BIOINFORMATICS ORIGINAL PAPER

Vol. 29 no. 16 2013, pages 2009–2016
doi:10.1093/bioinformatics/btt328

Systems biology

Advance Access publication June 6, 2013

Selection of human tissue-specific elementary flux modes using gene expression data

Alberto Rezola[1], Jon Pey[1], Luis F. de Figueiredo[2,3], Adam Podhorski[1], Stefan Schuster[2], Angel Rubio[1,*] and Francisco J. Planes[1,*]

[1]Biomedical Engineering Department, CEIT and Tecnun, University of Navarra, 20009 San Sebastian, Spain, [2]Department of Bioinformatics, School of Biology and Pharmacy, Friedrich Schiller University, 07743 Jena, Germany and [3]Chemoinformatics and Metabolism Team, EMBL-European Bioinformatics Institute, Wellcome Trust Genome Campus, CB10 1SD Hinxston, UK

Associate Editor: Martin Bishop

ABSTRACT

Motivation: The analysis of high-throughput molecular data in the context of metabolic pathways is essential to uncover their underlying functional structure. Among different metabolic pathway concepts in systems biology, elementary flux modes (EFMs) hold a predominant place, as they naturally capture the complexity and plasticity of cellular metabolism and go beyond predefined metabolic maps. However, their use to interpret high-throughput data has been limited so far, mainly because their computation in genome-scale metabolic networks has been unfeasible. To face this issue, different optimization-based techniques have been recently introduced and their application to human metabolism is promising.

Results: In this article, we exploit and generalize the *K*-shortest EFM algorithm to determine a subset of EFMs in a human genome-scale metabolic network. This subset of EFMs involves a wide number of reported human metabolic pathways, as well as potential novel routes, and constitutes a valuable database where high-throughput data can be mapped and contextualized from a metabolic perspective. To illustrate this, we took expression data of 10 healthy human tissues from a previous study and predicted their characteristic EFMs based on enrichment analysis. We used a multivariate hypergeometric test and showed that it leads to more biologically meaningful results than standard hypergeometric. Finally, a biological discussion on the characteristic EFMs obtained in liver is conducted, finding a high level of agreement when compared with the literature.

Contact: fplanes@tecnun.es or arubio@ceit.es

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on February 6, 2013; revised on May 30, 2013; accepted on June 1, 2013

1 INTRODUCTION

With the expansion of high-throughput molecular experimental technologies in the past decade, particularly genomics and transcriptomics, a vast amount of data is available to the scientific community (Brazma, 2003; Edgar, 2002). Their analysis in the context of metabolic pathways is essential to uncover their underlying functional structure, and it is a widely used practice in the field of bioinformatics and systems biology (Curtis *et al.*, 2005). To this end, different approaches have been proposed to date. Some of the existing tools explain high-throughput **'omics' data** in the light of predefined metabolic maps (Dahlquist *et al.*, 2002; Goffard and Weiller, 2007; Mlecnik *et al.*, 2005). Despite their wide application (Ghazalpour *et al.*, 2005; Konradi *et al.*, 2004; Naylor *et al.*, 2005), these methods are limited, as predefined maps do not capture the wide variety of complex metabolic states. Instead, the use of unbiased mathematical pathway concepts based on genome-scale metabolic networks is suitable for this purpose. Among these pathway concepts, path-finding techniques from the field of graph theory have been a recurrent approach (Zien *et al.*, 2000). In contrast to their low computational expense, they face important theoretical issues because they do not consider stoichiometric mass balancing constraints, as recently emphasized in de Figueiredo *et al.* (2009b) and Pey *et al.* (2011). For this reason, other more general pathway concepts have been used to interpret 'omics' data, particularly elementary flux modes (EFMs) (Schuster and Hilgetag, 1994; Schuster *et al.*, 2000) and recently elementary flux patterns (Kaleta *et al.*, 2009a). In this article, we focus on the **EFM** approach.

Despite its early definition, the interpretation of 'omics' data in light of EFMs has been limited so far (Schwarz *et al.*, 2005; Schwartz *et al.*, 2007). The reason is that the number of EFMs explodes in a combinatorial fashion as the network size increases and classical approaches fail to compute them in large networks (Pfeiffer *et al.*, 1999; Terzer and Stelling, 2008). That is the case for the human genome-scale metabolic network (Duarte *et al.*, 2007), which involves several thousand reactions and metabolites. To face this issue, Schwartz *et al.* (2007) computed a set of EFMs for each human Kyoto Encyclopedia of Genes and Genomes (KEGG) map, which is a reductionist approach.

A promising strategy is to compute a subset of EFMs using recently developed optimization techniques (de Figueiredo *et al.*, 2009a; Rezola *et al.*, 2011). In the present work, a generalized version of the algorithm presented in de Figueiredo *et al.* (2009a) is introduced and used to determine **a subset of EFMs in the human genome-scale metabolic network** (Duarte *et al.*, 2007) that captures a wide variety of pathways in human metabolism. In particular, aside from recovering an extensive number of metabolic pathways previously reported in the literature, more

*To whom correspondence should be addressed.

solutions do arise in our subset of EFMs, which should allow us to describe more accurately the repertoire of metabolic states.

In light of our subset of EFMs, we analyzed data from Shlomi *et al.* (2008). In that work, based on **gene expression data**, reactions from Duarte *et al.* (2007) are classified into highly, moderately and lowly expressed in different **healthy human tissues**. We used this **reaction classification** to determine **characteristic EFMs** for each tissue based on statistical enrichment analysis.

Three-level discrete classification of reactions prevented us from using more established enrichment techniques, such as Gene Set Enrichment Analysis (GSEA) (Subramanian *et al.*, 2005). Instead, we introduced a **multivariate hypergeometric test**, which considers highly expressed and lowly expressed reactions in the same statistical score. With this test, we aim to obtain as characteristic those EFMs enclosing an elevated number of highly expressed reactions and as few as possible of lowly expressed reactions. This fact differs from the approach presented in Schwartz *et al.* (2007), which uses the standard hypergeometric test with only one feature (either highly or lowly expressed reactions). We show in different tissues that this improvement provides a more accurate picture of their characteristic EFMs and, therefore, their key metabolic pathways.

2 METHODS

In this section, we first introduce the (*K*,*d*)-shortest EFM algorithm and describe its theoretical advance with respect to the one presented in de Figueiredo *et al.* (2009a). Then, we detail our statistical framework to determine characteristic EFMs based on gene expression data.

2.1 *(K,d)*-shortest EFMs

Consider a metabolic network that comprises R reactions and C metabolites. The set of reactions is typically classified into reversible and irreversible reactions. Reversible reactions are here divided into two irreversible steps that represent forward and backward reactions. We define the set $B=[(\alpha, \beta)|$ reaction α and reaction β are the reverse of each other, $\alpha<\beta]$. For each reaction r ($r=1,\ldots,R$), we define a flux variable v_r and a binary variable z_r, where $z_r=1$ if $v_r>0$; 0 otherwise. The set of metabolites is divided into two subsets, namely external (*Ext*) and internal (*Int*) metabolites. For internal metabolites, it is assumed that no accumulation or depletion is possible; therefore, the steady state condition holds. We denote s_{cr} the stoichiometric coefficient associated with metabolite c ($c=1,\ldots,C$) in reaction r ($r=1,\ldots,R$). External metabolites are typically considered as sources/sinks and (sometimes) as cofactors.

The (*K*,*d*)-shortest EFM approach is formulated below via mixed-integer linear programming (MILP). In essence, our method computes the K EFMs that (i) involve a certain reaction j and (ii) differ in at least d reactions. K, d and j are user-defined. The main difference with respect to our previous work (de Figueiredo *et al.*, 2009a) is in the use of d, which before was implicitly set to 1. We here allow $d>1$ to generate a more diverse set of EFMs. Variables, constraints and objective function are summarized and described below.

$$\min \sum_{r=1}^{R} Z_r$$

$$s.t.$$

$$\sum_{r=1}^{R} s_{cr} v_r = 0, c \in Int \tag{1}$$

$$v_r \geq z_r \text{ and } v_r \leq M z_r, r = 1, \ldots, R \tag{2}$$

$$z_\alpha + z_\beta \leq 1, (\alpha, \beta) \in B \tag{3}$$

$$z_j = 1 \tag{4}$$

$$\sum_{r=1}^{R} Z_r^k z_r \leq \sum_{r=1}^{R} (Z_r^k) - d, k = 1, ..., K-1$$
$$z_r = \{0,1\}, v_r \geq 0 \tag{5}$$

EFMs are steady state flux modes that satisfy a simplicity condition (Schuster *et al.*, 2000). This simplicity condition was termed the non-decomposability condition and essentially ensures that flux modes cannot be further reduced, i.e. they are the simplest possible. As introduced in de Figueiredo *et al.* (2009a), minimizing the number of active reactions in the solution (as we are doing here) guarantees that the non-decomposability condition is satisfied.

Equation (1) enforces steady state condition for internal metabolites. Equation (2) links reaction variables, namely if $z_r=1$ then $v_r>0$ and if $z_r=0$ then $v_r=0$; similarly if $v_r>0$ then $z_r=1$ and if $v_r=0$ then $z_r=0$. M is a large positive scalar that provides an upper bound for reaction fluxes. Equation (3) prevents forward and backward reactions from appearing together in the solution. Equation (4) forces reaction j to appear in the solution. As Equation (5) constitutes the main novelty of the (*K*,*d*)-shortest EFMs procedure, we explain below the meaning of this constraint in detail (see de Figueiredo *et al.*, 2009a; Rezola *et al.*, 2011).

Let Z_r^k be the binary solution associated with the k-shortest EFM ($k=1,\ldots,K-1$), where $Z_r^k=1$ if reaction r is active in the k-shortest EFM, 0 otherwise. The left side of Equation (5) determines the number of reactions in the new solution (z_r) that were active in the k-shortest EFM ($k=1,\ldots,K-1$), and the right side is the number of reactions that were active in the k-shortest EFM ($k=1,\ldots,K-1$) less d. The inequality states that the number of reactions repeating from the k-shortest EFM ($k=1,\ldots,K-1$) in the new solution should be smaller than the total number of active reactions in the k-shortest EFM ($k=1,\ldots,K-1$) minus d. If $d=1$, as in de Figueiredo *et al.* (2009a), Equation (5) ensures that, once we solve our model, the k-shortest EFM ($k=1,\ldots,K-1$) is prevented from appearing again in the new solution. If we fix $d>1$, our solution elimination constraints are more restrictive and prevent more solutions aside from the k-shortest EFM ($k=1,\ldots,K-1$) from appearing.

In summary, if $d=1$, we compute the K-shortest EFMs. However, if $d>1$, we start from the shortest but we will also reach long EFMs. This feature provides more diversity to computed EFMs, which is precisely the objective with this approach. Finally, various software tools are available to perform this task. We used IBM ILOG CPLEX®.

2.2 EFMs enrichment analysis

Assume that, based on gene expression data, reactions are grouped into three sets: highly, moderately and lowly expressed. We denote R_H, R_M and R_L the total number of highly, moderately and lowly expressed reactions in the network, respectively, and R_G the number of reactions whose reaction expression is known, namely $R_G=R_H+R_M+R_L$.

As noted above, EFMs are the simplest (steady state) flux modes. In other words, an EFM defines a subset of reactions whose flux is proportional with respect to a scalar quantity. Based on the above, we denote H_e, M_e and L_e the total number of highly, moderately and lowly expressed reactions in an EFM e ($e=1,\ldots,E$), respectively, and T_e the total number of active reactions with known discrete expression status in an EFM e, namely $T_e=H_e+M_e+L_e$.

2.2.1 Multivariate hypergeometric testing Given a set of E EFMs, we aim at finding a characteristic subset of EFMs for a given cell/tissue in certain (physiological) conditions, which are assumed to be represented by measured gene expression data. Clearly, such characteristic EFMs

should involve an elevated number of highly expressed reactions (H_e) and as few as possible lowly expressed reactions (L_e). To consider an EFM as characteristic, its associated gene expression data must be statistically significant. For this purpose, we need to ensure that the pair (H_e, L_e) in an EFM e of T_e reactions with known expression status can hardly arise by chance in the context of the whole network, which is defined by the tern (R_H, R_M, R_L), and therefore it is dependent on gene expression data. For this purpose, we pose the following hypothesis test:

H_0: EFM e is not characteristic, and therefore it is independent of gene expression data.
H_1: EFM e is characteristic, and therefore it is dependent on gene expression data.

Obtaining EFMs of length T_e with i highly and j lowly expressed reactions by chance is represented here as a random extraction of T_e balls without replacement in an urn. The probability mass function of a multivariate random variable describing the number of highly and lowly expressed reactions in an EFM (random extraction), $X=(i,j)$, follows a multivariate hypergeometric distribution with parameters R_G, R_H, R_L and T_e, as introduced in Equation (6).

$$P(X(R_G,R_H,R_L,T_e)=(i,j))=\frac{\binom{R_H}{i}\binom{R_L}{j}\binom{R_G-R_H-R_L}{T_e-i-j}}{\binom{R_G}{T_e}} \quad (6)$$

To decide whether (or not) the null hypothesis is rejected for a given EFM e of length T_e with H_e highly and L_e lowly expressed reactions, we determined the P-value (p_e) as the probability of obtaining the same or a better expression status by chance [Equation (7)]. Note here that an equal/better outcome than the pair (H_e, L_e) satisfies two conditions: (i) the number of lowly expressed reactions is less or equal than L_e; and (ii) the number of highly expressed reactions is greater or equal than H_e, whereas keeping unchanged the length of the EFM (T_e). These conditions require the use of one side hypothesis test.

$$p_e=\sum_{i=H_e}^{\min(R_H,T_e)}\sum_{\substack{j=0\\(i+j\le T_e)}}^{L_e}P(X(R_G,R_H,R_L,T_e)=(i,j)) \quad (7)$$

This rule excludes solutions with fewer highly expressed reactions. In addition, it seems functionally appropriate to establish that a worse solution involves more lowly expressed reactions, even with more highly expressed reactions.

Note here that P-values were computed using the '*BiasedUrn*' package in *R*. In this package, *dMFNCHypergeo* function returns the probability mass function for the multivariate Fisher's hypergeometric distribution shown in Equation (6).

2.2.2 Multiple hypothesis testing In single hypothesis testing, a result is called significant if the associated P-value is smaller than a significance level α (often $\alpha=0.05$ in the scientific literature). This determines the type I error rate, i.e. the probability of rejecting the null hypothesis when it is actually true. Extending this concept when several hypotheses are tested simultaneously (as done here) is studied in the field of multiple hypothesis testing. In this situation, the definition of an error measure according to the rate of false positives is more complex and refers to the probability of rejecting a null hypothesis when all (or $\hat{\pi}_0$) null hypotheses are actually true. For this purpose, we used the false discovery rate (FDR) approach presented in Storey and Tibshirani (2003), which maximizes the statistical power and maintains the conservative estimation of false positives:

$$\overline{FDR}(t)=\frac{\hat{\pi}_0\cdot m\cdot t}{R(t)} \quad (8)$$

where m is the total number of hypotheses tested, $R(t)$ is the number of P-values below t ($t\in[0,1]$) and $\hat{\pi}_0$ is the estimated rate of truly null hypothesis. This FDR approach assigns a q-value q_e to each hypothesis (similar to the P-value). The q-value is the FDR measure of significance and defines the minimum FDR that can be attained when calling a P-value (p_e) significant, as observed in Equation (9).

$$\hat{q}_e(p_e)=\min_{t\ge p_e}\overline{FDR}(t) \quad (9)$$

A more accurate estimation of $\hat{\pi}_0$ constitutes the main gain in statistical power with respect to control FDR (Benjamini and Hochberg, 1995), which assumes $\hat{\pi}_0=1$ by considering all hypotheses as truly null. Note here that we are posing one-sided test statistics, as reflected in our definition of P-value in Equation (7). For this scenario, a robust and conservative estimation of $\hat{\pi}_0$ is provided in Pounds and Cheng (2006), as shown below.

$$\begin{cases}\hat{\pi}_0=\min(1,2\bar{a})\\ \bar{a}=\frac{1}{m}\sum_{i=1}^{m}2\min(p_e,1-p_e)\end{cases} \quad (10)$$

A fundamental assumption in the method described above is that the joint distribution of the true null P-values must follow a uniform distribution on (0,1), Uniform (0,1) (Storey and Tibshirani, 2003). This property does not hold in our case, due to underlying discreteness in input data and dependencies among EFMs. When uniformity property in the joint distribution of true null P-values is violated, observed P-values must be (somehow) corrected to satisfy this property (Forner *et al.*, 2008). For this purpose, we first empirically verified this issue using bootstrapping, namely by calculating the P-value of each EFM in N different random permutations of the set of highly expressed reactions (R_H) and lowly expressed reactions (R_L), while maintaining R_G unchanged. We fixed $N=50$, as empirical experience shows that the null joint distribution behaves similarly for larger values of N. Then, we transformed empirical null P-values (p_i^e) into a Uniform (0,1), according to Equation (11).

$$p_i^{e,new}=\frac{\#\{p_j^{e,old}<p_i^{e,old}\}+\#\{p_j^{e,old}=p_i^{e,old}\}\cdot 0.5}{\#\{p_j^{e,old}\}},\ i=1,\ldots,N\cdot E \quad (11)$$

Then, after checking that $p_i^{e,new}$ follows a Uniform (0,1), the same transformation is done to observed P-values (p_e), as in Equation (12).

$$p_e^{new}=\frac{\#\{p_j^{e,old}<p_e^{old}\}+\#\{p_j^{e,old}=p_e^{old}\}\cdot 0.5}{\#\{p_j^{e,old}\}},\ e=1,\ldots,E \quad (12)$$

This procedure guarantees that the null distribution of transformed P-values (p_e^{new}) follows a Uniform (0,1), and therefore the mentioned FDR approach can be applied.

3 RESULTS

The (K,d)-shortest EFM algorithm is applied here to determine a subset of EFMs in human metabolism aiming at capturing a diverse and representative set of pathways. We integrated this subset of EFMs with expression data of 10 human tissues (liver, brain, heart, kidney, lung, pancreas, prostate, spleen, thymus and skeletal muscle) and determined their characteristic EFMs. Finally, the technical performance of our approach is analyzed and a biological discussion focused on liver is accomplished.

3.1 EFMs in human metabolic network

The human metabolic network presented in Duarte *et al.* (2007) comprises 2766 metabolites and 3742 reactions. Reversible reactions are split into two different reactions. Overall, we have 4892

non-negative fluxes. We assumed a general growth medium, as defined in Duarte *et al*. (2007). As our main objective is to find key biosynthetic and degradation routes for different nutrients, we neglected the balancing of highly present cofactors (see Supplementary Material S1), and therefore they were considered external metabolites. This is a common practice in pathway analysis that substantially simplifies the interpretation of EFMs and reduces computation time (de Figueiredo *et al*., 2009a; Rezola *et al*., 2011).

To obtain a diverse set of metabolic pathways from the human metabolic network, we used the (*K*,*d*)-shortest EFM procedure, as described in Methods section. To illustrate its performance, we computed five EFMs with $d=1$ and $d=5$, forcing the citrate synthase reaction to be involved in all EFMs (Fig. 1).

It can be observed that the (*K*,*d*)-shortest procedure obtains more varied EFMs than the standard *K*-shortest method, involving a higher number of reactions. Importantly, the (*K*,*d*)-shortest approach captures the tricarboxylic acid cycle among its first five solutions, whereas the *K*-shortest EFM method does not, which reflects a clear advantage. To fix an appropriate *d*-value, a sensitivity analysis can be found in Supplementary Material S2. Based on this analysis and empirical evidence, $d=5$ guarantees diversity in the computed set of EFMs.

We applied the (*K*,*d*)-shortest EFM method with $K=d=5$ for each reaction in Duarte *et al*. (2007). Therefore, we computed five EFMs per reaction, which would comprise a list of 24 460 EFMs. However, because of inconsistent reactions, infeasibilities and repeats, we obtained a final list of 5876 distinct EFMs from the human metabolic network. Details of EFMs can be found in Supplementary Material S1. For this task, we used two PCs with 2.00 GHz CPU and 12 GB of RAM during approximately 1 week. We observed that the lengths of the resulting EFMs are distributed between 2 and 52 reactions, which supports that our approach leads to both short and long EFMs.

To validate that our search procedure captures a diverse set of human metabolic pathways, we counted the number of HumanCyc pathways appearing in our subset of EFMs. HumanCyc is a well-known bioinformatics database that provides a repository of human metabolic pathways (Trupp *et al*., 2010). At the time when our analysis was conducted, HumanCyc included 282 biosynthesis, degradation and energy metabolism pathways. Some of these pathways involve reactions either not found or inconsistent (not able to carry flux) in Duarte *et al*. (2007). After filtering these unpredictable pathways, we obtained 180 feasible metabolic pathways. Our set of EFMs fully recovered 142 HumanCyc pathways. In particular, 2484 of our 5876 EFMs contained at least one entire HumanCyc pathway. This indicates that half of our EFMs (>50%) do not completely contain a known metabolic pathway and novel pathways may arise. In light of these results, we can conclude that our set of EFMs captures a high variety of metabolic pathways in human cells, and therefore it can be used similarly to a functional database where experimental information (such as 'omics' data) is integrated. This is precisely what we do in the next subsection.

3.2 Characteristic EFMs in human tissues

Given the set of EFMs determined in the previous subsection, we aim at selecting a characteristic subset of them for each of the 10 healthy human tissues analyzed. For this purpose, we used the reaction classification accomplished in Shlomi *et al*. (2008) based on gene expression data. In that work, a gene is considered highly (or lowly) expressed in a certain healthy tissue if it is substantially uniformly expressed (or non-expressed) in all GeneNote entries (Shmueli *et al*., 2003) and in HPRD (Mishra *et al*., 2006), otherwise it is considered moderately expressed. Based on this classification of genes and their Boolean rules for enzymes (Duarte *et al*., 2007), reactions are similarly classified into highly, moderately and lowly expressed in each tissue and then they are mapped into our set of EFMs. Details of this mapping for each EFM in each different tissue, which defines a particular subset of highly, moderately and lowly expressed reactions, can be found in Supplementary Material S1. We then determined the *P*-value (p_e) for each EFM based on multivariate hypergeometric distribution, as described in Methods section. The complete set of *P*-values for our set of human EFMs in each different tissue can be found in Supplementary Material S1.

In particular, the use of a multivariate hypergeometric distribution allows us to consider highly and lowly expressed reactions in a unique statistical score, and therefore our analysis is not limited to a single feature, as found in others approaches. In particular, in Schwartz *et al*. (2007), the standard hypergeometric distribution was used to determine the statistical significance of EFMs via BlastSets (Barriot *et al*., 2004). In that work, lowly and moderately expressed reactions are grouped, and therefore the enrichment of only highly expressed reactions in EFMs is evaluated. This, however, may lead to incorrect conclusions from the functional point of view, e.g. an EFM *e* with expression status ($H_e=5$, $M_e=0$, $L_e=5$) would be a better outcome than ($H_e=4$, $M_e=6$, $L_e=0$), which is counterintuitive. Our multivariate hypergeometric approach directly deals with this issue and allows us to treat differently lowly and moderately expressed reactions.

For illustration, Figures 2A and 2B show the number of lowly, moderately and highly expressed reactions in the top 100 EFMs in skeletal muscle tissue using the standard and multivariate hypergeometric distribution, respectively. The effect of the multivariate hypergeometric approach is visually observed in Figure 2, as the number of lowly expressed reactions (length of light grey bars) in the best ranked EFMs is substantially reduced, whereas the number of highly expressed reactions remains similar. This result shows that our approach provides a more biologically meaningful subset of significant EFMs with respect to the standard hypergeometric distribution.

Because we are testing several hypotheses simultaneously, *P*-values must be corrected. We applied the FDR approach presented in Storey and Tibshirani (2003), which determines a *q*-value (q_e) for each EFM. As noted in Methods section, before *q*-values calculation, *P*-values were transformed to satisfy that distribution of true null hypotheses is uniform, as assumed in FDR approach. As in other works (Gennarino *et al*., 2012; Menashe *et al*., 2012), we fixed an FDR of 20% for each tissue, i.e. we extracted a subset of characteristic EFMs whose FDR is <20% ($q_e \leq 0.2$). This implies that <20% of these EFMs may have appeared by chance, and therefore they are not really tissue specific. The list of *q*-values for each EFM in the different tissues can be found in Supplementary Material S1.

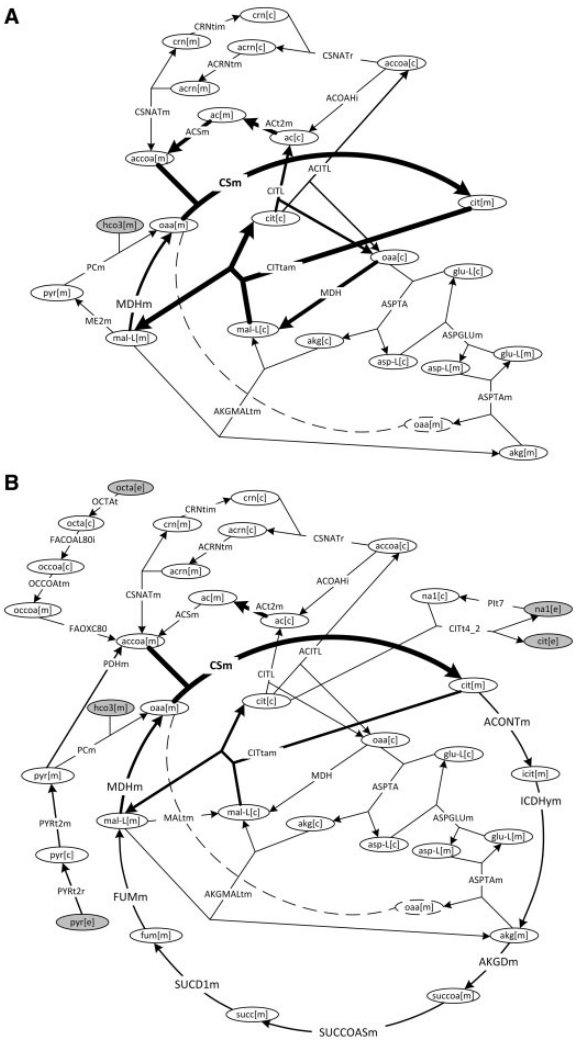

Fig. 1. The five shortest EFMs involving citrate synthase reaction (CSm) with (**A**) $d=1$ and (**B**) $d=5$. Ellipses represent metabolites and arrows represent reactions. Each metabolite is depicted with its corresponding compartment shown in brackets: [e], extracellular; [m], mitochondrial and [c], cytosol. Grey and white ellipses represent external and internal metabolites, respectively. Nomenclature for metabolites and reactions was taken from Duarte *et al.* (2007). Thickness of arrows (reactions) is proportional to the number of times they are involved in the five shortest EFMs. Individual EFMs involved are shown in Supplementary Material S2

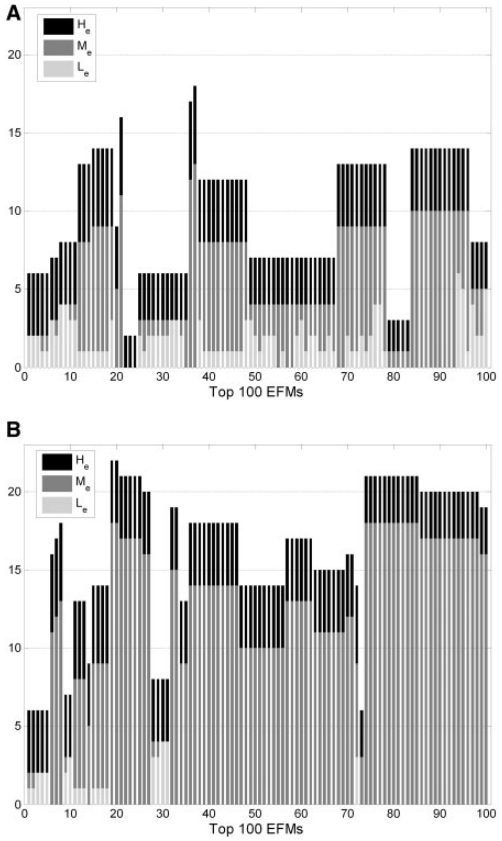

Fig. 2. Top 100 EFMs with the lowest *P*-values in skeletal muscle tissue using (**A**) standard hypergeometric distribution and (**B**) our multivariate hypergeometric distribution. The length of black, grey and light grey bars represents the number of highly (H_e), moderately (M_e) and lowly (L_e) expressed reactions in an EFM, respectively

Table 1 summarizes the number of characteristic EFMs obtained in each of our 10 analyzed tissues. It is important to note that we found characteristic EFMs in all the tissues, which validates (based on expression data) that our set of EFMs is more than a random selection of reactions and it is an appropriate concept to capture metabolic pathways in different conditions.

Table 1 shows a distinct number of characteristic EFMs among tissues. One may think that this number depends on their number of highly (R_H) and lowly (R_L) expressed reactions, which differs in each tissue. However, this effect was neutralized in our FDR procedure by selecting the null hypothesis so that the number of characteristic EFMs is independent from R_H and R_L. This was done by determining the mean number of characteristic EFMs obtained when R_H and R_L are randomly permuted a number of times (see Methods section). A clear example is observed in liver and heart, as we obtained more characteristic EFMs in heart with fewer R_H and more R_L. Instead, variations in Table 1 arise from a different level of functional association of highly and lowly expressed reactions in each tissue, namely a greater connectivity of highly (lowly) expressed reactions in a tissue increases (decreases) its number of characteristic EFMs. Based on our set of EFMs, we analyzed this issue in heart and lung, as they have a very similar R_H and R_L and a substantially different number of characteristic EFMs (728 and 10, respectively). Results emerging from this analysis support this hypothesis (see Supplementary Material S2), which reflect (i) metabolism in lung is less specific than in other tissues or (ii) our set of EFMs does not capture its key pathways. This issue requires further research, mainly by computing a larger set of EFMs.

Characteristic EFMs allow us to capture tissue-specific metabolic functions. For example, in Supplementary Material S2 we discuss two of our EFMs involved in nitrogen metabolism in the context of liver and brain and show their differences in the use of urea cycle. A global comparison of characteristic EFMs among tissues threw interesting insights (see Supplementary Material S2). In particular, we found that many of the characteristic

Table 1. Number of highly (R_H) and lowly (R_L) expressed reactions and number of characteristic EFMs in each human tissue

Tissue	R_H	R_L	Number of characteristic EFMs (FDR < 20%)
Liver	423	184	394
Brain	199	408	292
Heart	206	379	728
Kidney	313	221	113
Lung	187	386	10
Pancreas	156	356	86
Prostate	56	310	387
Spleen	48	432	536
Thymus	29	368	367
Muscle	168	395	71

EFMs in skeletal muscle were also present in heart tissue, while not the contrary. The same was found between muscle and brain. Two clusters with a similar set of characteristic EFMs were observed: (i) pancreas and kidney; (ii) thymus, spleen and prostate. Liver highly differs from the rest of tissues. Finally, we also determined tissue-specific metabolites, i.e. metabolites active in only one of the tissues, based on their resulting characteristic EFMs. It was found that 7 of 10 tissues do have a significant number of specific metabolites (see Supplementary Material S3), some of which are known to have a role in those tissues. For example, phosphatidylinositol-3,4-bisphosphate is a secondary messenger in neurons (Tanaka *et al*., 1999), whereas 1D-myo-inositol 1,4-bisphosphate is a degradation product of a Ca^{2+} channel regulator (c.f. Trinquet *et al*., 2006).

3.3 Characteristic EFMs in liver

A detailed analysis of the characteristic EFMs and metabolites in liver showed that the main key functionalities of this organ were captured. The majority of the EFMs take part in steroid metabolism (178 EFMs), fatty acid β-oxidation (75 EFMs), one-carbon metabolism (27 EFMs) and urea cycle (15 EFMs). For a more detailed classification of them, see Supplementary Material S1.

Approximately two thirds of the EFMs associated with the steroid metabolism are involved in bile acid synthesis, which is one of the main functions of hepatocytes (Monte, 2009; Gille *et al*., 2010). Given that our subset of EFMs mainly includes short EFMs, the precursor found for the synthesis of bile acids is vitamin D_3 instead of cholesterol (Monte, 2009) or fatty acid (Gille *et al*., 2010). However, these EFMs can be combined with the EFMs that produce vitamin D_3 from acetate, which is obtained in fatty acid β-oxidation, and reach a similar solution found by Gille *et al*. (2010), for the synthesis of bile acids. The other one-third of the EFMs associated with steroid metabolism carry out the synthesis of cholesterol, 7-α,24(S)-dihydroxycholesterol, 7-α,27-dihydroxycholesterol of the prohormone vitamin D_3 and its derivate 25-hydroxyvitamin D_3.

The EFMs associated with the urea cycle carry out the degradation of various amino acids: glutamine, alanine, ornithine and asparagine. On the other hand, the EFMs associated with the one-carbon metabolism are mainly involved in the catabolism of histidine to glutamate or glutamine. However, some of these EFMs also capture the conversion of glutamate to 4-aminobutanoate, though this mainly occurs in the brain (Jakobs *et al*., 1993). Finally, 22 substrate cycles were found among the characteristic EFMs of liver.

4 DISCUSSION

In this work, a new modeling framework based on EFMs is introduced to provide a metabolic context to gene expression data. To this end, we first determined a diverse subset of EFMs in the human genome-scale metabolic network. We are aware that our subset of EFMs is not complete; however, it contains a wide number of reported metabolic pathways in HumanCyc database and previously uncharacterized metabolic pathways. Therefore, our subset of EFMs constitutes a rich template of metabolic pathways available to the scientific community where experimental 'omics' data can be mapped for functional interpretation.

To compute our subset of EFMs, we introduced a generalized version of the *K*-shortest EFM approach presented in de Figueiredo *et al*. (2009a). Despite the good performance of the (*K*,*d*)-shortest EFMs procedure, our subset of EFMs can be further improved in size and quality to obtain a better picture of human metabolic processes. For this purpose, other algorithms should be explored in the future (Kaleta *et al*., 2009b; Chan and Ji, 2011).

We also determined a characteristic subset of EFMs in different tissues based on expression data. To guarantee statistical significance, we used a multivariate hypergeometric distribution, which allows us to consider highly expressed and lowly expressed reactions in a unique statistical score. We demonstrated that this test leads to more biologically meaningful EFMs than the standard hypergeometric approach, as illustrated for skeletal muscle.

Our approach distinguishes between highly, moderately and lowly expressed genes based on absolute levels of microarray data. Thus, we need to define two different thresholds: one for highly and one for lowly expressed genes. In this article, we directly took data from Shlomi *et al*. (2008). However, for different microarray datasets, threshold choice can be done using recently introduced approaches that allow a robust estimation as to whether (or not) a gene is expressed in a particular sample (McCall *et al*., 2011; Shi *et al*., 2010). These techniques open the application of our approach for a wide number of datasets.

In this article, our approach was applied to extract active characteristic EFMs. However, finding repressed EFMs can be easily done by slightly changing the definition of *P*-value. In addition, we can use our approach to obtain differential EFMs between two conditions based on fold changes. However, several theoretical issues arise, e.g. genes consistently highly expressed may have a small fold change. To overcome this issue, we propose to combine both approaches and compare (i) characteristic EFMs for each condition based on absolute expression levels and (ii) differential EFMs between two conditions based on fold-change data.

Our work complements effort in the last years to generate tissue-specific metabolic networks (Shlomi *et al*., 2008;

Gille *et al*., 2010; Jerby *et al*., 2010). We found strong variations among tissues, observing a high number of characteristic EFMs in several tissues (heart, liver, prostate, spleen, thymus, brain and kidney), whereas a limited number of them in others, particularly in lung. Based on characteristic EFMs, we captured tissue-specific functions and metabolites. Interestingly, with this analysis we showed that our set of EFMs allows us to more accurately predict global metabolic properties, mainly because they are determined from genome-scale networks, which provide a more holistic view of metabolism than canonical (HumanCyc) pathways.

To validate our approach, we analyzed the resulting characteristic EFMs in liver. We found that key metabolic roles of liver, like steroid metabolism and fatty acid β-oxidation, were correctly captured. Other characteristic EFMs are associated with pathways that have been described in literature, such as the glucuronidation of bilirubin (Hauser *et al*., 1984), the methylation of phosphatidylethanolamine to phosphatidylcholine (Vance and Ridgway, 1988), the methylation of NAD+, which plays a role in hepatocyte growth (Hoshino *et al*., 1984), or the tyrosine oxidation to acetoacetate (Knox and LeMay-Knox, 1951). This shows the predictive power of our approach. Other important characteristic pathways of liver, such as pathways for lipoprotein metabolism, were not determined because of limitations in the metabolic network used (Duarte *et al*., 2007), as recently highlighted in Gille *et al*. (2010).

Finally, our work constitutes an attempt to integrate experimental data into EFMs and elucidate their physiological relevance, an open challenge in systems biology. We believe that this effort will help to extend the application of EFMs to different questions in biotechnology and health, where metabolism does play an important role.

ACKNOWLEDGEMENTS

We would like to thank Prof. Sebastian Böcker for the valuable discussion about the (*K*,*d*)-shortest EFM approach and anonymous reviewers for helpful comments and suggestions that improved the original manuscript.

Funding: Funds were received from Asociación de Amigos de la Universidad de Navarra [to A.R.]; the Basque Government [to J.P.]; and the German Ministry of Education and Research (BMBF) within the Virtual Liver [0315758 to L.F.F.].

Conflict of Interest: none declared.

REFERENCES

Barriot,R. *et al*. (2004) New strategy for the representation and the integration of biomolecular knowledge at a cellular scale. *Nucleic Acids Res*., **32**, 3581–3589.

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Statist. Soc. B*, **57**, 289–300.

Brazma,A. (2003) ArrayExpress-a public repository for microarray gene expression data at the EBI. *Nucleic Acids Res*., **31**, 68–71.

Chan,S.J. and Ji,P. (2011) Decomposing flux distributions into elementary flux modes in genome-scale metabolic networks. *Bioinformatics*, **27**, 2256–2262.

Curtis,R.K. *et al*. (2005) Pathways to the analysis of microarray data. *Trends Biotechnol*., **23**, 429–435.

Dahlquist,K.D. *et al*. (2002) GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways. *Nat. Genet*., **31**, 19–20.

Duarte,N.C. *et al*. (2007) Global reconstruction of the human metabolic network based on genomic and bibliomic data. *Proc. Natl Acad. Sci. USA*, **104**, 1777–1782.

Edgar,R. (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res*., **30**, 207–210.

De Figueiredo,L.F. *et al*. (2009a) Computing the shortest elementary flux modes in genome-scale metabolic networks. *Bioinformatics*, **25**, 3158–3165.

De Figueiredo,L.F. *et al*. (2009b) Response to comment on "Can sugars be produced from fatty acids? A test case for pathway analysis tools". *Bioinformatics*, **25**, 3330–3331.

Forner,K. *et al*. (2008) Universal false discovery rate estimation methodology for genome-wide association studies. *Hum. Hered*., **65**, 183–194.

Gennarino,V.A. *et al*. (2012) Identification of microRNA-regulated gene networks by expression analysis of target genes. *Genome Res*., **22**, 1163–1172.

Ghazalpour,A. *et al*. (2005) Genomic analysis of metabolic pathway gene expression in mice. *Genome Biol*., **6**, R59.

Gille,C. *et al*. (2010) HepatoNet1: a comprehensive metabolic reconstruction of the human hepatocyte for the analysis of liver physiology. *Mol. Syst. Biol*., **6**, 411.

Goffard,N. and Weiller,G. (2007) PathExpress: a web-based tool to identify relevant pathways in gene expression data. *Nucleic Acids Res*., **35**, W176–W181.

Hauser,S. *et al*. (1984) Subcellular distribution and regulation of hepatic bilirubin UDP- glucuronyltransferase. *J. Biol. Chem*., **259**, 4527–4533.

Hoshino,J. *et al*. (1984) Nicotinamide methylation and its relation to NAD synthesis in rat liver tissue culture: Biochemical basis for the physiological activities of 1-methylnicotinamide. *Biochim. Biophys. Acta*, **801**, 250–258.

Jakobs,C. *et al*. (1993) Inherited disorders of GABA metabolism. *J. Inherit. Metab. Dis*., **16**, 704–715.

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

Kaleta,C. *et al*. (2009a) Can the whole be less than the sum of its parts? Pathway analysis in genome-scale metabolic networks using elementary flux patterns. *Genome Res*., **19**, 1872–1883.

Kaleta,C. *et al*. (2009b) EFMEvolver: Computing elementary flux modes in genome-scale metabolic networks. In: Grosse,I. *et al*. (ed.) *P. 14th German Conf. Bioinformatics*. Gesellschaft f. Informatik e.V., Halle, pp. 180–190.

Knox,W.E. and LeMay-Knox,M. (1951) The oxidation in liver of l-tyrosine to acetoacetate through p-hydroxyphenylpyruvate and homogentisic acid. *Biochem. J*., **49**, 686.

Konradi,C. *et al*. (2004) Molecular evidence for mitochondrial dysfunction in bipolar disorder. *Arch. Gen. Psychiat*., **61**, 300–308.

McCall,M.N. *et al*. (2011) The Gene Expression Barcode: leveraging public data repositories to begin cataloging the human and murine transcriptomes. *Nucleic Acids Res*., **39**, D1011–D1015.

Menashe,I. *et al*. (2012) Large-scale pathway-based analysis of bladder cancer genome-wide association data from five studies of European background. *PLoS One*, **7**, e29396.

Mishra,G.R. *et al*. (2006) Human protein reference database–2006 update. *Nucleic Acids Res*., **34**, D411–D414.

Mlecnik,B. *et al*. (2005) PathwayExplorer: web service for visualizing high-throughput expression data on biological pathways. *Nucleic Acids Res*., **33**, W633–W637.

Monte,M.J. (2009) Bile acids: Chemistry, physiology, and pathophysiology. *World J. Gastroentero*., **15**, 804.

Naylor,T.L. *et al*. (2005) High resolution genomic analysis of sporadic breast cancer using array-based comparative genomic hybridization. *Breast Cancer Res*., **7**, R1186–R1198.

Pey,J. *et al*. (2011) Path finding methods accounting for stoichiometry in metabolic networks. *Genome Biol*., **12**, R49.

Pfeiffer,T. *et al*. (1999) METATOOL: for studying metabolic networks. *Bioinformatics*, **15**, 251–257.

Pounds,S. and Cheng,C. (2006) Robust estimation of the false discovery rate. *Bioinformatics*, **22**, 1979–1987.

Rezola,A. *et al*. (2011) Exploring metabolic pathways in genome-scale networks via generating flux modes. *Bioinformatics*, **27**, 534–540.

Schuster,S. *et al*. (2000) A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks. *Nat. Biotechnol*., **18**, 326–332.

Schuster,S. and Hilgetag,C. (1994) On elementary flux modes in biochemical reaction systems at steady state. *J.f Biol. Syst*., **2**, 165–182.

Schwartz,J.-M. *et al.* (2007) Observing metabolic functions at the genome scale. *Genome Biol.*, **8**, R123.
Schwarz,R. *et al.* (2005) YANA - a software tool for analyzing flux modes, gene-expression and enzyme activities. *BMC bioinformatics*, **6**, 135.
Shi,W. *et al.* (2010) Estimating the proportion of microarray probes expressed in an RNA sample. *Nucleic Acids Res.*, **38**, 2168–2176.
Shlomi,T. *et al.* (2008) Network-based prediction of human tissue-specific metabolism. *Nat. Biotechnol.*, **26**, 1003–1010.
Shmueli,O. *et al.* (2003) GeneNote: whole genome expression profiles in normal human tissues. *C. R. Biol.*, **326**, 1067–1072.
Storey,J.D. and Tibshirani,R. (2003) Statistical significance for genomewide studies. *Proc. Natl Acad. Sci. USA*, **100**, 9440–945.
Subramanian,A. *et al.* (2005) Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA*, **102**, 15545–15550.
Tanaka,K. *et al.* (1999) Evidence That a Phosphatidylinositol 3,4,5-Trisphosphate-binding Protein Can Function in Nucleus. *J. Biol. Chem.*, **274**, 3919–3922.
Terzer,M. and Stelling,J. (2008) Large-scale computation of elementary flux modes with bit pattern trees. *Bioinformatics*, **24**, 2229–2235.
Trinquet,E. *et al.* (2006) d-myo-Inositol 1-phosphate as a surrogate of d-myo-inositol 1,4,5-tris phosphate to monitor G protein-coupled receptor activation. *Anal. Biochem.*, **358**, 126–135.
Trupp,M. *et al.* (2010) Beyond the genome (BTG) is a (PGDB) pathway genome database: HumanCyc. *Genome Biol.*, **11**, O12.
Vance,D.E. and Ridgway,N.D. (1988) The methylation of phosphatidylethanolamine. *Prog. Lipid Res.*, **27**, 61–79.
Zien,A. *et al.* (2000) Analysis of gene expression data with pathway scores. *P. Int. C. Intelligent Syst. Mol. Biol. (ISMB)*, **8**, 407–417.