. Our approach was based on the mathematical optimization model presented in the Model SEED (Henry *et al.*, 2010). However, we adapt this model to incorporate metaproteomic data and obtain a context-specific metanetwork in which the identified proteins from the multiple organisms making up the community are coexpressed. To this end, we also include important technical differences. In particular, we use a 3-step iterative procedure based on linear programming and mixed integer linear programming.

Our approach is an alternative to previously reported methods (dos Santos *et al.*, 2013; Khandelwal *et al.*, 2013; Zomorrodi and Maranas, 2012), where the role of each organism is explicitly represented in a different metabolic compartment and, therefore, their relationships can be directly analysed. These methods require the genome-scale metabolic network of each organism in the community as input data, which, in consortia involving a high number of species as we have here, is typically not available; therefore, we turn to a metanetwork approach, which involves several assumptions. First, we need a consensus biomass equation that represents the metabolic requirements of the community to support growth. With metametabolomics approaches being developed, it is expected that consensus biomass equations will be refined in the near future. Second, a free exchange of metabolites between species is allowed, as boundaries between individual organisms are not defined. However, a metanetwork could serve as a basis to disentangle the role of each organism in the community, as suggested in Section 3.4. More sophisticated approaches need to be developed for this task, for example, analysing the role of a single organism in the context of the entire metanetwork.

Our approach was applied to draft the context-specific metabolic networks of two different naphthalene-enriched communities (Guazzaroni *et al.*, 2013). Analysis of the resilience to single-reaction elimination and the ability to grow on different sources suggests that CN2 metabolism is more varied than CN1. Then, we used KEGG maps to obtain a global picture of the reconstructed draft networks. We were able to capture the overall functional differences between

CN1 and CN2 at the metabolic level. We showed that CN1 and CN2 utilize different metabolic pathways to synthesize essential metabolites for growth. In particular, we hypothesized an important role for the fluorobenzoate degradation pathway in CN1 and for geraniol metabolism in CN2. Experimental validation was conducted and good agreement with our computational predictions was observed.

On the other hand, we showed that these metabolic differences lead to a different pathway organization at the organismal level. For example, while naphthalene degradation (map00626) seems to be supported by *Achromobacter* in both CN1 and CN2, *Mesorhizobium septentrionale* and *Pseudoxanthomonas japonensis* may be involved in an alternative pathway in CN1. In addition, while metagenomic sequences outlined the broad metabolic capabilities of the abundant populations present in an adapted community, proteomics-guided metabolic reconstructions allowed us to focus on the pathways that were actually expressed and refine the assignment of roles for community members not only in naphthalene degradation but also in the assimilation of the low molecular weight compounds produced from it.

These results show that network-based methods represent a promising strategy for exploiting the value of data and the available bioinformatics tools, allowing us to obtain a better understanding of biological systems. As the available meta-omics data from scientific studies at different levels are increasing, reconstruction procedures will play an important role in disentangling contexts-specific metabolic phenotypes. The approach presented here can be extended to meta-genomic and meta-transcriptomic data and will clearly benefit from the availability of meta-metabolomic data, mainly to address the failure to detect all different enzymes (ECs) that catalyze different reactions. Amending our approach to include these data is straightforward.

Funding

This work was supported by the Basque Government [to L.T. and J.P.], a Severo Ochoa fellowship from FICYT (Principado de Asturias) [to I.L], the Spanish Ministry of Economy and Competitiveness (projects CSD2007-00005, BIO2011-25012 and BIO2013-48933), the European Community projects MAGICPAH (FP7-KBBE-2009-245226), ULIXES (FP7-KBBE-2010-266473) and KILLSPILL (FP7-KBBE-2012-312139), the European Regional Development Fund (ERDF), and a DFG funded project within the framework of SPP1319 [to FA.H., J.Seif. and M.v.B.].

Conflict of Interest: none declared.

References

Agren,R. *et al.* (2012) Reconstruction of genome-scale active metabolic networks for 69 human cell types and 16 cancer types using INIT. *PLoS Comput. Biol.*, **8**, e1002518.

Agren,R. *et al.* (2013) The RAVEN toolbox and its use for generating a genome-scale metabolic model for *Penicillium chrysogenum*. *PLoS Comput. Biol.*, **9**, e1002980.

Åkesson,M. *et al.* (2004) Integration of gene expression data into genome-scale metabolic models. *Metab. Eng.*, **6**, 285–293.

Arike,L. *et al.* (2012) Comparison and applications of label-free absolute proteome quantification methods on *Escherichia coli*. *J. Proteomics*, **75**, 5437–5448.

Bachmann,H. *et al.* (2013) Availability of public goods shapes the evolution of competing metabolic strategies. *Proc. Natl Acad. Sci.*, **110**, 14302–14307.

Bairoch,A. (2000) The ENZYME database in 2000. *Nucleic Acids Res.*, **28**, 304–305.

Becker,S.A. and Palsson,B.O. (2008) Context-specific metabolic networks are consistent with experiments. *PLoS Comput. Biol.*, **4**, e1000082.

Beloqui,A. *et al.* (2008) Recent trends in industrial microbiology. *Curr. Opin. Microbiol.*, **11**, 240–248.

Bordbar,A. and Palsson,B.O. (2012) Using the reconstructed genome-scale human metabolic network to study physiology and pathology. *J. Intern. Med.*, **271**, 131–141.

Borenstein,E. (2012) Computational systems biology and in silico modeling of the human microbiome. *Brief. Bioinform.*, **13**, 769–780.

Boyle,P.M. and Silver,P.A. (2012) Parts plus pipes: synthetic biology approaches to metabolic engineering. *Metab. Eng.*, **14**, 223–232.

Brochado,A.R. *et al.* (2012) Impact of stoichiometry representation on simulation of genotype-phenotype relationships in metabolic networks. *PLoS Comput. Biol.*, **8**, e1002758.

Carter,K.K. *et al.* (2012) Pathway engineering via quorum sensing and sRNA riboregulators—interconnected networks and controllers. *Metab. Eng.*, **14**, 281–288.

Clemente-Soto,A.F. *et al.* (2014) Potential mechanism of action of meso-dihydroguaiaretic acid on mycobacterium tuberculosis H37Rv. *Molecules*, **19**, 20170–20182.

Curran,K.A. and Alper,H.S. (2012) Expanding the chemical palate of cells by combining systems biology and metabolic engineering. *Metab. Eng.*, **14**, 289–297.

dos Santos,F.B. *et al.* (2013) Towards metagenome-scale models for industrial applications—the case of Lactic Acid Bacteria. *Curr. Opin. Biotechnol.*, **24**, 200–206.

Feist,A.M. and Palsson,B.O. (2010) The biomass objective function. *Curr. Opin. Microbiol.*, **13**, 344–349.

Greenblum,S. *et al.* (2012) Metagenomic systems biology of the human gut microbiome reveals topological shifts associated with obesity and inflammatory bowel disease. *Proc. Natl Acad. Sci.*, **109**, 594–599.

Guazzaroni,M.-E. and Ferrer,M. (2011) Metagenomic approaches in systems biology. In: de Bruijn,F.J. (ed.) *Handbook of Molecular Microbial Ecology I: Metagenomics and Complementary Approaches*. Wiley-Blackwell, John Wiley & Sons, Inc., pp. 475–489.

Guazzaroni,M.-E. *et al.* (2013) Metaproteogenomic insights beyond bacterial response to naphthalene exposure and bio-stimulation. *ISME J.*, **7**, 122–136.

Henry,C.S. *et al.* (2010) High-throughput generation, optimization and analysis of genome-scale metabolic models. *Nat. Biotechnol.*, **28**, 977–982.

Hidde Boersma,F.G. *et al.* (2004) A 19F NMR study of fluorobenzoate biodegradation by *Sphingomonas* sp. HB-1. *FEMS Microbiol. Lett.*, **237**, 355–361.

Jerby,L. *et al.* (2010) Computational reconstruction of tissue-specific metabolic models: application to human liver metabolism. *Mol. Syst. Biol.*, **6**, 401.

Kästner,M. (2000) Degradation of aromatic and polyaromatic compounds. In: Rehm,H.J., and Reed,G. (eds) *Biotechnology, Environmental Processes*.Wiley Vch, Germany, pp. 211–271.

Khandelwal,R.A. *et al.* (2013) Community flux balance analysis for microbial consortia at balanced growth. *PLoS ONE*, **8**, e64567.

Kinross,J. *et al.* (2011) Gut microbiome-host interactions in health and disease. *Genome Med.*, **3**, 14.

Letunic,I. *et al.* (2008) iPath: interactive exploration of biochemical pathways and networks. *Trends Biochem. Sci.*, **33**, 101–103.

Lu,X.-Y. *et al.* (2011) Bacteria-mediated PAH degradation in soil and sediment. *Appl. Microbiol. Biotechnol.*, **89**, 1357–1371.

Mahadevan,R. and Schilling,C.H. (2003) The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. *Metab. Eng.*, **5**, 264–276.

Mascarelli,A.L. (2009) Geomicrobiology: low life. *Nature*, **459**, 770–773.

McCloskey,D. *et al.* (2013) Basic and applied uses of genome-scale metabolic network reconstructions of *Escherichia coli*. *Mol. Syst. Biol.*, **9**, 661.

McQuaker,N.R. and Gurney,M. (1977) Determination of total fluoride in soil and vegetation using an alkali fusion-selective ion electrode technique. *Anal. Chem.*, **49**, 53–56.

Moran,M.A. *et al.* (2013) Sizing up metatranscriptomics. *ISME J.*, **7**, 237–243.

Nogales,J. *et al.* (2008) A genome-scale metabolic reconstruction of *Pseudomonas putida* KT2440: iJN746 as a cell factory. *BMC Syst. Biol.*, **2**, 79.

Palsson,B. (2009) Metabolic systems biology. *FEBS Lett.*, **583**, 3900–3904.

Pey,J. *et al.* (2013) A network-based approach for predicting key enzymes explaining metabolite abundance alterations in a disease phenotype. *BMC Syst. Biol.*, **7**, 62.

Poblete-Castro,I. *et al.* (2013) In-silico-driven metabolic engineering of *Pseudomonas putida* for enhanced production of poly-hydroxyalkanoates. *Metab. Eng.*, **15**, 113–123.

Rezola,A. *et al.* (2013) Selection of human tissue-specific elementary flux modes using gene expression data. *Bioinformatics*, **29**, 2009–2016.

Röling,W.F. *et al.* (2010) Systems approaches to microbial communities and their functioning. *Curr. Opin. Biotechnol.*, **21**, 532–538.

Seifert,J. *et al.* (2013) Bioinformatic progress and applications in metaproteogenomics for bridging the gap between genomic sequences and metabolic functions in microbial communities. *Proteomics*, **13**, 2786–2804.

Shlomi,T. *et al.* (2008) Network-based prediction of human tissue-specific metabolism. *Nat. Biotechnol.*, **26**, 1003–1010.

Thiele,I. and Palsson,B.O. (2010) A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nat. Protoc.*, **5**, 93–121.

Vitkin,E. and Shlomi,T. (2012) MIRAGE: a functional genomics-based approach for metabolic network model reconstruction and its application to cyanobacteria networks. *Genome Biol.*, **13**, R111.

Yamada,T. *et al.* (2011) iPath2.0: interactive pathway explorer. *Nucleic Acids Res.*, **39**, W412–W415.

Zamboni,N. *et al.* (2008) anNET: a tool for network-embedded thermodynamic analysis of quantitative metabolome data. *BMC Bioinformatics*, **9**, 199.

Zomorrodi,A.R. and Maranas,C.D. (2012) OptCom: a multi-level optimization framework for the metabolic modeling and analysis of microbial communities. *PLoS Comput. Biol.*, **8**, e1002363.

Zomorrodi,A.R. *et al.* (2012) Mathematical optimization applications in metabolic networks. *Metab. Eng.*, **14**, 672–686.