BIOINFORMATICS ORIGINAL PAPER Vol. 27 no. 14 2011, pages 1957–1963
doi:10.1093/bioinformatics/btr271

Systems biology

Advance Access publication May 5, 2011

The strength of chemical linkage as a criterion for pruning metabolic graphs

Wanding Zhou[1,*] and Luay Nakhleh[2,*]

[1]Department of Bioengineering and [2]Department of Computer Science, Rice University, Houston, TX, USA

Associate Editor: Trey Ideker

ABSTRACT

Motivation: A metabolic graph represents the connectivity patterns of a metabolic system, and provides a powerful framework within which the organization of metabolic reactions can be analyzed and elucidated. A common practice is to prune (i.e. remove nodes and edges) the metabolic graph prior to any analysis in order to eliminate confounding signals from the representation. Currently, this pruning process is carried out in an *ad hoc* fashion, resulting in discrepancies and ambiguities across studies.

Results: We propose a biochemically informative criterion, the *strength of chemical linkage* (*SCL*), for a systematic pruning of metabolic graphs. By analyzing the metabolic graph of *Escherichia coli*, we show that thresholding *SCL* is powerful in selecting the conventional pathways' connectivity out of the raw network connectivity when the network is restricted to the reactions collected from these pathways. Further, we argue that the root of ambiguity in pruning metabolic graphs is in the continuity of the amount of chemical content that can be conserved in reaction transformation patterns. Finally, we demonstrate how biochemical pathways can be inferred efficiently if the search procedure is guided by *SCL*.

Contact: wz4@rice.edu; nakhleh@rice.edu

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

Received on December 22, 2010; revised on April 13, 2011; accepted on April 22, 2011

1 INTRODUCTION

Graph representation of a metabolic network connectivity map provides a simple representation of certain relationships among the network's entities. Analyses of such graphs have provided various insights into the properties of metabolic networks, yet not without controversy. For example, the finding of a short average path length in metabolic networks [e.g. Jeong *et al.* (2000); Wagner and Fell (2001)] has been challenged in that it was based on the 'raw' metabolic graphs, without first *pruning* them (Arita, 2005). To illustrate the concept of 'pruning', consider the following reaction analyzed in Ma and Zeng (2003):

$$N\text{-acetylornithine} + \text{L-glutamate} \Leftrightarrow \text{L-ornithine} + N\text{-acetyl-L-glutamate} \quad (1)$$

In their network assembly, the authors only linked *N*-acetylornithine to L-ornithine and L-glutamate to *N*-acetyl-L-glutamate and omitted the link between L-glutamate and L-ornithine and the link between *N*-acetylornithine and *N*-acetyl-L-glutamate. Depending on the edge semantics of the network and the subsequent analyses, this pruning step may or may not make a difference. From the perspective of *causality of biochemical transformation* and *pathway inference* (Ma and Zeng, 2003), this pruning makes sense: no chemical content is conserved between L-ornithine and L-glutamate, and the acetyl group that is conserved between *N*-acetylornithine and *N*-acetyl-L-glutamate is not sufficiently representative for linking the two chemical compounds. However, if one is modeling the *propagation of perturbation* in the concentration of metabolites [e.g. Wagner and Fell (2001)], then not removing these last two edges makes sense, since they do capture how a perturbation to certain metabolites may spread throughout the metabolic system. Under this semantics, metabolic graphs are built by connecting all the metabolites that participate in a reaction (Holme, 2009), and subsequently analyzed, without post-processing the connectivity (Ravasz *et al.*, 2002), to elucidate properties on information transfer, network robustness and resilience, etc. In this article, we focus on the first of the edge semantics, namely causality of biochemical transformation.

To build metabolic graphs for pathway inference, all metabolites participating in a reaction are connected to form the *raw graph*, and then, via *connectivity pruning*, edges that may result in the inference of biochemically implausible pathways [Ma and Zeng (2003); van Helden *et al.* (2002)] are pruned [hypergraph-based pathway inference techniques, such as the network expansion (Ebenhh *et al.*, 2004), require different treatment and are beyond the scope of this article]. Several methods exist for pruning metabolic graphs including hub deletion (Diaz-Mejia *et al.*, 2007), removal of currency metabolites (Herrgard *et al.*, 2008; Zhao *et al.*, 2007), manual curation (Zhao *et al.*, 2006) and RPair typing (Faust *et al.*, 2009). However, the ambiguity inherent in these *ad hoc* methods, and the lack of a systematic one, may confound analyses of metabolic graphs (Tanaka, 2005; Zhao *et al.*, 2006; Zhu and Qin, 2005). Here, we propose a simple criterion, the *strength of chemical linkage* (*SCL*), for systematic pruning of metabolic graphs. By analyzing the metabolic graph of *Escherichia coli*, we demonstrate the power of this criterion in yielding biochemically meaningful pathways. Further, we characterize the commonly used pruning heuristics in terms of the strength of chemical linkage, and discuss the ambiguity in these methods and the superiority of using the *SCL* criterion. Finally, we demonstrate the utility of the criterion in pruning the search tree used in pathway inference methods to gain in accuracy and efficiency compared with other graph-based search heuristics [e.g. Croes *et al.* (2006)].

*To whom correspondence should be addressed.

Table 1. Inference of aMAZE pathways (Lemer *et al.*, 2004)

Pathway name	Length	Rank
Arginine Catabolism	3	1
Arginine Utilization	4	1
Chorismate Biosynthesis	4	1
Glucuronate Catabolism	3	1
Lysine Biosynthesis	7	1
Threonine Biosynthesis	3	1
Oxidative Pentose Phosphate Pathway	4	1
Glycolysis	6	2
Methionine Biosynthesis	5	95

'Length' is the number of reactions from the source to the target compound in the reference pathway. 'Rank' is the place of the reference pathway as identified by Algorithm *InferPathway*.

2 METHODS

2.1 Reaction data and reference pathways

The 1383 reaction equations were obtained from KEGG Ligand database (Kanehisa and Goto, 2000). For each reaction with any gene in *E.coli* annotated to produce an enzyme that catalyzes the reaction, we assembled a graph connecting every pair of metabolites that sit on opposite sides of the reaction (the raw graph). Both reaction-enzyme mapping information and enzyme-gene mapping information were downloaded from KEGG. Following common practice (Lee *et al.*, 2006), we removed any reaction that appeared in the reference pathway and yet did not have a definite gene annotation; e.g. reaction R07765 has only the generic EC number 1.3.1.- even though there are genes in *E.coli* that are annotated for that EC number. For reactant pairs that exist in the KEGG RPair database, we used information on the molecule alignment. 'Markush structures' and groups with label 'R' were taken as one atom. For reactant pairs that do not exist in the database, we manually set the alignment number to 0. This treatment is dependent on the coverage of RPair database to all possible reactant pairs with non-zero alignments. We manually verified that the coverage is satisfactory. Out of 1383 reactions that exist in the *E.coli* network, 1104 reactions needed to be treated with additional specification to connections with alignment number 0. Out of 2642 connections with alignment number 0 added in these reactions, 98 connections have actual non-zero chemical linkage. The percentage of unsatisfactory connections was less than 4%. Moreover, from a closer inspection, many of these linkages are hard to process because of the use of generic compounds and unbalanced reactions in the KEGG Ligand database (Blum and Kohlbacher, 2008; Poolman *et al.*, 2006).

Reference pathways were obtained from the KEGG KGML pathway files. Reactions that exist in the reference pathways but are not validated by the presence of clearly defined enzymatic information, and thus do not appear in the total set of reactions from which the raw metabolic graph is assembled, were removed. Annotated pathways for the pathway inference validation were obtained from the aMAZE database (Lemer *et al.*, 2004). We excluded those pathways that are duplicates in terms of using the same sequence of compounds and those that have fewer than two steps. Two other manually curated reference pathways were also included. The name of the pathways are listed in Table 1. The main metabolites and reactions of these pathways are listed in the Supplementary Material.

2.2 The *SCL* criterion

We define the *strength of chemical linkage*, or *SCL*, for two reactants as the proportion of chemical content conserved between them in a reaction, normalized by the maximum chemical content of either of the two reactants. If there is more than one mechanism that involve the same two reactants, *SCL* takes the maximum result computed over all such mechanisms. For example, in some rare cases, two reactants can be converted to one another via more than one mechanism even in one reaction; e.g. C00022-C00900 in R00006 [the C and R labels are standard indices used in KEGG (Kanehisa and Goto, 2000)]. Chemical content can be quantified in many ways; in this article, we use the *absolute atom counting*. Under this quantification, given the set $\mathcal{C}(A)$ of non-hydrogen atoms in molecule A (due to the fact that hydrogen is not generally considered the backbone of biochemical compounds), the chemical content of the compound is simply $|\mathcal{C}(A)|$. Atoms are mapped in the reaction according to the true chemistry. While in this study we use the KEGG RPair database for molecule alignments, *SCL* depends only on the physics of the real chemical reaction and is independent of the data source.

Formally, for a compound pair (A,D) that sit on two sides of a reaction (e.g. $A+B \Leftrightarrow C+D$), we define

$$\begin{aligned} SCL_{\text{self}}(A|D) &= |\mathcal{C}(A)\cap\mathcal{C}(D)|/|\mathcal{C}(A)| \\ SCL &= |\mathcal{C}(A)\cap\mathcal{C}(D)|/\max(|\mathcal{C}(A)|,|\mathcal{C}(D)|) \\ &= \min(SCL_{\text{self}}(A|D), SCL_{\text{self}}(D|A)). \end{aligned} \tag{2}$$

$SCL_{\text{self}}(A|D)$ measures the contribution of chemical content from D to A, or equivalently, how much chemical content of A comes from D. A high $SCL_{\text{self}}(A|D)$ value indicates a greater importance in chemical composition of D when A is produced/consumed in the reaction. When compound D is clear from the context (e.g. when there is only one reactant on the other side), we write $SCL_{\text{self}}(A)$. Further, in this context, we use the definition $SCL_{\text{other}}(A)=SCL_{\text{self}}(D|A)$. High *SCL* is an indication of stronger chemical causality in terms of chemical content between the two reactants in a specific reaction.

2.3 A pathway inference method

In order to demonstrate the quality of *SCL*-based pruning, we introduce a simple, *SCL*-based algorithm for identifying a set of pathways from a source to a target metabolite in a given metabolic graph; see Algorithm 1. The algorithm first identifies a set of candidate paths within a maximum length in a breadth-first manner (Lines 5–10), extending paths (Lines 9–10) only using edges with an *SCL* value higher than a certain threshold T (in our study, we use $t=0.4$; see Section 3.1 for a discussion of this choice). After exploring all the valid paths of a certain length, all paths are reordered according to the minimum SCL_{self} value of all steps along the path (Line 11). Only the top (if there are more than) $N=1000$ paths are saved for further exploration in the next round (Line 12). Finally, the paths are ranked by the minimum SCL_{self} value of any edge they contain, where paths with lower values are ranked higher (Line 13).

Algorithm 1: InferPathway.

Input: Source metabolite: s; Target metabolite: t; Threshold of *SCL*: T; Maximum path length: L;
Output: A list of ranked pathways $\mathcal{S}_{\text{Result}}$.

1 $\mathcal{S}_{\text{ToExplore}} \leftarrow$ a path that contains only t;
2 $\mathcal{S}_{\text{Result}} \leftarrow \emptyset$;
3 **while** $\mathcal{S}_{\text{ToExplore}} \neq \emptyset$ *&& path length* $< L$ **do**
4 $\mathcal{S}_{\text{Temp}} \leftarrow \emptyset$;
5 **foreach** *path p in* $\mathcal{S}_{\text{ToExplore}}$ **do**
6 **foreach** *neighbor n of the last node l in p* **do**
7 **if** $n=s$ **then**
8 $\mathcal{S}_{\text{Result}} \leftarrow p$ extended by n;
9 **else if** $n \notin p$ *&&* $SCL_{self}(l|n) > T$ **then**
10 Extend p using n and add the new path into $\mathcal{S}_{\text{Temp}}$;
11 Sort $\mathcal{S}_{\text{Temp}}$ by the minimum SCL_{self} on each path;
12 $\mathcal{S}_{\text{ToExplore}} \leftarrow$ top 1000 paths in $\mathcal{S}_{\text{Temp}}$;
13 Sort $\mathcal{S}_{\text{Result}}$ by the minimum SCL_{self} on each path;
14 **return** $\mathcal{S}_{\text{Result}}$;

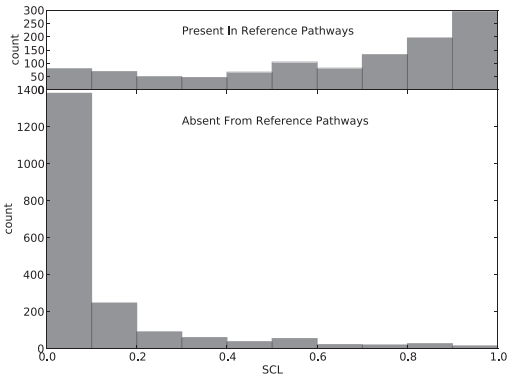

Fig. 1. The distribution of *SCL* values based on the 1124 edges in the raw metabolic network of *E.coli* that are present in (top panel) and 1957 ones that are absent from (bottom panel) the reference pathways.

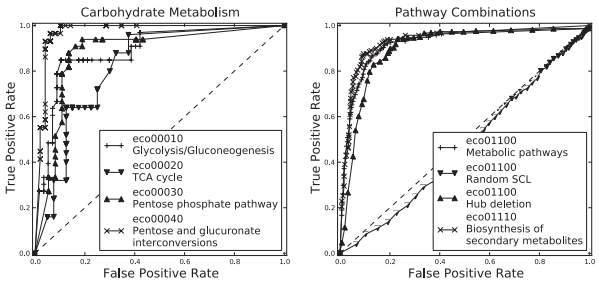

Fig. 2. ROC curves based on thresholding the *SCL* values. Each curve is based on the raw graph restricted to a particular pathway map and on 50 threshold values evenly distributed in the range [0,1]. Similar results have been obtained on other pathway maps (see Supplementary Material). The right panel is based on a combination of pathways where the ROC curve of hub deletion and one based on random *SCL* assignment are also shown with up and down triangles, respectively. The ROC curve on hub deletion is obtained by tuning the degree threshold of the hub definition, thus changing the presence/absence of the network connections.

3 RESULTS

To assess the quality of *SCL*, we conducted three tasks on *E.coli*'s metabolic graph. First, we computed the distribution of *SCL* values for edges in the raw graph, categorized based on the presence/absence in the reference pathways to assess the power of *SCL* in selecting connectivities. Second, we explored pathway inference guided by *SCL* values to assess its power in finding compound-to-compound linear/cyclic biochemical pathways. Third, we studied existing metabolic graph pruning methods in the light of the *SCL* criterion.

3.1 The distribution of *SCL* values

Figure 1 shows the distribution of *SCL* values of raw graph edges present/absent in the reference pathways. Most of the edges present in the reference pathways have high *SCL* values, whereas a great majority of those missing from the reference pathways have *SCL* values lower than 0.2. These results show that *SCL* is strongly correlated with the absence/presence in reference pathways; stated differently, *SCL* can be used as a selection criterion for deciding which connections to include in a reference pathway.

Next, we studied whether pruning the raw graph based on thresholding the *SCL* values (keeping only edges with *SCL* values no less than a threshold T) produces the conventional pathway connectivities. Pathway maps are organized into five categories according to the hierarchy provided by the KEGG pathway database: carbohydrate metabolism, nucleotide metabolism, lipid metabolism, amino acid metabolism and metabolism of cofactors and vitamins. In addition, we considered two higher level pathway unions provided by KEGG, namely eco01100 (Metabolic Pathways) and eco01110 (biosynthesis of secondary metabolites). Receiver operating characteristic (ROC) curves for some of the pathways in the carbohydrate metabolism and the pathway unions are shown in Figure 2.

For a specific pathway, we count the number of edges from the raw metabolic network that are present in the reference pathway. Positives (P) [Negatives (N)] are defined as the connections that do (do not) exist in the reference pathway. For each value of the threshold T, only edges with *SCL* values $\geq T$ are kept, and the rest are removed. The true positives (TP) are the positives that also exist in the thresholded network. The false positives (FP) are the positives that do not exist in the thresholded network. The true positive rate (TPR) equals $|\mathrm{TP}|/|\mathrm{P}|$ and the false positive rate (FPR) equals $|\mathrm{FP}|/|\mathrm{N}|$. Notice that when $T=0$, TP = P and FP = N, giving TPR = FPR = 1. On the other hand, when $T>1$, TP = FP = $\emptyset$, giving TPR = FPR = 0. In each panel, T increases in the direction from the upper-right corner to the lower-left corner. The concave shapes of the ROC curves indicate that thresholding *SCL* has strong power of selection for connections that appear in pathways, as opposed to those that are missing. This further validates our implicit reasoning that conventional pathway connectivity is a reflection of the chemical linkage strength.

While no *SCL* threshold seems to exist for perfect retrieval of established biochemical pathways, our detailed study of the Glycolysis/Gluconeogenesis pathway (with threshold $T=0.4$) revealed two main reasons (beside the issue with KEGG's RPair database coverage) behind the false cases. The first reason is the presence of reactant pairs with special roles. Not all the reactant pairs actively participate in the mass circulation, but are required for, e.g. energy dependencies. For example, ATP–ADP drives a reaction toward a certain direction (Ma and Zeng, 2003). Many such reactant pairs have strong chemical linkage with each other, yet weak linkage with other reactants. They are commonly perceived as carriers of small chemical moieties, such as proton (NAD, NADH), phosphate (ATP, ADP; Protein–histidine, Protein *N*-phospho-L-histidine) and acetyl group (CoA, Acetyl-CoA). These reactant pairs with special roles usually cause false positives, i.e. reactant pairs absent from the reference pathways but with a high chemical linkage. Nonetheless, these false positives are completely tolerable and, in our view, are even better to be preserved in the network. For pathways where these reactant pairs are used for non-mass circulation reasons (e.g. eco00010), they are usually disconnected from the bulk network component due to a low *SCL* value with other reactants (see Supplementary Material). Therefore, their presence would not confuse the pathway inference by creating biochemically unintuitive shortcuts. Moreover, it makes sense to preserve these reactant pairs in the network and the reference pathways since they represent the way how energy is consumed. For example, when ADP is used to make ATP, such cycling would be unclear if the consumption of ATP is missing from the network. More importantly,

being a 'carrier' is in itself ambiguously defined, since all reactant pairs 'carry something'. In order to be a 'carrier', same reactant pair should appear in multiple reactions to 'load' and 'unload' chemical groups. We have observed that although there exist certain reactant pairs that participate in a large number of reactions, there is in general no clear-cut boundary for being a 'carrier' (see Supplementary Material).

The second reason is the inappropriate quantification of chemical content by *absolute atom counting*. For example, in the following reaction [pyruvate + thiamin diphosphate ⇔ 2-(alpha-hydroxyethyl)thiamine diphosphate + CO_2], reactant pair (CO_2, pyruvate) is missing from the reference pathway and its presence in the thresholded network can potentially give rise to shortcuts between pyruvate and other irrelevant compounds via CO_2. The total number of non-hydrogen atoms of pyruvate is 6 and that of CO_2 is 3. Although they share three atoms in the reaction, two of them are oxygen and only one is carbon, which is traditionally considered to be the backbone of Pyruvate. Therefore, when all the non-hydrogen atoms are included $SCL = 3/6 = 0.5$. However, the value becomes $SCL = 1/3 \sim 0.3$ when only carbon atoms are considered, since Pyruvate has three carbons, CO_2 has one, and they share one carbon in the reaction. By counting only carbons, the false positive connection can be avoided. This not only reflects the fact that the atoms are treated as biochemically different, but also suggests that alternatives to *absolute atom counting* (e.g. by counting only carbons) might improve the performance of the criterion on certain (but not all) reactions.

Further, a large collection of atoms may form some chemical group that functions as a single unit. In the context of one particular pathway, their detailed composition, creation and degradation is not relevant. But, again, if we count only the number of non-hydrogen atoms, the criterion might be biased. For example, in reaction [ATP + acetate + CoA ⇔ AMP + Diphosphate + Acetyl-CoA], the reactant pair (CoA, acetyl-CoA) is falsely present (false positive) while the pathway reactant pair (acetate, acetyl-CoA) is missing (false negative). The 48 non-hydrogen atoms are conserved in the former reactant pair, while only 3 are conserved in the latter. Under *absolute atom counting*, the former reactant pair is stronger in chemical linkage while the latter is weaker. However, biochemically, all 48 atoms in the former comes from CoA which functions as a single unit in the context of glycolysis pathway, while the three atoms conserved in the latter contribute to three out of four non-hydrogen atoms in acetate. If we count all atoms in CoA as 1, we obtain $SCL = 1/4 = 0.25$ for the (Acetate, Acetyl-CoA) pair and $SCL = 3/4 = 0.75$ for the (CoA, acetyl-CoA) pair, which would avoid both false cases. However, to do this throughout the metabolic graph, finer delineation of functional groups for metabolites is needed, which we target as future work.

By scanning the threshold from 0 to 1 with increment 0.01, we find that the range of threshold value that minimizes the false cases (both false positive and false negative) is from 0.38 to 0.39. Nevertheless, as shown in Figure 2, when individual pathway is concerned, there does not exist a threshold value of *SCL* that suits all (see Supplementary Material). We found that the range of optimal threshold in different pathways varies not only in magnitudes, but also in lengths. Some pathways reach optimal pruning under a wide range of threshold values. Besides, the optimal threshold of some pathways can be explained by their biochemical function. For example, pathways involved in fatty acid metabolism have a lower threshold. This reflects the fact that links in these pathways are responsible for the extension of a long fatty acid chain by one small residue, which is weak in the sense of relative mass conservation.

3.2 Using *SCL* in pathway inference

In order to investigate the effectiveness of *SCL* in pathway inference, we applied Algorithm *InferPathway* (see Section 2) to source/target pairs of eight reference *E.coli* pathways obtained mostly from the aMAZE database (Lemer *et al.*, 2004). Results are shown in Table 1. In Figure 3, we show two of the reference pathways which are correctly returned by our method as the top result, namely, the Lysine Biosynthesis and Oxidative Pentose Phosphate Pathway, as well as the two pathways that differ from our top results namely, the Glycolysis and Methionine Biosynthesis pathways.

For Glycolysis (iii of Fig. 3), the only difference between the reference pathway and our top result (ranked second; see Table 1) is the use of reaction R01827 instead of a combination of reactions R04779 and R01070. The shortcut is a documented step in the KEGG pathway map, but only in the Pentose Phosphate pathway map, indicating that the exclusion of the single reaction step is of manual, rather than biochemical, origins. For the Methionine biosynthesis pathway (iv of Fig. 3), our method fails to return the annotated pathway as the top result. The only step that is consistently missing from our inference is the shortcut from O-succinyl-L-homoserine to L-homocystein without passing through Cystathionine, as is the case in the annotated pathway. Two clarifications are in order here. First, the shortcut step is structurally valid but infeasible in terms of free energy—information that is not incorporated into the reaction equation. The reaction's direction is to the left, as Hydrogen sulfide takes the gas form under room temperature and leaves the system quickly once formed: [O-succinyl-L-homoserine + Hydrogen sulfide ⇔ L-Homocysteine + Succinate]. Second, the missing step in the annotated path, composed of the two reactions [O-succinyl-L-homoserine + L-cysteine ⇔ Cystathionine + Succinate] and [Cystathionine + H_2O ⇔ L-Homocysteine + NH3 + Pyruvate] has a low SCL_{self} value of 0.5. This is because to add just an SH (mercapto group), a cysteine is recruited and pyruvate is released subsequently. However, when computing SCL_{self} of the link from O-succinyl-L-homoserine to L-cystathionine, the entire cysteine contributes to the number of non-conserved atoms, although only SH (1 non-hydrogen atom) is preserved after the subsequent step. The major part of cysteine (6 non-hydrogen atoms) is released as pyruvate. This issue can be solved by tracking the identity of each atom and recording for each intermediate metabolite the set of its atoms that are conserved all the way to the target.

We also studied the capability of the algorithm to find cyclic pathways by applying it to contiguous metabolites on the TCA cycle pathway (v in Fig. 3). The pathway connections corresponding to the top nine results returned are shown. Indeed, all intermediate metabolites, except for the Succinyl-CoA, in the TCA cycle are recovered as well as all other connections that are documented in the pathway maps of KEGG. The low *SCL* value (0.13) on one of the two missing steps involving Succinyl-CoA is due to a reason same as discussed above, namely that CoA contains many atoms yet functions as one unit.

A satisfying consequence of this pruning strategy is its capability of not only getting pathways but also rejecting cases where in

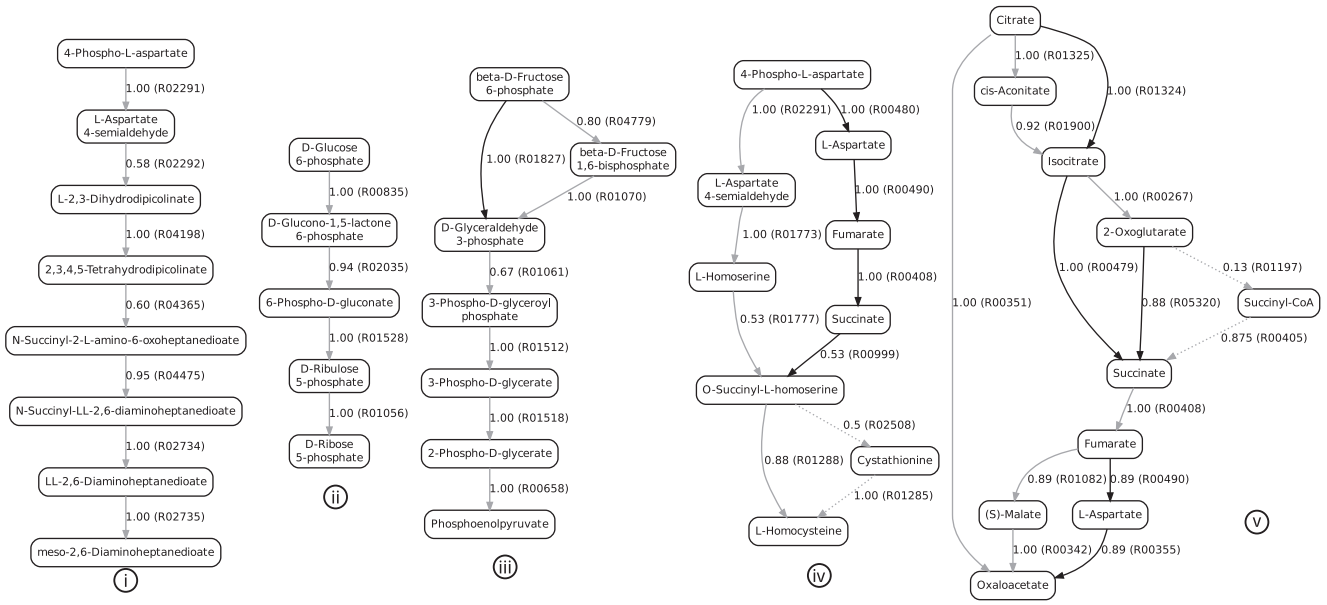

Fig. 3. Pathways inferred by applying algorithm *InferPathway* onto the *E.coli* metabolic network. Nodes are labeled by the metabolite name used in KEGG. Edges are labeled by the SCL_{self} of the link, along with the reaction ID used in KEGG of one of the reactions that make available the transition shown in parentheses. The solid connections correspond to the top 9 results returned from our method. (i) Lysine Biosynthesis. (ii) Pentose Phosphate Pathway. (iii) Glycolysis. Dim solid connections are the annotated pathway and the second result returned by our method. The top result uses the shortcut that is shown in dark color. (iv) Methionine Biosynthesis. Dark solid connections are the top result returned by our method. The annotated pathway differs from the first result by the dashed connections. The second result differs from the first result by the dark solid connections. (v) TCA (tricarboxylic acid) cycle. The annotated TCA cycle is shown in dim color (the dashed links are missing from the result).

between the given source and target there is no linear unbranched pathway that is biochemically meaningful. If the synthesis of a metabolite requires contributions from many different sources, this advantage would be reflected in our method as a low minimum SCL_{self} of all paths returned. To further illustrate the efficiency of pathway inference guided by SCL_{self}, we compare the search efficiency by only considering non-zero *SCL* (or equivalently, the presence/absence of annotations for reactant pairs in the RPair database), by pruning of extension using the SCL_{self}, by the pruning of exploration using minimum SCL_{self} on the path and by a combination of both pruning. We found first that no matter how large the results, minimum SCL_{self} always sorts out the reference pathway to a high rank (2 in case of glycolysis) while sorting using the path length does not (170 ties with the highest rank being 65). With the same accuracy of the reference pathway, the result set (Fig. 4) and total number of node visits (see Supplementary Material) is greatly reduced when our pruning strategies are applied. Same observations have been obtained from other pathways. These results combined demonstrate the utility of *SCL* in not only sorting out the best pathway from the result set but also being effective in pruning the search tree of path finding.

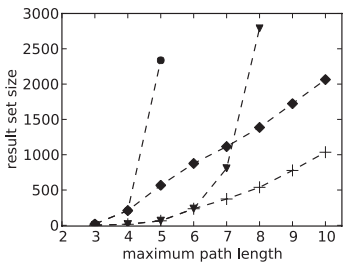

Fig. 4. Comparison of pruning strategies. The size of results computed on Glycolysis pathway as the maximum path length increases. ○: No *SCL* pruning. Only presence and absence of RPair is used. ◇: Pruning of path exploration using minimum *SCL* on the pathway (see Methods). ∇: Pruning of path extension using SCL_{self}. +: Combination of both path exploration pruning and path extension pruning. In all cases, the reference pathway ranks the second in the result set (see Table 1).

3.3 Existing pruning methods in the light of *SCL*

Here, we compare existing pruning methods in terms of the *SCL* criterion, and discuss the necessity and superiority of *SCL*.

3.3.1 Hub deletion (Diaz-Mejia et al.*, 2007) and currency metabolites (Herrgard et al., 2008)* Figure 5 shows that the average *SCL* value of a graph increases as hubs (nodes of high degree) are removed from the graph, which is in agreement with the rationale behind hub deletion (Diaz-Mejia *et al.*, 2007). Despite its similar effectiveness in pruning the network connection (red curve in Fig. 2), the degree of a metabolite depends only on the global layout of the network which has little meaning in the local chemistry of each reaction. Indeed, the lack of smoothness of the curves indicates a poor correlation between the node degree and the *SCL* [similarly shown by Faust *et al.* (2009)]. Hub deletion is known to suffer from several issues (Arita, 2005; Zhao *et al.*, 2006)), one of which is the coarse grainedness in the sense that connections are pruned by deleting metabolites as well as all their connections. Accordingly, we

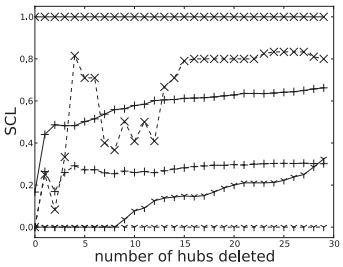

Fig. 5. Change in *SCL* values as more hubs are deleted by decreasing order of their degrees. Solid line: presence in the pruned graph. Dashed line: absence from the pruned graph. ×: upper quartile (75% in *SCL*). +: median (50% in *SCL*). *Y*: lower quartile (25% in *SCL*).

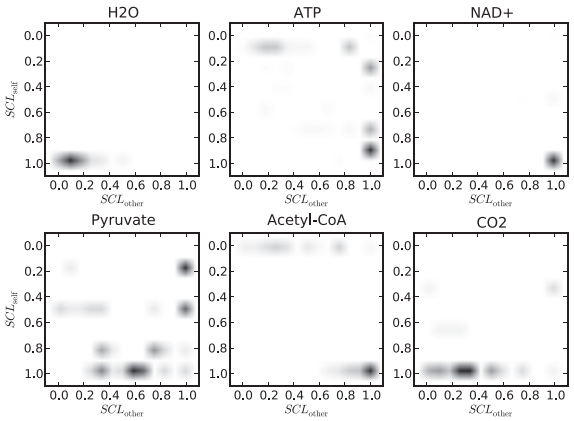

Fig. 6. The *SCL* signature for six metabolites: H_2O, ATP, NAD^+, Pyruvate, Acetyl-CoA, and CO_2. The darker the dot for a metabolite *C*, the more reactions exist in *E.coli*'s metabolic network with connections of ($SCL_{\text{other}}(C)$; $SCL_{\text{self}}(C)$) combination.

have observed that not all the connections of these hub metabolites are of low *SCL* values.

Along the same line, some metabolites, defined in an *ad hoc* fashion, and referred to as *pool metabolites* (Deville *et al.*, 2003) or *currency metabolites* (Ma and Zeng, 2003), which largely coincide with the hub metabolites (Croes *et al.*, 2006) are often defined for network pruning. A widely used example is ATP (Blum and Kohlbacher, 2008; Zhao *et al.*, 2006). Although ATP in many cases serves as a carrier of phosphate groups and an energetic driver of reactions, and it is also actively involved in the mass circulation of nucleotide metabolism. The versatility of ATP can be demonstrated by a *signature* based on SCL_{self} and SCL_{other} (Fig. 6). The functionality of serving as a carrier of small chemical moieties is reflected in the signature by two groups of dots, one in the top-left corner and the other in the bottom-right corner. The others correspond to other functions of ATP. Some of these involve a high SCL_{other} value—an indication of contribution to the mass circulation. For the same reason, we see in Figure 5 that as more and more hubs are deleted from the graph, more connections of high *SCL* values are eliminated (the increase in the dashed curves) as well. This suggests that a graph with high *SCL* values is hard to obtain without losing important connectivity information of the graph. From Figure 6, we also observe that some pool metabolites serve multiple functions (as indicated by multiple dots in the signature; e.g. ATP and pyruvate) while others are functionally specific (as indicated by a single dot in the signature; e.g. NAD^+ and H_2O). Compounds that are usually released as part of other bulk metabolites have dot(s) only in the bottom-left corner of their signatures (e.g. H_2O, CO_2 and NH_3).

3.3.2 Manual curation (Ma and Zeng, 2003; Zhao et al.*, 2006; Zhu and Qin, 2005)* In addition to being labor intensive and error prone, we show that pruning by manual curation may also be ambiguous. Consider reaction (1) above. The reactant pair *N*-acetylornithine and *N*-acetyl-L-glutamate is missing because the acetyl group is not sufficient to represent the link between the two compounds. Now, consider reaction

$$\text{Acetyl-CoA} + \text{formate} \Leftrightarrow \text{CoA} + \text{pyruvate} \qquad (3)$$

In this case, acetyl-CoA loses the acetyl group to formate to form pyruvate. The question is: should we eliminate the connection between acetyl-CoA and pyruvate? In this case, the acetyl group is important, as 2/3 of the carbon backbone of pyruvate comes from it. Hence, this connection should be kept in the network. Although reactions (1) and (3) show exactly the same mass transformation pattern [(Acetyl-Group I, Group II)] on one side of the reaction and (Group I, Acetyl-Group II) on the other side), we make different decisions on whether to prune away the connection between acetyl-Group I and acetyl-Group II. In fact, the intrinsic nature of chemical transformations implies that the amount of chemical moieties that are conserved in any reactant pairs is arbitrary (Fig. 1). A combination of the *SCL* pruning criterion with the objective quantification of chemical content can help ameliorate this problem.

3.3.3 RPair types (Faust et al.*, 2009)* Pruning methods also include filtering specific class(es) of reactant pairs inside a reaction (Faust *et al.*, 2009). The KEGG RPair database provides such classification by assigning reactant pairs to five different categories: 'main', 'cofac', 'trans', 'ligase' and 'leave" (Kotera *et al.*, 2004). The categorization is reaction dependent: one reactant pair may be of different types in different reactions. The typing is based on the classification of the enzymes (e.g. oxidoreductase, transferase, etc.) that catalyze the reaction and the role of the reactant pair in the reaction. However, the five types are manually curated, thus resulting in the same problems as discussed above.

In Figure 7, we plotted the distribution of *SCL* values of reactant pairs in the five categories. Reactant pairs of 'main' and 'cofac' tend to have higher *SCL* values, while reactant pairs of 'leave' and 'trans' tend to have lower *SCL* values. Pathway connections are composed of reactant pairs belonging to different categories, although the dominant categories are 'main' and 'trans' (Fig. 7). This is in agreement with the previous practice of using only these two types for metabolic network assembly (Faust *et al.*, 2009). Reactant pairs of type 'ligase' have a symmetric distribution, since this type consists of reactions where a large piece of chemical compound is decomposed into two components giving rise to two reactant pairs complementary in *SCL* values. For example, reaction [ATP + Deamino-NAD^+ + NH3 ⇔ AMP + Diphosphate + NAD^+] gives rise to both ATP-AMP (high *SCL*) and ATP-Diphosphate (low *SCL*) as reactant pairs of type 'ligase'.

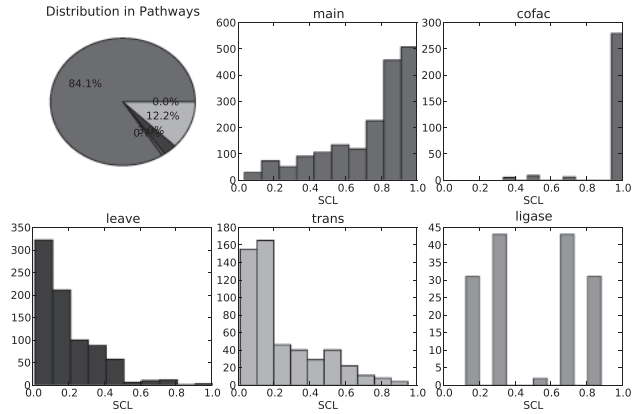

Fig. 7. Distribution of *SCL* values of the five RPair types. The pie chart shows the distribution of RPair types in the connections that are annotated in KEGG reference pathway. Colors in the pie chart correspond to the colors in the histograms. The histograms shows for a given RPair type its distribution of *SCL* values based on all connections that are annotated with RPairs in the metabolic network of *E.coli*.

4 DISCUSSION

In this article, we introduced the strength of chemical linkage, or *SCL*, as a criterion for pruning metabolic graphs. The use of the conserved chemical content according to the actual reaction mechanism in this work is in contrast to the previous approach where only structural similarity between the two reactants is considered (Rahman *et al.*, 2005). Although we also use the RPair database, unlike (Faust *et al.*, 2009), whose pruning depends on the presence/absence of RPairs or the certain classes of reactant pairs annotated from KEGG based on the type of enzyme that catalyze the reaction (Kotera *et al.*, 2004), we used only the information for molecule alignments. Such information can also be obtained from other sources, such as Crabtree *et al.* (2010).

We showed that the *SCL* criterion is biochemically intuitive and has power of selection for the conventional pathway connectivity when thresholded. False positive and false negative cases are caused mainly by improper quantification of chemical content as well as flaws in the data. The utility of using *SCL* on pruning the searching tree in pathway inference was evaluated. Biochemically meaningful pathways can be found by implementing a simple search program using the *SCL* criterion. Further, we compared several commonly used connectivity pruning heuristics and *ad hoc* methods, such as hub deletion and manual curation. We found that *SCL* values reflect the rationale behind these heuristics, yet the *SCL* is more objective, systematic and robust to annotation error. Many ambiguities of these heuristics are rooted in lacking an objective criterion and quantification of chemical content.

Note that although, we focus here only on graphs whose nodes are compounds, *SCL* can also be adapted in the assembly of networks whose nodes are reactions or reaction-derived entities such as enzyme class or genes. This is done by considering nodes for linking reactions, only ones that appear in sufficiently strong reactant pairs. One potential improvement to the quantification of chemical content is to partition every chemical compound into functionally independent groups. The amount of chemical content is measured in terms of the number of such functionally independent groups, instead of the absolute number of non-hydrogen atoms. Further, the partition of a compound into functionally independent groups is flexible yet objective, relying on at most a delineation of a set of specific pathway-related reactions. Two atoms in a compound are considered in the same group if they are linked by covalent bond(s) that does not break in all the chemical transformations under that delineation.

Funding: National Science Foundation (grant number CCF-0622037); National Library of Medicine (grant number R01LM009494); an Alfred P. Sloan Research Fellowship.

Conflict of Interest: none declared.

REFERENCES

Arita,M. (2005) Scale-freeness and biological networks. *J. Biochem.*, **138**, 1–4.

Blum,T. and Kohlbacher,O. (2008) Using atom mapping rules for an improved detection of relevant routes in weighted metabolic networks. *J. Comput. Biol.*, **15**, 565–576.

Crabtree,J.D. *et al.* (2010) An open-source java platform for automated reaction mapping. *J. Chem. Inform. Model.*, **50**, 1751–1756.

Croes,D. *et al.* (2006) Inferring meaningful pathways in weighted metabolic networks. *J. Mol. Biol.*, **356**, 222–236.

Deville,Y. *et al.* (2003) An overview of data models for the analysis of biochemical pathways. *Brief. Bioinform.*, **4**, 246–259.

Diaz-Mejia,J. *et al.* (2007) A network perspective on the evolution of metabolism by gene duplication. *Genome Biol.*, **8**, R26.

Ebenhh,O. *et al.* (2004) Structural analysis of expanding metabolic networks. *Genome Inform.*, **15**, 35–45.

Faust,K. *et al.* (2009) Metabolic pathfinding using rpair annotation. *J. Mol. Biol.*, **388**, 390–414.

Herrgard,M.J. *et al.* (2008) A consensus yeast metabolic network reconstruction obtained from a community approach to systems biology. *Nat. Biotechnol.*, **26**, 1155–1160.

Holme,P. (2009) Model validation of simple-graph representations of metabolism. *J. R. Soc. Inter.*, **6**, 1027–1034.

Jeong,H. *et al.* (2000) The large-scale organization of metabolic networks. *Nature*, **407**, 651–654.

Kanehisa,M. and Goto,S. (2000) KEGG: Kyoto Encyclopaedia of Genes and Genomes. *Nucleic Acids Res.*, **28**, 27–30.

Kotera,M. *et al.* (2004) RPAIR: a reactant-pair database representing chemical changes in enzymatic reactions. *Genome Inform.*, **15**, P062.

Lee,T. *et al.* (2006) Biowarehouse: a bioinformatics database warehouse toolkit. *BMC Bioinformatics*, **7**, 170.

Lemer,C. *et al.* (2004) The aMAZE LightBench: a web interface to a relational database of cellular processes. *Nucleic Acids Res.*, **32** (Suppl. 1), D443–D448.

Ma,H.W. and Zeng,A.P. (2003) Reconstruction of metabolic networks from genome data and analysis of their global structure for various organisms. *Bioinformatics*, **19**, 270–277.

Poolman,M.G. *et al.* (2006) Challenges to be faced in the reconstruction of metabolic networks from public databases. *IEEE Proc. Sys. Biol.*, **153**, 379–384.

Rahman,S.A. *et al.* (2005) Metabolic pathway analysis web service (Pathway Hunter Tool at CUBIC). *Bioinformatics*, **21**, 1189–1193.

Ravasz,E. *et al.* (2002) Hierarchical organization of modularity in metabolic networks. *Science*, **297**, 1551–1555.

Tanaka,R. (2005) Scale-rich metabolic networks. *Phys. Rev. Lett.*, **94**, 168101.

van Helden,J. *et al.* (2002) Graph-based analysis of metabolic networks. *Ernst Schering Res Found Workshop*, **38**, 245–274.

Wagner,A. and Fell,D.A. (2001) The small world inside large metabolic networks. *Proc. R. Soc. Lond. Ser. B Biol. Sci.*, **268**, 1803–1810.

Zhao,J. *et al.* (2006) Hierarchical modularity of nested bow-ties in metabolic networks. *BMC Bioinformatics*, **7**, 386.

Zhao,J. *et al.* (2007) Modular co-evolution of metabolic networks. *BMC Bioinformatics*, **8**, 311.

Zhu,D. and Qin,Z.S. (2005) Structural comparison of metabolic networks in selected single cell organisms. *BMC Bioinformatics*, **6**, 1471–2105.