BIOINFORMATICS ORIGINAL PAPER

Vol. 30 no. 7 2014, pages 975–980
doi:10.1093/bioinformatics/btt653

Systems biology

Advance Access publication November 21, 2013

Refining carbon flux paths using atomic trace data

Jon Pey[1], Francisco J. Planes[1,*] and John E. Beasley[2,*]

[1]CEIT and TECNUN, University of Navarra, Manuel de Lardizabal 15, 20018 San Sebastian, Spain and [2]Mathematical Sciences, Brunel University, Kingston Lane, UB8 3PH, Uxbridge, UK

Associate Editor: Prof. Mario Albrecht

ABSTRACT

Motivation: Pathway analysis tools are a powerful strategy to analyze ‘omics’ data in the field of systems biology. From a metabolic perspective, several pathway definitions can be found in the literature, each one appropriate for a particular study. Recently, a novel pathway concept termed carbon flux paths (CFPs) was introduced and benchmarked against existing approaches, showing a clear advantage for finding linear pathways from a given source to target metabolite.

CFPs are simple paths in a metabolite–metabolite graph that satisfy typical constraints in stoichiometric models: mass balancing and thermodynamics (irreversibility). In addition, CFPs guarantee carbon exchange in each of their intermediate steps, but not between the source and the target metabolites and consequently false positive solutions may arise. These pathways often lack biological interest, particularly when studying biosynthetic or degradation routes of a metabolite. To overcome this issue, we amend the formulation in CFP, so as to account for atomic fate information. This approach is termed atomic CFP (aCFP).

Results: By means of a side-by-side comparison in a medium scale metabolic network in *Escherichia Coli*, we show that aCFP provides more biologically relevant pathways than CFP, because canonical pathways are more easily recovered, which reflects the benefits of removing false positives. In addition, we demonstrate that aCFP can be successfully applied to genome-scale metabolic networks. As the quality of genome-scale atomic reconstruction is improved, methods such as the one presented here will undoubtedly be of value to interpret ‘omics’ data.

Contact: fplanes@ceit.es or John.Beasley@brunel.ac.uk

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

Received on June 21, 2013; revised on November 5, 2013; accepted on November 6, 2013

1 INTRODUCTION

Systems biology aims at modeling a wide assortment of biological phenomena from a holistic molecular perspective (Kitano, 2002). Analysis of high-throughput ‘omics’ data is typically intricate and requires sophisticated techniques assisting their interpretation and understanding. For this purpose, pathway analysis tools constitute a powerful strategy, as there is strong experimental evidence supporting the view that many cellular processes are organized into pathways (Stelling *et al*., 2002). In particular, biochemistry textbooks detail a wide list of well-known (canonical) metabolic pathways, which can now be computationally accessed from different databases and analyzed in different scenarios (Kanehisa *et al*., 2012; Keseler *et al*., 2009).

The outbreak of genome-scale metabolic networks has evidenced that it is not possible to represent all the metabolic phenotypes using a small set of canonical pathways. Intending to increase the number and diversity of pathways, a number of mathematical pathway definitions can be found in the literature (Planes and Beasley, 2008). For each scenario under study, selecting the appropriate definition is crucial so as to obtain relevant insights (Faust *et al*., 2009; de Figueiredo *et al*., 2009a). Defining a pathway concept representing too many biological mechanisms may be counterproductive due to the underlying computational demand. In this light, two sides of the same coin are the pathway definitions based on graph theory (Arita, 2000) and elementary flux modes (EFMs) (Schuster *et al*., 2000). Although the former is calculated based on classical graph theory (Dijkstra, 1959), the latter requires less efficient algebraic techniques (de Figueiredo *et al*., 2009b; Kamp and Schuster, 2006; Rezola *et al*., 2013; Terzer and Stelling, 2008). However, EFMs fulfill additional biological constraints, e.g. mass balance, leading to more biologically relevant results. It is worth mentioning that, in addition to EFMs, we can find in the literature similar pathway definitions, such as generating flux modes (Rezola *et al*., 2011) or extreme pathways (Price *et al*., 2003). Furthermore, additional examples of metabolic pathway definitions based on graph theory can also be found (Lim and Wong, 2012).

Recently, carbon flux paths (CFPs) (Pey *et al*., 2011) were introduced so as to account for the advantages of both definitions. This approach searches for paths in a graph of metabolites, where arcs represent an effective carbon exchange in at least one enzyme-catalyzed reaction. In addition, as in the case of EFMs, additional biophysical constraints are imposed, namely mass balance and thermodynamic restrictions. CFPs approach is an effective tool to determine linear pathways from a given source to a target metabolite, as it overcomes combinatorial issues found in EFMs approach. CFPs also take advantage of the inherent versatility of Integer Linear Programming (ILP) to integrate gene expression and protein abundance data as well as metabolomics data (Pey *et al*., 2013).

Despite the potential and versatility of CFPs, there is still room for improvement. In particular, while the underlying graph of metabolites guarantees carbon exchange in each step of the path, this is not always achieved between the source and target metabolites and false-positive solutions may arise. As discussed in Pey *et al*. (2011), these solutions lack biological significance, especially when studying biosynthetic or degradation mechanisms.

*To whom correspondence should be addressed.

In order to overcome this issue, the CFP formulation is here extended to ensure that a fixed number of carbon atoms from the source reach the target metabolite, in line with Pitkänen *et al.* (2009) and Heath *et al.* (2010). To do so, we integrate atomic tracing data for reactions into the model. In other words, for a biochemical reaction we require precise information indicating in which position in a product each carbon atom from a substrate ends. This information is growing day by day and is currently available in different reconstructions. These reconstructions can be manually curated (Suthers *et al.*, 2007) or computed by automatic methods (Mu *et al.*, 2007; Ravikirthi *et al.*, 2011). The use of *in silico* (automatically) generated databases entails sacrificing accuracy, but increases the number of reactions that can be considered. Based on this information, we show that the atomic extension of CFP presented here (aCFP) provides more biologically relevant pathways than the original CFP definition.

2 METHODS

For a metabolic network under consideration, we denote R and C the set of reactions and compounds, respectively. We define S_{cr} as the stoichiometric coefficient associated with metabolite c in reaction r.

The constraints corresponding to the metabolic fluxes (v_r) remain as in Pey *et al.*, (2011). Metabolites are divided into two sets: internal (I) and external (E). For metabolites in I, steady-state condition is applied, namely Equation (1). Metabolites in E are not necessarily balanced, but consumption is only allowed to the subset of metabolites in the growth media (E_m), see Equation (2). We also include the binary variable z_r, being equal to one if reaction r is active and zero otherwise. This relationship is represented in Equation (3) where N is a large scalar representing the maximum flux value. Finally, note that each reversible reaction is split into two irreversible reactions where B = {(λ,μ)| reaction λ and reaction μ are the reverse of each other}. Equation (4) prevents a solution with a reversible reaction activated in both directions.

$$\sum_{r \in R} S_{cr} v_r = 0, \forall c \in I \tag{1}$$

$$\sum_{r \in R} S_{cr} v_r \geq 0, \forall c \in E, c \notin E_m \tag{2}$$

$$z_r \leq v_r \leq N z_r, \forall r \in R \tag{3}$$

$$z_\lambda + z_\mu \leq 1, \forall (\lambda, \mu) \in B \tag{4}$$

We briefly introduce a subset of the path constraints included in Pey *et al.* (2011). As we shall see below, constraints removed from the previous formulation are not required when atomic tracing data is incorporated into the model.

We have binary variable u_{ij}, being equal to one if the arc linking the metabolites i and j is present in the obtained pathway and zero otherwise. Note here that we only include in the graph arcs between those metabolites exchanging carbon atom(s) in at least one reaction in R.

Equation (5) ensures that one arc leaves the source (α) and one arc reaches the target (β) metabolites. Similarly, Equation (6) guarantees that no arcs reach and leave α and β. This equation is not applied when looking for a cyclic pathway, namely $\alpha = \beta$. Equation (7) prevents revisiting a metabolite so branched solutions are not allowed in the calculated pathway.

$$\sum_{j \in C} u_{\alpha j} = \sum_{i \in C} u_{i \beta} = 1 \tag{5}$$

$$\sum_{i \in C} u_{i \alpha} = \sum_{j \in C} u_{\beta j} = 0 \tag{6}$$

$$\sum_{i \in C} u_{ik} \leq 1, \forall k \in C \tag{7}$$

In Pey *et al.* (2011), we forced carbon exchange between intermediate metabolites in the path, however, carbon flux between α and β was not ensured. We overcome this issue by going one level of complexity further, namely modeling the fate of carbon atoms. The information describing the movement of a carbon atom is represented by atom mapping matrices (AMM) (Zupke and Stephanopoulos, 1994). These matrices have been manually annotated for medium scale metabolic networks comprising hundreds of reactions (Suthers *et al.*, 2007). There are also algorithms calculating the fate of an atom within a reaction in genome-scale metabolic networks (Latendresse *et al.*, 2012; Ravikirthi *et al.*, 2011). However, further research is required to improve the accuracy of these algorithms.

Below we show how to extend CFPs so as to assure effective carbon exchange between α and β by including carbon fate information into the model. Note how a graph of metabolites exchanging carbon atoms (Fig. 1B) is less informative than a graph where nodes are carbon atoms (Fig. 1C). We describe below this carbon atom graph.

Let us define A as the set of carbon atoms in the metabolic network under study and A_c the set containing all the carbon atoms of a particular metabolite c. Figure 1D includes the sets described above for the Pyruvate kinase (PYK) reaction except for adenosine triphosphate (ATP) and adenine dinucleotide phosphate (ADP).

In Pey *et al.* (2011), d_{ijr} was defined to be equal to one if at least one carbon atom from metabolite j reaches metabolite j in reaction r. Similarly to d_{ijr}, we introduce here the coefficient e_{lmr}, being equal to one if carbon atom l reaches m in reaction r and zero otherwise. Both coefficients are closely related, in particular, if $l \in A_i$, $m \in A_j$ and $e_{lmr} = 1$ then $d_{ijr} = 1$, whereas the opposite is not always true. Therefore, if $\exists\, r \in R$ / $e_{lmr} = 1$, we define an arc between l and m.

In analogy with u_{ij}, we introduce the binary variable w_{lm}, which is equal to one if the atomic arc connecting carbon atoms l and m is active in the obtained path and zero otherwise.

Equation (8) ensures that the number of incoming arcs is equal to the number of outgoing arcs for each carbon atom in A except for those belonging to the source and target metabolites, i.e. α and β. Equation (8) encompasses original formulation in Pey *et al.* (2011), where this constraint was applied in the metabolic graph.

In addition, we relate u_{ij} and w_{lm} through Equation (9). In essence, this constraint ensures that if w_{lm} is active in the path and $l \in A_i$ and $m \in A_j$, then u_{ij} is also in the path. Notice that Equation (9) links in one equality constraint the carbon atom level with the metabolite level. Q fixes the number of atoms reaching the target from the source metabolite, in consequence Q is an integer >0. The activation of a metabolic arc (u_{ij}) implies the activation of Q of its underlying atomic arcs (w_{lm}). Therefore, with Equations (5)–(9), we guarantee in each intermediate metabolic step that Q carbons from the source are transported.

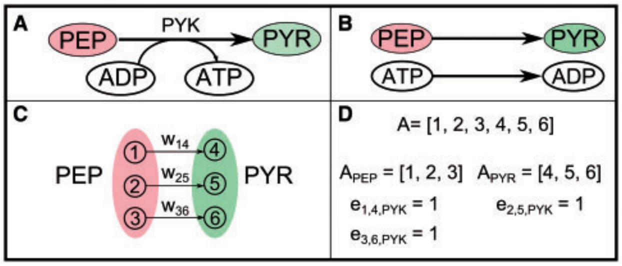

Fig. 1. (**A**) *Pyruvate kinase* (PYK, EC 2.7.1.40); (**B**) corresponding carbon arcs; (**C**) carbon graph; (**D**) corresponding atomic coefficients. For the sake of simplicity the atomic fate of ATP/ADP was not included in Figure 1C and D

It is important to notice how constraints (5)–(9) allows us to determine a metabolic path that ensures carbon exchange from source α to target β.

By means of Equation (10), we impose the constraint that if an atomic arc turns out to be active in the solution, at least one metabolic reaction containing it should be active. Note that, through this equation, we have connected the path and stoichiometric layers.

Equation (11) is the objective function that guides the optimization process. As in Pey *et al.* (2011), we will search for the shortest path. In particular, we will minimize the number of active arcs appearing in the pathway.

$$\sum_{l\in A} w_{ln} = \sum_{m\in A} w_{nm}, \forall n \in A, n \notin \{A_\alpha \cup A_\beta\} \tag{8}$$

$$\sum_{l\in A_i} \sum_{m\in A_j} w_{lm} = Qu_{ij}, \forall i \in C; \forall j \in C \tag{9}$$

$$\sum_{r\in R,\, e_{lmr}=1} z_r \geq w_{lm}, \forall l \in A; \forall m \in A \tag{10}$$

$$\min \sum_{i\in C} \sum_{j\in C} u_{ij} \tag{11}$$

Integer programing allows us to enumerate a predefined number of paths (K) in an ascending length order by including Equation (12), where U_{ij}^k is equal to u_{ij} in the kth solution. Here we enumerate paths at the metabolic level (u_{ij}) rather than at the atomic level (w_{lm}). In particular, note how a particular metabolic path may comprise many underlying atomic pathways. For example, in Figure 1C there are three atomic pathways between Phosphoenolpyruvic acid (PEP) and Pyruvate (PYR), but only one metabolic path. A deeper discussion highlighting the differences can be found in Section 3. Aiming at obtaining each metabolic pathway only once, we enumerate them at the metabolic level, namely by means of u_{ij}.

$$\sum_{i\in C} \sum_{j\in C,\, j\neq i} U_{ij}^k u_{ij} \leq \sum_{i\in C} \sum_{j\in C,\, j\neq i} U_{ij}^k - 1, k = 1, \ldots, K-1 \tag{12}$$

Summarizing, Equations (1)–(12) form the atomic extension to the CFP approach, namely aCFP.

3 RESULTS

3.1 Recovering canonical pathways with aCFP

As for appraising the performance of the methodology introduced here, we perform a side-by-side comparison between aCFP and CFP. For completeness, we included the results coming from a pure atomic graph, i.e. path resulting from the carbon graph (Fig. 1C). We referred to this last group as atomic paths (AP) approach. In order to analyze the performance of AP approach, we modified Equation (12) to enumerate at the atomic level rather than at the metabolic level; i.e. based on w_{lm} instead of u_{ij}.

We used the imPS1485 model, which constitutes one of the more extensive metabolic networks including manually curated AMMs (Suthers *et al.*, 2007). In particular, this network involves several metabolic pathways in *Escherichia coli*, comprising 317 reactions and 270 metabolites after including exchanges and splitting reversible reactions. To build the metabolite graph for CFP as defined in Pey *et al.* (2011), we used AMMs in Suthers *et al.* (2007), as highlighted in methods section ($l\in A_i$, $m\in A_j$ and $e_{lmr}=1$ then $d_{ijr}=1$). Results shown below were obtained using the general growth media defined in the imPS1485 model.

The performance of aCFP was evaluated in the light of 18 canonical metabolic pathways. These pathways are a subset of the 40 pathways used in the validation of the CFP approach (Pey *et al.*, 2011), particularly those included in the imPS1485 metabolic model. Two reasons led us to discard a canonical pathway from the original set of 40 pathways: (i) absence of some of its intermediate metabolites and/or reactions in the imPS1485 model; (ii) inability to flow in steady state. In order to maximize the number of pathways considered, we performed some slight modifications. For instance, some reactions are lumped in imPS1485 and, in consequence, some intermediate steps of the canonical pathway must be lumped so as to match with the corresponding pathway in imPS1485. So, 18 is the maximum number we can consider given the limitations in imPS1485. The full set of pathways can be found in the Supplementary Material.

Overall, each one of these 18 pathways is defined by a sequence of several metabolites, so we assume that the pathway is recovered when the calculated solution and the canonical pathway are the same in terms of this sequence of intermediate metabolites. Note also that the set of constraints is particularized for each canonical pathway, defined by a distinct pair of source/target metabolites, i.e. α and β.

As discussed in the Section 2, ILP allows us to enumerate pathways in increasing length order. It is typical that solutions with the same length exist; these will be enumerated in an arbitrary order depending on several factors, e.g. ILP solver, the computer, the order of the constraints. Therefore, the side-by-side comparison should be carried out in terms other than based on the position in which the canonical pathway is recovered (here by position we mean the solution at which the pathway is found, e.g. position 10 means that the pathway is found at the 10th solution). In particular, we perform the side-by-side comparison using a position interval in which the solutions of equal length to the canonical pathway are computed. For that we introduce p^- and p^+, which represent the position of the first and last computed solution with equal length to the canonical pathway. The reader should note here that the lower the values of p^- and p^+ and the narrower the interval $[p^-, p^+]$, the better the method is at recovering canonical pathways. For the sake of clarity, we represent p^- and p^+ by means of a bar graph. In addition, the position in which the canonical pathway was recovered is represented with a dot. Note that if p^- is equal to p^+, the bar is reduced to a point which must correspond to that in which the canonical pathway under request was recovered, i.e. indicated with a dot.

We include three bars per pathway, namely one per approach: aCFP white bar, CFP dark gray bar and AP light gray bar. When recovering pathways with aCFP we impose that, at least, half of the carbon atoms from the source reach the target when the source has more carbons than the target or *vice versa* when the target has more carbons, except in the case of the pentose phosphate pathway in which only two carbons from the source reaches the target by means of the canonical pathway. Finally, this comparison is carried out for each of the 18 pathways and is presented in Figure 2.

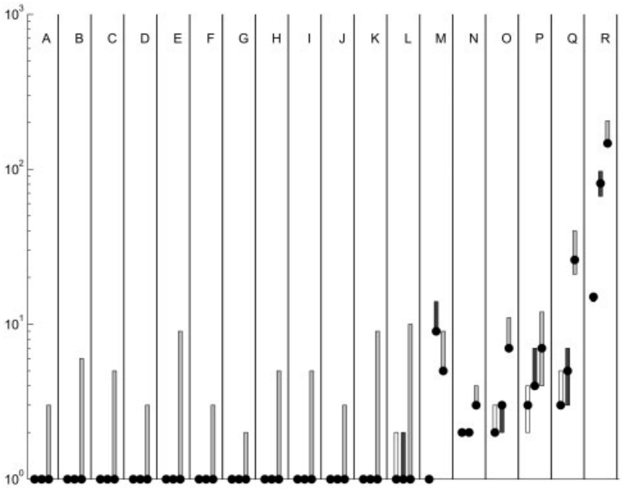

Fig. 2. A side-by-side comparison between aCFP, CFP and AP presented in terms of p^- and p^+, which represent the position of the first and last computed solution with equal length to the canonical pathway. Bars extend from p^- to p^+. The first (white) bar corresponds to aCFP, the second (dark gray) bars to CFP and the third (light gray) bars to AP. Black dot represents the solution in which the pathway was recovered. Ordinate axis is log scaled. (**A**) Gluconeogenesis; (**B**) glycogen biosynthesis; (**C**) proline biosynthesis; (**D**) serine biosynthesis; (**E**) tirosine biosynthesis; (**F**) Entner–Doudoroff; (**G**) anaerobic respiration; (**H**) arginine degradation; (**I**) proline degradation; (**J**) biosynthesis of cysteine; (**K**) phenylalanine biosynthesis; (**L**) glutamate biosynthesis cycle; (**M**) arginine biosynthesis; (**N**), threonine degradation; (**O**) pentose phosphate pathway; (**P**) glycolysis; (**Q**) glyoxylate cycle; (**R**) TCA cycle

For example, taking the pathway in Figure 2Q as an example, note that aCFP recovers the corresponding canonical solution, namely the *Glyoxylate cycle*, between the third and the fifth solutions. In other words, under aCFP there are only $(5 - 3 + 1) = 3$ pathway solutions with the same length as the *Glyoxylate cycle* pathway. In particular, under the simulated computational conditions, this cycle was recovered as the third solution (black dot). CFP approach will recover it within the intervals third to seventh, being recovered in the fifth solution. Finally, AP needs 26 solutions before recovering the *Glyoxylate cycle*, obtaining pathways of the same length between the 21st and the 40th solutions.

First, we analyze the differences between aCFP and CFP. Both approaches show a similar behavior, however the white bars for aCFP are always equal or better than the dark gray bars for CFP. In particular, M and R pathways are always recovered earlier via aCFP, whereas in P and Q pathways aCFP outperforms CFP in most cases. This confirms that including the AMM information into CFP helps in calculating the canonical solutions faster. In contrast, significant differences arise when enumerating paths at the atomic level rather than at the metabolic level. It is clear that AP needs more solutions to assure that the canonical pathway is recovered.

Aiming at understanding the mechanisms leading to the differences in Figure 2, we now examine the first solutions obtained by CFP and aCFP, respectively, when recovering the *arginine biosynthesis pathway* (Fig. 3A and B), corresponding to the pathway M in Figure 2. In particular, we highlight the mechanisms captured by aCFP and overlooked by CFP.

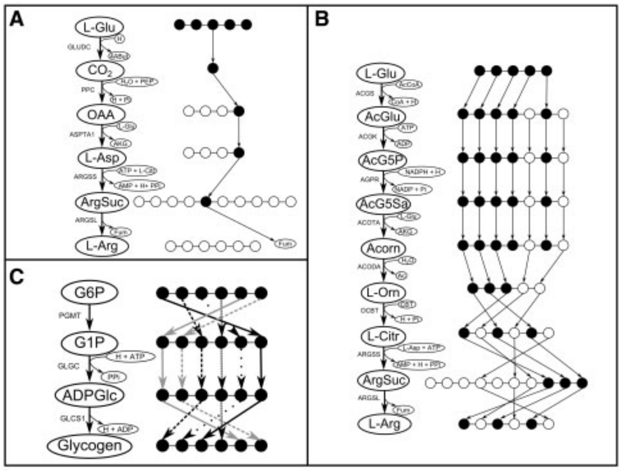

Fig. 3. Pathway calculated by (**A**) CFP and (**B**) aCFP when recovering the arginine biosynthesis pathway and imposing that at least three carbons from L-Glu reaches L-Arg. (**C**) atomic pathways arising from the glycogen biosynthesis

As discussed above, CFP guarantees effective carbon exchange between its constituents but not between the source and the target metabolites. This is illustrated in Figure 3A, where it can be observed that no carbon atom from *L-Glu* (source) reaches *L-Arg* (target). In Figure 3B we present the first pathway obtained by the aCFP approach, which is precisely the canonical pathway. This pathway conserves three atoms from the source to the target metabolite. It should be pointed out that mappings from Suthers *et al.* (2007) may involve errors, i.e. in Figure 3A the third carbon of *L-Glu* is being decarboxylated which usually takes place at the first carbon. Note, however, that developing an error-free atomic fate database goes beyond the scope of this article.

Figure 3C shows six atomic paths for the same metabolic solution, namely the *glycogen biosynthesis* (Fig. 2B). However, most applications in systems biology demand information regarding the active metabolites and enzymes. Therefore, enumerating solutions as done by aCFP prevents redundancy without loss of relevant information.

Summarizing, we have shown here that extending CFP to prevent solutions with no carbon flux between the source and target metabolites facilitates the recovery of canonical pathways. We expect the differences between aCFP and other approaches to increase with the size of the metabolic network under study and when all the carbon fates are included. However, several questions arise as to the feasibility of applying aCFP to genome-scale metabolic networks. In the next section, we shall inquire if extending the scope of aCFP to large metabolic networks is computationally feasible.

3.2 Genome-scale metabolic network

We considered here the *E. Coli* genome-scale metabolic reconstruction presented in Feist *et al.* (2007). We adopted this network on account of the recent publication of its computationally generated atomic fate reconstruction (Ravikirthi *et al.*, 2011).

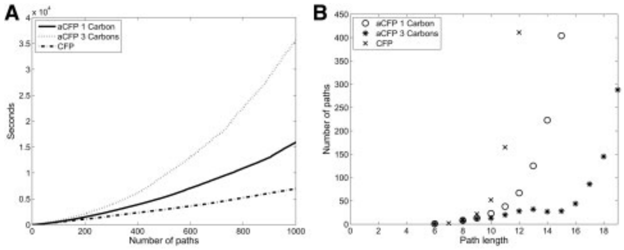

Fig. 4. (**A**) Accumulated computational time required to calculate 1000 paths between D-Glc and Pyr. (**B**) number of paths of a particular length ($Q = 1$ and $Q = 3$)

Unfortunately, these *in silico* methods still require further research so as to provide a completely error-free atomic reconstruction. However, the data in Ravikirthi *et al.*, 2011 can be used for testing the computation times of aCFP in genome-scale networks, which are not severely affected by atomic mapping errors. In addition, in order to reduce the impact of errors in the comparison between aCFP and CFP, the carbon graphs for both approaches were defined based on the same set of AMMs, keeping in mind that if $l \in A_i$, $m \in A_j$ and $e_{lmr} = 1$ then $d_{ijr} = 1$.

Summarizing, this metabolic network comprises a total of 1972 metabolites and 3234 reactions (after the splitting of reversible reactions) leading to a system of aCFP equations with around 75 000 constraints to be satisfied by 35 000 variables. Figure 4A compares the accumulated computational time required by aCFP and CFP to calculate 1000 paths between glucose (*D-Glc*) and Pyruvate (*Pyr*). These pathways were calculated in a regular laptop (Intel Core i5-2.53 GHz and 4GB of RAM) using the IBM ILOG CPLEX 12.5 solver, which reduces the problem to 13 834 constraints and 12 660 variables through automatic algebraic manipulation/substitution. Particularly, Figure 4A shows that aCFP calculated 1000 paths in ~4 h, on average one path every 15 s. Although aCFP turned out to be around two and five times slower than CFP when $Q = 1$ and $Q = 3$, respectively, it is clear that this result shows that aCFP can be applied in genome-scale networks as the computational time required is not excessive.

For the same biological question, i.e. pathways from *D-Glc* to *Pyr* in *E. coli* genome-scale metabolic network, Figure 4B represents the number of solutions of a particular length using aCFP ($Q = 1$ and $Q = 3$) and CFP, respectively. Note that as we are enumerating pathways in ascending length order, the computation of solutions of a particular length implies the previous computation of shorter solutions. Differences are clear. For example, we found 411 paths of length 12 with CFP, 67 with aCFP when $Q = 1$ and only 28 when $Q = 3$. These differences increase as pathway lengths increase, e.g. CFP was unable to enumerate all paths of length 13 within the first 1000 solutions, while aCFP computed all solutions of length 19 for $Q = 3$.

After a general inspection of Figure 4B, we can conclude that the number of feasible pathways is dramatically reduced after including atomic mapping information. In particular, the pathways calculated by CFP but neglected by aCFP are, precisely, those not guaranteeing carbon exchange between the source and the target metabolites. It is important to note that such pathways are not biologically relevant. We also emphasize the importance of Q, which determines both the quality and quantity of feasible pathways. We found that as Q increases, the number of feasible pathways is also considerably reduced.

Overall, we can conclude that aCFP improves on CFP in terms of the ability to generate relevant pathways as the network size increases.

4 DISCUSSION

Pathway analysis tools are a powerful strategy to study metabolic processes in the context of 'omics' data. We recently introduced a novel pathway concept termed CFPs, which was benchmarked with existing approaches, finding a clear progress over the state of the art (Pey *et al.*, 2011). CFPs approach has been shown to be particularly effective to determine linear pathways from a given source to a target metabolite. In other works, we illustrate the versatility of CFPs approach for different applications. In the first work we extract key pathways based on gene and protein expression data that underlies the acetate overflow issue in *E. coli* cell cultures. In the second work, based on CFPs, we interpret metabolomics data in different neurological disorders (Pey *et al.*, 2013).

Despite the potential of CFPs, improvements are still possible. In particular, while the underlying graph of metabolites guarantees carbon exchange in each intermediate step of the path, this is not always achieved between the source and target metabolites and false-positive solutions may arise. These pathways often lack biological interest, particularly when studying biosynthetic or degradation routes of a metabolite. To overcome this issue, we amend the formulation in Pey *et al.* (2011), so as to account for atomic fate information. This approach was termed aCFP.

In Section 3, by means of a side-by-side comparison in a medium scale metabolic network, we show that aCFP provides more biologically relevant pathways than CFP, because canonical pathways are more easily recovered, which reflects the removal of false positives. Note here that CFP represents a state-of-the-art approach and it is challenging to obtain a notable improvement over a high-quality approach. In particular, in our 18 canonical pathway analysis, CFP (and aCFP) recovered the solution in 12 cases. Consequently for these 12 pathways CFP cannot be outperformed by any other approach. Nevertheless, in those six cases in which CFP did not recover the canonical pathway in the first solution, aCFP improved the results obtained by CFP in four out of the six cases, some of them significantly (e.g. arginine biosynthesis pathway).

The limited size of the metabolic network considered in Section 3.1 makes it difficult to obtain large differences between aCFP and CFP. However, in Section 3.2 we used a large-sized genome-scale metabolic network and found remarkable differences between aCFP and CFP in finding pathways between glucose and pyruvate, as observed in Figure 4.

Results in Section 3.2 also demonstrate that aCFP can be successfully applied to genome-scale metabolic networks. However, the performance of aCFP is intimately tied to the quality of the atomic reconstruction under consideration. Although there are high-quality manually curated medium-scale atomic reconstructions, genome-scale atomic models generated by means of *in silico* methods still require further research so as to provide atomic reconstructions free from errors. This work will be also valuable for those conducting and modeling isotope labeling experiments, the most quantitative technology to infer reaction fluxes (Pey *et al*., 2012).

There are other biological mechanisms demanding pathways assuring exchange of different atoms, e.g. reactive oxygen species (ROS) in different applications in health (Selivanov *et al*., 2011); nitrogen assimilation in bacteria (Amon *et al*., 2010). aCFP can be easily adapted so as to guarantee any atom exchange, simply by introducing the atomic reconstructions into the model presented here.

Overall, combining CFP with high-quality atomic reconstructions further improves the accuracy of predicted pathways. Therefore, as high-quality atomic reconstructions become available, aCFP substitutes the CFP approach, as 'omics' experimental information can be integrated in a similar manner and we expect to obtain more reliable results. With the increasing amount of 'omics' data, tools such as the one presented here will be essential to obtain relevant insights from them.

Funding: The work of Jon Pey was supported by the Basque Government.

Conflict of Interest: none declared.

REFERENCES

Amon,J. *et al*. (2010) Common patterns–unique features: nitrogen metabolism and regulation in Gram-positive bacteria. *FEMS Microbiol. Rev*., **34**, 588–605.

Arita,M. (2000) Metabolic reconstruction using shortest paths. *Simulat. Pract. Theory*, **8**, 109–125.

Dijkstra,E.W. (1959) A note on two problems in connection with graphs. *Numerische Mathematik*, **1**, 269–271.

Faust,K. *et al*. (2009) In response to "Can sugars be produced from fatty acids? A test case for pathway analysis tools". *Bioinformatics*, **25**, 3202–3205.

Feist,A.M. *et al*. (2007) A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. *Mol. Syst. Biol*., **3**, 121.

De Figueiredo,L.F. *et al*. (2009a) Can sugars be produced from fatty acids? A test case for pathway analysis tools. *Bioinformatics*, **25**, 152–158.

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

Heath,A.P. *et al*. (2010) Finding metabolic pathways using atom tracking. *Bioinformatics*, **26**, 1548–1555.

Kamp,A. and Schuster,S. (2006) Metatool 5.0: fast and flexible elementary modes analysis. *Bioinformatics*, **22**, 1930–1931.

Kanehisa,M. *et al*. (2012) KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res*., **40**, D109–D114.

Keseler,I.M. *et al*. (2009) EcoCyc: A comprehensive view of Escherichia coli biology. *Nucleic Acids Res*., **37**, D464–D470.

Kitano,H. (2002) Systems biology: a brief overview. *Science*, **295**, 1662–1664.

Latendresse,M. *et al*. (2012) Accurate atom-mapping computation for biochemical reactions. *J. Chem. Inf.Model*., **52**, 2970–2982.

Lim,K. and Wong,L. (2012) CMPF: Class-switching minimized pathfinding in metabolic networks. *BMC Bioinformatics*, **13**, S17.

Mu,F. *et al*. (2007) Carbon-fate maps for metabolic reactions. *Bioinformatics*, **23**, 3193–3199.

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

Pey,J. *et al*. (2012) Integrating tracer-based metabolomics data and metabolic fluxes in a linear fashion via elementary carbon modes. *Metab. Eng*., **14**, 344–353.

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.

Pitkänen,E. *et al*. (2009) Inferring branching pathways in genome-scale metabolic networks. *BMC Syst. Biol*., **3**, 103.

Planes,F.J. and Beasley,J.E. (2008) A critical examination of stoichiometric and path-finding approaches to metabolic pathways. *Brief. Bioinform*., **9**, 422–436.

Price,N.D. *et al*. (2003) Network-based analysis of metabolic regulation in the human red blood cell. *J. Theor. Biol*., **225**, 185–194.

Ravikirthi,P. *et al*. (2011) Construction of an E. Coli genome-scale atom mapping model for MFA calculations. *Biotechnol. Bioeng*., **108**, 1372–1382.

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

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

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

Selivanov,V.A. *et al*. (2011) Reactive oxygen species production by forward and reverse electron fluxes in the mitochondrial respiratory chain. *PLoS Comput. Biol*., **7**, e1001115.

Stelling,J. *et al*. (2002) Metabolic network structure determines key aspects of functionality and regulation. *Nature*, **420**, 190–193.

Suthers,P.F. *et al*. (2007) Metabolic flux elucidation for large-scale models using ^{13}C labeled isotopes. *Metab. Eng*., **9**, 387–405.

Terzer,M. and Stelling,J. (2008) Large-scale computation of elementary flux modes with bit pattern trees. *Bioinformatics*, **24**, 2229–2235.

Zupke,C. and Stephanopoulos,G. (1994) Modeling of isotope distributions and intracellular fluxes in metabolic networks using atom mapping matrices. *Biotechnol. Prog*., **10**, 489–498.