Functional Similarities of Reaction Sets in Metabolic Pathways

Ferhat Ay
Computer and Information Science and Engineering
University of Florida, Gainesville, FL 32611
fay@cise.ufl.edu

Tamer Kahveci
Computer and Information Science and Engineering
University of Florida, Gainesville, FL 32611
tamer@cise.ufl.edu

ABSTRACT

Analyzing metabolic pathways by means of their steady states has proven to be accurate and efficient for practical purposes. The models such as elementary flux modes (EFMs) and extreme pathways (EPs) define the boundaries of the metabolic flux cone that is the set of all steady states of a pathway. However, the contributions of the subsets of pathway components in this flux cone so far has not been characterized mathematically. Also, the functional similarities of different component sets (e.g., sets of reactions) has not been expressed as a function of the steady states of metabolic pathways. Here, we aim to fill this gap by proposing a model that quantifies the impact of a set of components on the steady states of a pathway by using EFMs. At a high level, we model the impact of a given component set as the change in the flux cone when all the elements of that set are inhibited. Furthermore, given two sets of components from different pathways, we measure their functional similarity as the similarity of their impacts on corresponding pathways. Computing this functional similarity is a computationally challenging task as it requires finding the volumes of the intersection and the union of two polyhedral cones in high dimensional space. These volumes cannot be expressed in closed form. In this paper, we develop a novel method that first transforms the polyhedral cones to polytopes and then uses minimum enclosing balls to approximate this intersection efficiently. Our experiments on real metabolic pathways demonstrate that our method is of great use for both measuring the impacts of pathway components and identifying functionally similar component sets.

1. INTRODUCTION

In the last decades, significant amount of research has been done on identification and reconstruction of biological pathways such as gene regulatory [15, 26, 29], protein interaction [30] and metabolic pathways [3, 7]. This resulted in a wealth of interaction data which attracted attention of many computational biologists. Among these, metabolic pathways is an essential class that represents how compounds are transformed from one to another through metabolic reactions. Comparative analysis of these pathways is necessary to identify component sets that perform similar biological functions in different pathways. Identifying these similarities is of utmost importance for many applications such as drug target identification [4, 22], phylogeny reconstruction [8, 9], metabolic engineering [24] and biocommodity engineering [13]. Often homological and topological features of the pathways have been used for this purpose. However, the biological functions of two reaction sets can be equivalent even when their size, connectivity, intermediate products and catalyzing enzymes are different [1, 2]. Therefore, neither the topological features (e.g., centrality) nor the homological similarities (e.g, enzymes similarity, compound similarity) can provide sufficient information for identifying such functional similarities.

Other than homology and topology based methods, another common way to analyze metabolic pathways is to identify their metabolic capabilities in terms of their steady states. A *steady state* of a pathway is a feasible flux distribution that represents a possible long term outcome of that pathway. The steady states of a pathway determine the set of functions it can perform. These states define a polyhedral cone in a high dimensional space where fluxes of the pathway corresponds to dimensions. Figure 1 depicts an example of a metabolic flux cone of a hypothetical pathway with three fluxes.

A number of models have been proposed to analyze the metabolic capabilities of a pathway, such as elementary flux modes (EFMs) [20], extreme pathways (EPs) [25], flux balance analysis (FBA) [23] and minimal metabolic behaviors (MMBs) [12]. All these models are different interpretations of the flux cone. They compute this flux cone using the stoichiometric constraints of the reactions of the underlying pathway in terms of its extreme rays emanating from the origin of the high dimensional flux space. We elaborate on these models and some important properties of the metabolic flux cone in Section 2.

At this point, we explain the concept of EFM on a real example as it is an integral part of the rest of the paper.

EXAMPLE 1. *Figure 2 illustrates a metabolic pathway with 11 reactions and 7 compounds. Schuster* et al. *[18] identified six meaningful EFMs of this pathway. Each EFM is an 11 dimensional vector (i.e., one dimension per reaction). These EFMs are* e_1 *=[1 0 1 1 0 1 1 0 0 1 0], e_2 =[1 0 1 1 0 1 1 1 0 0 1], e_3 =[0 1 -1 0 1 -1 0 1 1 -1 0], e_4 =[0 1 -1 0 1 -1 0 0 1 0 -1], e_5 =[0 0 0 0 0 0 0 1 0 -1 1] and e_6 =[0 0 0 0 0 0 1 0 1 1 -1]. Here, each e_i represents an EFM and 0, 1 and -1s denote flux values for the reactions from r_1 (leftmost) to r_{11} (rightmost). For reversible reactions, such as r_3, r_6, r_{10} and r_{11}, -1 indicates that the reaction direction is from right to left (i.e., inverse reaction).*

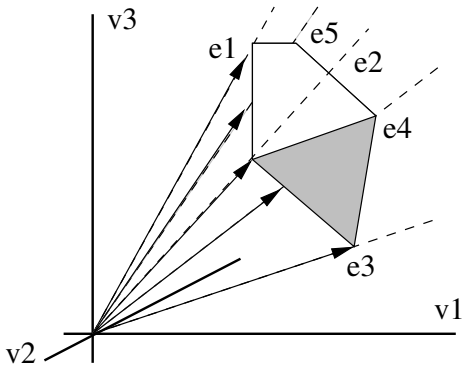

Figure 1: An illustration of the impact of a reaction set R on pathway $\mathcal{P}$. $\mathcal{P}$ has five elementary flux modes (EFMs) denoted by $e_1, \cdots, e_5$. After inhibiting all reactions in R, only four of these remain feasible, namely e_1, e_2, e_4 and e_5. The impact of R on $\mathcal{P}$ is the change in the flux cone represented by the shaded polyhedral cone bounded by e_2, e_3 and e_4.

Biologically, e_1 corresponds to glycolytic pathway; e_2 represents the anaplerotic path from glucose (Glu) to oxaloacetate (OAA); e_3 and e_4 are gluconeogenesis pathways starting from pyruvate (Pyr) production and oxaloacetate respectively; e_5 represents the conversion from pyruvate to oxaloacetate; and e_6 is on the route to synthesis of several amino acids in glucose-poor environment [18]. Each of these EFMs are non-decomposable (i.e., minimal) in the sense that if any of nonzero flux values for e_i is set to zero, then e_i is not an EFM anymore. □

Each component (i.e., enzyme, compound, reaction) of a pathway can contribute to the set of possible steady states of that pathway. The above models (EFMs, EPs, etc.) are key tools in examining the overall steady state behavior. They, however, do not provide any information on how much each component or a component set contributes to these steady states. Analyzing and quantifying the impacts of these components provide a better understanding of the pathway. When the reactions of a pathway are of interest, several existing approaches measure the impact of a reaction as the number of its neighbors (centrality) [14], the number of compounds it uniquely produces or consumes (UP/UC) [17] and the number of EFMs that it participates in (participation) [5, 16]. However, none of these methods characterize the biological functions of a reaction set as a function of the steady states of its pathway. Characterization of the impact of a reaction set on the overall metabolic capabilities of a pathway is an exciting task as it is of great use to model the outcomes of different perturbations for metabolic engineering applications.

In this paper, we develop a systematic way to characterize and compute the functional similarity between two given reaction sets of metabolic pathways. We achieve this goal in two steps. We elaborate on these steps later in this section and in the rest of the paper.

- **Step 1.** Given a metabolic pathway $\mathcal{P}$ and a subset R of its reactions, we calculate the *impact* of R on the steady states of $\mathcal{P}$.
- **Step 2.** Given pathways $\mathcal{P}$, $\bar{\mathcal{P}}$ and subsets of their reactions R and $\bar{R}$ respectively, we compute the *functional similarity* of R and $\bar{R}$ in terms of their impacts on $\mathcal{P}$ and $\bar{\mathcal{P}}$ respectively.

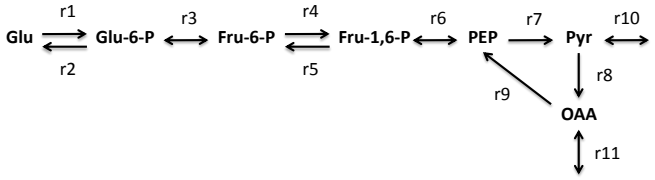

Figure 2: Simplified metabolic pathway of *Glycolysis and Gluconeogenesis* taken from Schuster *et al.* [18]. Glucose is assumed to be an external compound. This pathway contains 11 reactions denoted by r1, r2, $\cdots$, r11. Here, r10 and r11 are exchange reactions.

At a high level, we model the impact of a reaction set (R) in terms of the steady states of its pathway ($\mathcal{P}$). Specifically, we compute it as the portion of the flux cone of the original pathway that cannot be achieved without the reactions of R. To do this, we inhibit all the reactions of R and compute the new flux cone. The difference between the old and new flux cones is the impact of R on pathway $\mathcal{P}$. Figure 1 demonstrates the notion of impact on a hypothetical example. The grey portion illustrates the flux values that are reachable in steady state before inhibiting the reactions in R but not reachable after inhibiting them. If there is a steady state flux distribution that corresponds to a key biological function (e.g., optimal production of a certain compound that is essential for the survival) in this grey area, then the perturbation harms this function as this state could not be achieved any more. Intuitively, this suggests that R has an important role in that specific function.

EXAMPLE 2. *Here, we want to illustrate on a real example how the impacts of different components of a pathway can be significantly different. We use the same pathway as in Example 1 (see Figure 2). Consider the impacts of two irreversible reactions r_1 and r_7. When we inhibit r_1, the two EFMs e_1 and e_2 are no more feasible. This implies that the perturbed pathway is neither capable of using glycolytic pathway nor the anaplerotic path from glucose to oxaloacetate. On the other hand, it can still synthesize necessary amino acids without using glucose through e_6. However, if we inhibit r_7, it will render the EFM e_6 useless in addition to e_1 and e_2. In this case, the perturbed pathway cannot synthesize necessary amino acids neither with nor without using glucose. Hence, the effect of inhibiting r_7 is more likely to be lethal compared to r_1. We can see how this translates into our impact model by a closer look at EFMs e_1, e_2 and e_6. e_1 and e_2 are different for only 3 dimensions whereas e_6 differs from e_1 and e_2 in 6 and 7 dimensions respectively. This implies that the shift in the flux cone by r_7 which exterminates e_6 on top of e_1 and e_2 will have effect on 7 more dimensions than the shift caused by r_1. Therefore, in our model, the impact of r_7 on this pathway will be significantly larger than the impact of r_1. We discuss how the impacts of reaction subsets computed in this manner can be used to predict essential reactions in Section 5.* □

Utilizing the impacts as modeled above, we characterize the functional similarity between two reaction sets from potentially different pathways. We do this by first lifting the flux cones of both pathways to a higher dimensional flux space that is the union of flux spaces of these pathways. The term *lifting* denotes the transformation of a geometric object to a higher dimensional space. We explain why we use the union of flux spaces in Section 3. Here, it is necessary to know that this process preserves EFMs of the flux

cones. In other words, by lifting an EFM of an original pathway to the new flux space, we get an EFM of that pathway extended with additional fluxes. We, then, represent the functional similarity of two reaction subsets as the ratio of the volume of the intersection of the regions corresponding to their impacts on their pathways in the new flux space to that of their union.

Functional similarity modeled in this fashion is an accurate indicator of the similarities of biological roles of different reaction sets as it represents the ratio of the common steady states that are not possible without the contribution of these reaction sets in their pathways. The next step is to devise an efficient method that quantifies this notion. Here, we propose a novel method that utilizes EFMs of metabolic pathways and converts functional similarity calculation into a high dimensional geometric problem. The set of EFMs generates the space of steady state flux distributions for a pathway. In case of metabolic pathways with only irreversible fluxes, the set of EFMs that delineates the flux cone is unique. Additionally, the flux cone is polyhedral in this case and therefore it is finitely generated [12]. In other words, for a given pathway $\mathcal{P}$, its flux cone $C(\mathcal{P})$ is uniquely defined by the set of EFMs $E(\mathcal{P})$=$\{e_1, e_2, ..., e_n\}$. To compute the EFMs of a pathway, in our method, we use an existing implementation called *Metatool* [20]. First, we compute the EFMs of the query pathways, say $\mathcal{P}$ and $\bar{\mathcal{P}}$, by using Metatool. Each set of EFMs generate the *original flux cone* of the corresponding pathway. Then, for a given reaction subset R of a pathway $\mathcal{P}$, we first remove all the reactions of R from this pathway. Biologically, this corresponds to inhibiting the enzymes that catalyze these reactions. Then, we compute the region that represents the impact of R on $\mathcal{P}$. Due to the non-decomposability property of EFMs, the impact can also be represented as a polyhedral cone and has a set of EFMs that defines it. We denote this set by $E(\mathcal{P} - R)$. Similarly, we compute $E(\bar{\mathcal{P}} - \bar{R})$ for the impact of $\bar{R}$ on $\bar{\mathcal{P}}$.

At this point, we have two sets of EFMs in hand that define two polyhedral cones. Now, the problem of finding the functional similarity of R and $\bar{R}$ is equivalent to purely geometric problem of finding the intersection and the union of these two cones. This, however, is computationally difficult as there is no closed form solution to this problem. To tackle this, we first transform the polyhedral cones into polytopes by taking their intersections with a hyperplane in the positive quadrant (i.e., all flux values are non-negative). We show that this transformation preserves the ratio between the intersection and the union of the flux cones. Then, we compute the intersection of these polytopes as the intersection of their minimum enclosing balls (MEB). Finding MEB is a well-studied computational geometry problem and efficient solutions exist for problems with up to a few hundred dimensions. We use an efficient implementation of Fischer *et al.* [6] to compute MEBs. Finally, we normalize this intersection and convert it to a functional similarity score. We elaborate on each step of our method in Section 4.

Our experiments on real metabolic pathways show that the functional similarity we propose here is of important use for identifying reaction sets that perform biologically similar functions. Moreover, we observe that our definition of impact can provide biologically and statistically significant predictions of essential reaction sets.

The following summarizes our technical contributions:

- We build a mathematical model for the impact of a set of reactions on the steady state space of a metabolic pathway.
- We characterize the functional similarity of two sets of reactions in terms of their impacts on the metabolic capabilities of their corresponding pathways.
- We develop an efficient method to compute the functional similarity of component sets from different pathways.

The rest of the paper is organized as follows. Section 2 summarizes existing approaches for analyzing metabolic pathways. Section 3 describes how we model the impacts of reaction sets as well as the functional similarity between them. Section 4 presents our method that computes the functional similarity score. We report our experimental results in Section 5. Section 6 briefly concludes the paper.

2. BACKGROUND

There are several existing methods that model metabolic pathways by means of their steady states. Here, we briefly describe four commonly used models, namely elementary flux modes (EFMs) [20], extreme pathways (EPs) [25], flux balance analysis (FBA) [23] and minimal metabolic behaviors (MMBs) [12]. FBA differs from the other three models as it aims to find a steady state that maximizes a given objective function. All the other models define the flux cone of a pathway that contains all of its possible steady states. In order to use FBA we need to know the objective of the cells that we are analyzing. This objective is often maximizing biomass production in single cell organisms. However, in complex organisms different cells have different objectives and it is usually not possible to represent these objectives as a well-defined mathematical function. Furthermore, FBA does not identify suboptimal states which provide better understanding of the steady state flux distribution for perturbed pathways [21].

A relatively newer model named MMBs defines the flux cone using a constraint-based approach. This method uses an outer description of the flux cone which is more compact compared to an inner description. Instead of finding the extreme rays of a flux cone, MMBs identifies the sets of constraints that define the minimal proper faces of the cone. These sets of constraints are all subsets of nonnegativity constraints of irreversible reactions and they form the MMBs of the pathways. MMBs together with reversible metabolic space (RMS) uniquely determine the flux cone. Also, this model provides a test for determining whether a given flux distribution belongs to the cone. However, it does not provide a means of generating all the steady state flux distributions.

Two popular and closely related models that use an inner description of a flux cone are EFMs and EPs. An *EFM* is a minimal set of reactions that can operate at steady state with all irreversible reactions having nonnegative rates. An *EP* is an EFM that corresponds to a steady state which defines an extreme ray of the flux cone. For any metabolic pathway, the set of EPs is a subset of the set of EFMs and both of these sets generate the flux cone. Klamt *et al.* [11] point out that these two sets are often equal for realistic applications. In fact, they state that if all exchange reactions in a pathway are irreversible then the sets of EFMs and EPs coincide. Another very useful property of EFMs is that reconfiguring a pathway by splitting up all its reversible reactions into two irreversible reactions does not change the set of its EFMs. By using these two properties, for a pathway with no irreversible exchange reactions, we always get a flux cone in the first quadrant of high dimensional flux space that is generated by a set of extreme rays emanating from the origin. The set of these extreme rays is equivalent to the set of

EFMs as well as the set of EPs. The convex combinations of the elements of this set immediately generate all possible steady state flux distributions.

In our method, we use EFMs to characterize the flux cones. One reason behind that is, we need a model that can represent all the metabolic capabilities of a pathway rather than only the optimal steady state. Another reason is that, metabolic pathways we consider have no irreversible exchange fluxes and the set of EFMs coincide with the set of EPs. Hence, it is only to avoid confusion we use the name EFMs for that set. Furthermore, the non-decomposability property of EFMs (i.e., no non-empty, proper subset of the reactions of an EFM can lead to a steady state) is useful when considering different perturbations on the pathway. By using the set of EFMs of a pathway and analyzing the effects of different perturbations on this set, we devise a method that computes the impacts and the functional similarities of different reaction sets.

3. MODELING FUNCTIONAL SIMILARITY

In this section, we describe how we mathematically interpret the impact of a set of reactions on the metabolic capabilities of a given pathway. Also, we discuss how we model the functional similarity between different reaction sets.

Consider a metabolic pathway $\mathcal{P}$ with n reactions $U(\mathcal{P}) = \{r_1, r_2, \ldots, r_n\}$ and d fluxes $F(\mathcal{P}) = \{f_1, f_2, \ldots, f_d\}$. Let S be the stoichiometric matrix of $\mathcal{P}$ that consists of one column for each flux $f_i \in F(\mathcal{P})$ which identifies the input and output compounds of that flux. Also, let $v = [v_1, v_2, \cdots, v_d]^T$ be a flux vector that represents the state v in which each v_i is the value realized by flux f_i. Then, $S \cdot v$ computes the change in each flux from v to the next state. The solution space of the equation system $S \cdot v = 0$ is the set of all states in which the flux values stabilize. This set has infinite number of solutions which form a cone in high dimensional space. We can write this flux cone as the spanning set of the EFMs of the pathway. If $E(\mathcal{P}) = \{e_1, e_2, \ldots, e_t\}$ is the set of EFMs of pathway $\mathcal{P}$, then the flux cone $C(\mathcal{P})$ is:

$$C(\mathcal{P}) = span(E(\mathcal{P})) = \{v \mid v = \sum_{i=1}^{t} c_i e_i, c_i \geq 0\} \quad (1)$$

We want to consider the impact of a reaction set R on the original flux cone of pathway $\mathcal{P}$. In other words, we want to obtain a mathematical expression for the change in $C(\mathcal{P})$ when all the reactions in R are inhibited. We represent the pathway perturbed in this manner with $\mathcal{P} - R$. Also, we denote the flux cone and the EFMs of $\mathcal{P} - R$ with $C(\mathcal{P} - R)$ and $E(\mathcal{P} - R)$ respectively. Below is a formal statement of impact:

DEFINITION 1. (IMPACT) *Let $\mathcal{P}$ be a metabolic pathway with reaction set $U(\mathcal{P})$. The impact of a reaction subset $R \subseteq U(\mathcal{P})$ is:*

$$Impact(R, \mathcal{P}) = Span(E(\mathcal{P})) - Span(E(\mathcal{P} - R))$$

where $E(\mathcal{P})$ and $E(\mathcal{P} - R)$ are the set of EFMs of $\mathcal{P}$ and $\mathcal{P} - R$ respectively. □

Inhibiting reactions can only shrink the flux cone. Furthermore, from the non-decomposability property of EFMs, we know that all EFMs of $\mathcal{P} - R$ are also EFMs of $\mathcal{P}$. That is, $E(\mathcal{P} - R) \subseteq E(\mathcal{P})$. As a result, for the metabolic pathways with only irreversible fluxes Impact(R, $\mathcal{P}$) is also a polyhedral cone and has a set of EFMs that generates it. When reversible fluxes are present, we split them into two different fluxes in opposite directions and this property still holds. The set of EFMs that define Impact(R, $\mathcal{P}$) is a subset of the EFMs of the original pathway and we can construct it by checking for each $e_i \in E(\mathcal{P})$ whether it has any nonzero flux value for a reaction that is an element of R. If e_i has zero entries for all the reactions of R, then e_i is an EFM of Impact(R, $\mathcal{P}$). The set of these EFMs define the flux cone of this impact. Using this model of impact, we are now ready to present our characterization of functional similarity between different reaction sets.

Let $\mathcal{P}$, $\bar{\mathcal{P}}$ be two metabolic pathways. Clearly, these two pathways can have different fluxes as well as common ones. If all the fluxes are not common to both pathways, then $C(\mathcal{P})$ and $C(\bar{\mathcal{P}})$ lie on different spaces. To be able to compare these two cones and the impacts of perturbations on them, we need them to be in the same high dimensional flux space. One way to do it is to take only the common fluxes and project the sets of EFMs that generate the flux cones. However, again by non-decomposability property of EFMs, the projected EFMs may not be feasible anymore. This results in erroneous steady states which do not reflect the actual capabilities of the corresponding pathway. In order to bring the two flux cones to the same flux space without affecting the steady states they represent, we lift them to a higher dimensional space defined by the union of the fluxes of two pathways. We do this by simply extending the EFMs with zeros for non-common fluxes. The lifting process guarantees that the EFMs of both pathways stay as EFMs in the new flux space and the new flux cones reflect the metabolic capabilities correctly. We use the notation Impact(R, $\mathcal{P}|\ \bar{\mathcal{P}}$) to represent the impact of R on pathway $\mathcal{P}$ in the new flux space that is the union of the fluxes of $\mathcal{P}$ and $\bar{\mathcal{P}}$. We define the functional similarity between the reaction sets R of $\mathcal{P}$ and $\bar{R}$ of $\bar{\mathcal{P}}$ as the unexpectedness of the size of the intersections of their impacts in this new flux space. Formal definition is as follows.

DEFINITION 2. (FUNCTIONAL SIMILARITY) *Let $\mathcal{P}$ and $\bar{\mathcal{P}}$ be two metabolic pathways. Let R, $\bar{R}$ be two given subsets of the reactions of $\mathcal{P}$ and $\bar{\mathcal{P}}$ respectively. Also, let Impact(R, $\mathcal{P}|\ \bar{\mathcal{P}}$) and Impact($\bar{R}$, $\bar{\mathcal{P}}|\ \mathcal{P}$) denote the impacts of R and $\bar{R}$ in the flux space defined by the union of the fluxes of $\mathcal{P}$ and $\bar{\mathcal{P}}$ respectively. If s is an arbitrary flux distribution in this space, then we define the functional similarity of R and $\bar{R}$ as:*

$$Sim(R, \bar{R}|\mathcal{P}, \bar{\mathcal{P}}) = -\log\Big(1 - Pr\big(s \in Impact(R, \mathcal{P}|\bar{\mathcal{P}}) \wedge s \in Impact(\bar{R}, \bar{\mathcal{P}}|\mathcal{P}) \mid s \in Impact(R, \mathcal{P}|\bar{\mathcal{P}}) \vee s \in Impact(\bar{R}, \bar{\mathcal{P}}|\mathcal{P})\big)\Big)$$

□

Here, the symbols $\wedge$ and $\vee$ denote logical "AND" and "OR" operators respectively. Intuitively, the above definition states that two reaction sets from different pathways serve similar functions if the sets of steady states they contribute to their own pathways have a significant intersection. The probability value (i.e., Pr()) increases linearly in the interval [0, 1] with the ratio of common steady states of the impacts of R and $\bar{R}$. Its value becomes 1 when two impacts are identical.

Computation of the functional similarity as formulated above is a nontrivial problem. It requires finding the intersection and the

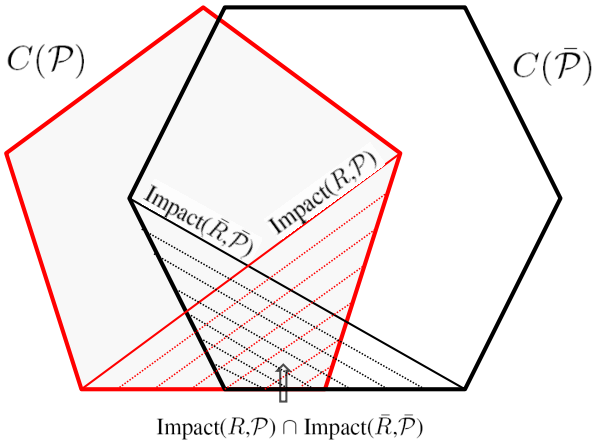

Figure 3: ***Pictorial description of functional similarity.*** **The pentagon on the left is a cross section of the flux cone of pathway $\mathcal{P}$. Similarly, the hexagon on the right represents the flux cone for $\bar{\mathcal{P}}$. The dashed areas denote the impacts of reaction sets R on $\mathcal{P}$ and $\bar{R}$ on $\bar{\mathcal{P}}$. Functional similarity of these impacts is determined by their intersection that is pointed by the arrow.**

union of two polyhedral cones in a high dimensional space. Neither of these problems has a closed form solution. In the next section, we propose an efficient method that allows us to compute the functional similarities of reactions sets for real sized pathways. In Section 5, we present experimental results that illustrate how our method performs on real metabolic pathways.

4. COMPUTING FUNCTIONAL SIMILARITY

Computing the functional similarity between two reaction sets (R of $\mathcal{P}$ and $\bar{R}$ of $\bar{\mathcal{P}}$) requires solving several challenging problems. First of all, we need to identify the impacts of both reaction sets on their pathways. Then, we need to find a high dimensional flux space in which EFMs of original pathways are preserved. In this new flux space, we need to compute the hyper-volume of the intersection of two polyhedral cones as well as the hyper-volume of their union. After solving all these problems, we can calculate the functional similarity between R and $\bar{R}$ in terms of their impacts on the metabolic capabilities of their pathways.

Algorithm 1 outlines our solution. Before going into details of each step of this algorithm, we explain our solution on a visual example. Figure 3 illustrates the cross-sections of two flux cones belonging to two hypothetical pathways both in three dimensional flux space. The cross-sections are obtained by intersecting $C(\mathcal{P})$ and $C(\bar{\mathcal{P}})$ with a two dimensional plane. Here, $C(\mathcal{P})$ and $C(\bar{\mathcal{P}})$ are generated by 5 and 6 EFMs respectively. The dashed areas of this cross-section denote the impacts of R and $\bar{R}$ on their pathways as labeled. Both Impact(R, $\mathcal{P}$| $\bar{\mathcal{P}}$) and Impact($\bar{R}$, $\bar{\mathcal{P}}$| $\mathcal{P}$) have 3 EFMs that generate them. The double dashed area represents the intersection of these two impacts. The bigger this intersection is, the more similar the impacts, hence the functions, of R and $\bar{R}$. In order for functional similarities of different reaction set pairs to be comparable, we normalize the size of this intersection by the size of the union of the impacts. In other words, what we are aiming to compute in Figure 3 is the ratio of the size of double dashed region to the size of the union of all dashed regions. For a three dimensional flux space, the problem is to find the intersection and the union of two areas in two dimensions. In general, if two flux cones lie in a d dimensional space the we need to find the intersection of two $d-1$ dimensional polytopes.

Next, we elaborate on each step of Algorithm 1.

Algorithm 1 Computing Functional Similarity

Input: Two reaction sets R of $\mathcal{P}$ and $\bar{R}$ of $\bar{\mathcal{P}}$
Output: Sim(R, $\bar{R}$ | $\mathcal{P}$, $\bar{\mathcal{P}}$)

1. Compute the EFMs of $C(\mathcal{P})$ and $C(\bar{\mathcal{P}})$ (i.e., $E(\mathcal{P})$, $E(\bar{\mathcal{P}})$).
2. Find the EFMs of Impact(R,$\mathcal{P}$) and Impact($\bar{R}$,$\bar{\mathcal{P}}$).
3. Compute $-log\Big(1-\big(\frac{\text{Impact}(R,\mathcal{P}|\bar{\mathcal{P}})\cap\text{Impact}(\bar{R},\bar{\mathcal{P}}|\mathcal{P})}{\text{Impact}(R,\mathcal{P}|\bar{\mathcal{P}})\cup\text{Impact}(\bar{R},\bar{\mathcal{P}}|\mathcal{P})}\big)\Big)$.
   I. Lift the EFMs of impacts into a higher dimensional flux space.
   II. Transform the polyhedral cones in this new flux space into polytopes by taking their intersection with a hyperplane.
   III. Bound the hyper-volumes of the created polytopes by finding their minimum enclosing balls (MEBs).
   IV. Find the hyper-volumes of the intersection and union of these MEBs and return the similarity score.

4.1 Finding the EFMs of Metabolic Pathways

In order to compute functional similarity of two reaction sets as a function of the steady states of the pathways, we first need to find the set of EFMs that generates the metabolic flux cone of a pathway. Identification of EFMs is a well-studied problem and a number of algorithms as well as their implementations are available [19, 20, 27, 28]. In this paper, we use a recent implementation of von Kamp *et al.* called *Metatool* [20]. We choose this tool as its an efficient implementation and is commonly used in the literature. Metatool uses the stoichiometric constraints and the reversibility of the reactions of a pathway in order to compute its EFMs. However, finding EFMs is a computationally expensive problem and even the most efficient algorithms cannot scale for pathways with more than 100 reactions. Therefore, this part becomes the bottleneck when we want to compute functional similarities of reaction sets of very large pathways.

4.2 Extracting EFMs of Impacts

After finding the EFMs of the original pathways $\mathcal{P}$ and $\bar{\mathcal{P}}$, the next step is to compute the impacts of reactions sets R and $\bar{R}$ on the corresponding flux cones of $\mathcal{P}$ and $\bar{\mathcal{P}}$ in terms of the EFMs. In other words, we want to find the set of EFMs $E(R)$ ($E(\bar{R})$) that represents the change in $C(\mathcal{P})$ ($C(\bar{\mathcal{P}})$) when all the reactions in R ($\bar{R}$) are inhibited.

We discuss how we compute the change in $C(\mathcal{P})$ due to inhibition of R next. The computation of that for $C(\bar{\mathcal{P}})$ and $\bar{R}$ is similar. As we explained in Section 3, the impact of a reaction set defines a polyhedral cone and the set of its EFMs is a subset of the EFMs of the original pathway. We construct this set of EFMs by checking for each EFM of the original pathway whether it is still feasible after inhibition of the corresponding reaction set. If an EFM does not remain feasible after inhibiting the reactions of R, then it is a member of the generating set of the polyhedral cone that represents the impact of R on $\mathcal{P}$.

Here, we want to demonstrate on a hypothetical example how we extract the EFMs of the impact of a reaction set R of a given pathway $\mathcal{P}$. Let $U(\mathcal{P})$ =$\{r_1,r_2,r_3\}$, $F(\mathcal{P})$ =$\{f_1,f_2,f_3,f_4\}$, $E(\mathcal{P})$ = $\{e_1, e_2, e_3, e_4, e_5\}$ denote the set of reactions, set of fluxes and the set of EFMs of pathway $\mathcal{P}$ respectively. Also, let r_1, r_2 be irreversible reactions corresponding to fluxes f_1, f_2 respectively, and r_3 be a reversible reaction that is split into two irreversible fluxes f_3 and f_4. Let the flux values of five EFMs of this pathway be given as:

	f_1	f_2	f_3	f_4
e_1	1	0	0	1
e_2	0	1	0	1
e_3	0	1	1	0
e_4	1	0	1	0
e_5	1	1	0	0

Consider the impact of the reaction r_1. The EFMs e_1, e_4 and e_5 are not feasible after inhibiting r_1 since f_1 has a nonzero flux value at these EFMs. Therefore, the set $E(\{r_1\}) = \{e_1, e_4, e_5\}$ generates the polyhedral cone that represent the impact of r_1 on $\mathcal{P}$. For r_3 there are two corresponding fluxes f_3 and f_4. Hence, the impact of inhibiting r_3 is generated by the EFMs that have either nonzero values for either f_3 or f_4 (i.e., $E(\{r_3\}) = \{e_1, e_2, e_3, e_4\}$).

Extracting the EFMs of the impacts in this manner is more efficient than computing them from scratch for every different reaction set. This important reduction in computational cost of our method is due to the non-decomposability property of EFMs. These EFMs generate all the steady states that the reaction set in consideration contributes to its metabolic pathway. Next, we describe how we use these steady states to define the functional similarity between two such reaction sets.

4.3 Calculating the Similarity Score

Once we extract the EFMs of the impacts of R and $\bar{R}$, it is conceptually easy to describe how we calculate the functional similarity between them. We measure the similarity as the unexpectedness of the size of the intersection of the cones representing these impacts. However, finding this similarity requires solving several computationally challenging problems. In the following, we summarize the steps we take in our solution to tackle these problems.

STEP I. As two pathways $\mathcal{P}$ and $\bar{\mathcal{P}}$ can have different fluxes, we first need to find a common flux space in which the metabolic flux cones of these pathways are comparable. At this point the first attempt would be to find the fluxes that are common to both pathways and project flux cones onto that space. As we explained in Section 3, this approach results in incorrect EFMs that do not generate the correct flux cones of the pathways. Therefore, we use lifting instead of projection that extends EFMs with zero entries to lift the flux cones into a higher dimensional space.

Here, we explain how the flux cones are lifted to the new flux space on a hypothetical example. Let $\mathcal{P}$, $\bar{\mathcal{P}}$ be two metabolic pathways with fluxes $F(\mathcal{P})$ =$\{f_1,f_2,f_3,f_4\}$ and $F(\bar{\mathcal{P}})$ =$\{f_1,f_3,f_5\}$. We denote new flux space that is the union of $F(\mathcal{P})$ and $F(\bar{\mathcal{P}})$ as $F(\mathcal{P}+\bar{\mathcal{P}})$= $\{f_1,f_2,f_3,f_4,f_5\}$. Let $\{e_1$= [1010], e_2= [0011], e_3 =[0100]$\}$ be the set of EFMs that generate the cone representing the impact of R on $\mathcal{P}$. We lift this cone to the flux space of $F(\mathcal{P}+\bar{\mathcal{P}})$ by adding zeros for f_5 to each EFM. The new set $\{e_1^{'}$= [10100], $e_2^{'}$= [00110], $e_3^{'}$ =[01000]$\}$ is the set of EFMs that generate the impact of R in the new flux space. Similarly, we can lift the impact of $\bar{R}$ on $\bar{\mathcal{P}}$ to this new space. The impacts of any two reaction sets from $\mathcal{P}$ and $\bar{\mathcal{P}}$ are comparable in the flux space of $F(\mathcal{P}+\bar{\mathcal{P}})$.

STEP II. Now, we have two polyhedral cones in the same flux space each representing the impact of a reaction set on its pathway. What we need at this point is the ratio of the hyper-volume of the intersection of these two cones to that of their union. As we mentioned earlier, there is no closed form expression for either of these hyper-volumes. In order to avoid computing these hyper-volumes directly, we transform the polyhedral cones to polytopes in our high dimensional flux space. We do this by intersecting the cones with a hyperplane in the positive quadrant. If d is the dimension of the flux space $F(\mathcal{P}+\bar{\mathcal{P}})$, then the intersection of a $d-1$ dimensional hyperplane with a polyhedral cone is a polytope in d dimensions. By construction, this polytope is convex and is easier to deal with compared to a polyhedral cone.

This transformation clearly changes the hyper-volumes of the regions that correspond to the impacts of different reaction sets. However, what we aim to find is neither these hyper-volumes nor the exact hyper-volumes of the intersection and union of these impacts. Instead, our goal is to calculate the ratio of the hyper-volume of the intersection to that of the union (See Algorithm 1). Theorem 1 proves that our transformation of the flux cones into polytopes preserves this ratio.

LEMMA 1. *Let $\vec{e}$ be an EFM of pathway $\mathcal{P}$ with d fluxes. Also, let $S : \vec{n} \cdot \vec{x} + c = 0$ denote a hyperplane where $\vec{n}$ and $\vec{x}$ are both d dimensional vectors representing the normal and a point on this hyperplane respectively. If $\vec{e}$ intersects S than this point is of the form $t\,\vec{e}$ for some $t \in \mathbb{R}$. Solving for t we get the intersection of the EFM $\vec{e}$ with hyperplane S as*

$$T_S(\vec{e}) = -\frac{c}{\vec{n} \cdot \vec{e}}\ \vec{e}$$

Proof: Omitted. □

THEOREM 1. *Let $C(R)$, $C(\bar{R})$ be two polyhedral cones in d dimensional space that are generated by the EFMs $E(R) = \{e_1, \ldots, e_a\}$ and $E(\bar{R}) = \{\bar{e_1}, \ldots, \bar{e_b}\}$ respectively. Let S be any hyperplane in this space and $H_S(R) = \{T_S(e_1), \ldots, T_S(e_a)\}$, $H_S(\bar{R}) = \{T_S(\bar{e_1}), \ldots, T_S(\bar{e_b})\}$ denote the polytopes created by the intersection of $C(R)$ and $C(\bar{R})$ with S respectively. Then,*

$$\frac{Volume(C(R) \cap C(\bar{R}))}{Volume(C(R) \cup C(\bar{R}))} = \frac{Volume(H_S(R) \cap H_S(\bar{R}))}{Volume(H_S(R) \cup H_S(\bar{R}))}$$

Proof: Let e_i be an EFM of $E(R)$. Consider all the hyperplanes S' that are parallel to S. The normal vector for any S' will be the same as the normal of S. This implies that the intersection point of e_i with S (i.e., $T_S(e_i)$) is a constant multiple (t) of its intersection point with S' (i.e., $T_{S'}(e_i)$). This constant is same for each e_i as $S' \| S$ and $T_S(e_i)$ = t $T_{S'}(e_i)$ for all $i \in [1, a]$. The intersection points of $E(R)$ with a hyperplane define the corners of a convex polytope on that hyperplane. Therefore, $H_S(R)$ and $H_{S'}(R)$ denote two convex polytopes that are scaled versions of

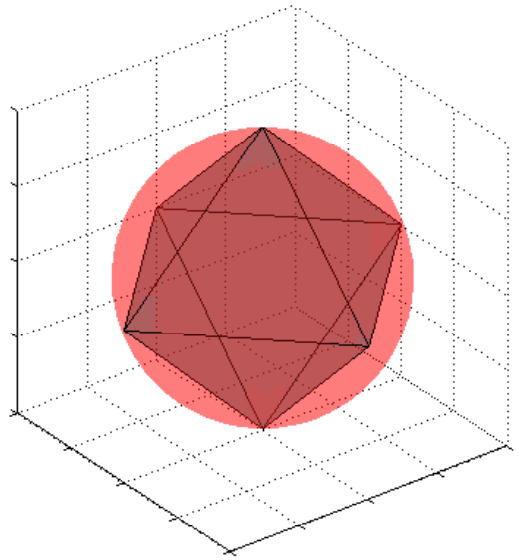

Figure 4: ***Minimum enclosing ball (MEB) of a polytope in three dimensional space.***

each other by a factor of t. Taking the ratio of the hyper-volumes of two such polytopes will cancel out the scaling factor. Hence,

$$\frac{Volume(H_S(R) \cap H_S(\bar{R}))}{Volume(H_S(R) \cup H_S(\bar{R}))} = \frac{Volume(H_{S'}(R) \cap H_{S'}(\bar{R}))}{Volume(H_{S'}(R) \cup H_{S'}(\bar{R}))}$$

when $S' \parallel S$.

Now, if we do integration over all S' that are parallel to S for the volumes of the polyhedral cones $C(R)$ and $C(\bar{R})$, we get:

$$Volume(C(R)) = \int_{S' \parallel S} Volume(H_{S'}(R))\, dS'$$
$$Volume(C(\bar{R})) = \int_{S' \parallel S} Volume(H_{S'}(\bar{R}))\, dS'$$

Using the above formulations:

$$Volume(C(R) \cap C(\bar{R})) =$$
$$\int_{S' \parallel S} \left[Volume(H_{S'}(R)) \cap Volume(H_{S'}(\bar{R}))\right] dS'$$
$$\int_{S' \parallel S} Volume(H_{S'}(R) \cap H_{S'}(\bar{R}))dS'$$

Computing $Volume(C(R) \cup C(\bar{R}))$ similarly and taking the ratio:

$$\frac{Volume(C(R) \cap C(\bar{R}))}{Volume(C(R) \cup C(\bar{R}))} = \int_{S' \parallel S} \frac{Volume(H_{S'}(R) \cap H_{S'}(\bar{R}))}{Volume(H_{S'}(R) \cup H_{S'}(\bar{R}))} dS'$$

$$= \int_{S'} \frac{tVolume(H_S(R) \cap H_S(\bar{R}))}{tVolume(H_S(R) \cup H_S(\bar{R}))} dS' = \frac{Volume(H_S(R) \cap H_S(\bar{R}))}{Volume(H_S(R) \cup H_S(\bar{R}))}$$

□

Theorem 1 states that we can convert our problem into a more familiar form without sacrificing accuracy. Our problem of calculating the ratio of the size of the intersection of the two polyhedral cones to that of their union is now transformed into finding this ratio for two convex polytopes.

STEP III. Even after transforming the two flux cones that represent the impacts into convex polytopes, computing the hyper-volumes of their intersection and union is still a difficult problem. We approximate the intersection and the union of these two polytopes by the intersection and the union of their minimum enclosing balls (MEBs). Figure 4 illustrates an example MEB that bounds a polytope in three dimensional space. In our case, we compute the MEBs of both polytopes in high dimensional space by using the method of Fischer *et al.* [6]. This method computes the MEBs of point sets instead of polytopes. Therefore, we treat each corner of our polytopes as a point and then compute the MEB of these points. This MEB is the same as the MEB of the actual polytope as this polytope is convex by our construction. Computing MEBs for large number of points can be done efficiently in spaces with up to a few hundred dimensions [6]. The MEBs we compute provide a simple parametric representation for the impacts of reaction sets in terms of a center and a radius in high dimensional flux space.

STEP IV. By using parametric representations of the MEBs, we find the intersection of the impacts of reaction sets as the points of the polytopes that lie in both MEBs. The union is trivial and is the set of all points of first polytope plus the ones of the second. We compute the MEBs of these points sets to get the radii of MEBs of the intersection and the union of the impacts. Let r_{int} and r_{un} denote these radii respectively. Then, we calculate the ratio of the volume of the MEB of the intersection of impacts to the volume of the MEB of their union as $\frac{c_d.r_{int}^d}{c_d.r_{un}^d}$ where d is the number of dimensions and c_d is a constant factor depending on d. Since the constants cancel out, we have the ratio $\frac{r_{int}^d}{r_{un}^d}$ as our approximation to $\frac{\text{Impact}(R, \mathcal{P}|\bar{\mathcal{P}}) \cap \text{Impact}(\bar{R}, \bar{\mathcal{P}}|\mathcal{P})}{\text{Impact}(R, \mathcal{P}|\bar{\mathcal{P}}) \cup \text{Impact}(\bar{R}, \bar{\mathcal{P}}|\mathcal{P})}$. Finally, we compute the functional similarity by subtracting this ratio from 1 and taking the minus logarithm of the result.

5. EXPERIMENTAL RESULTS

In this section, we experimentally evaluate the performance of our method. First, we measure the accuracy of our method in identifying reaction sets of different pathways that are functionally equivalent. We compare our functional similarity score with several existing similarity scores in this setting. Then, we test whether our method can be utilized to predict essential reactions of a given metabolic pathway. We modify our functional similarity score to define an essentiality score and we calculate its statistical significance.

Dataset: We use the metabolic pathways taken from KEGG pathway database. We use 12 different pathways of *E.coli K-12 MG1655*. The average number of reactions per pathway is 20.75. The largest pathway in our database (Pyrimidine metabolism) has 69 reactions, 60 compounds and 1,007 EFMs.

Environment: We run all the experiments on a desktop computer running Ubuntu 9.10 with one processor and 2 GB of RAM.

5.1 Identification of Functional Similarities

Our motivation for identifying functionally similar reaction sets is that metabolisms of different organisms can perform the same func-

tion through different sets of reactions. Here, we evaluate the accuracy of the similarity measure we describe (Section 4) for identifying reaction sets with similar biological functions.

In this experimental setting, we compare three existing similarity measures with the functional similarity we propose. The first similarity measure, *Degree similarity*, takes into account only the topological features of the pathways [14]. This measure considers two reaction sets from two possibly different pathways as similar if the in and out degrees of the reactions in these sets are similar. This suggests that the reaction sets with central reactions in one pathway tend to be similar to the reactions sets that contain reactions with high centrality in the other pathway. The second existing measure is the *UP/UC similarity* [17]. UP/UC suggests that the number of compounds that are uniquely produced or consumed by a reaction is a good indicator of the biological role of that reaction. If this number is similar for the reactions of two reaction sets we compare, then they are considered to have similar functions by the UP/UC measure. The last existing approach is *Reaction participation similarity* [5, 16]. Reaction participation uses the number of EFMs that a reaction participates in to measure the importance of that reaction for the pathway. Unlike the first two approaches, reaction participation similarity uses the steady state information of the metabolic pathways. However, it considers each EFM equally important and it only counts the number of EFMs that a reaction participates in. As a result, it can over or underestimate the contributions of different EFMs in the metabolic capabilities of a pathway.

In order to have a fair comparison between existing measures and the one we propose, we introduce functionally equivalent reaction sets by perturbing a given pathway. We randomly combine two neighbor reactions of the original pathway with a given probability. We do this for all reaction pairs to attain a perturbed instance of the pathway. This way we isolate the effects of other factors (e.g., pathway size, number of EFMs, etc.) and can create reaction sets in the perturbed pathway that precisely have the same function as another reaction set in the original pathway. A reaction in the perturbed instance is a combination of two reactions of the original pathway and is functionally equivalent to the set of these two reactions. We take each reaction from a perturbed pathway that represents a combination, then compare it with each connected reaction set of size at most two of the original pathway. We, then, calculate the recall of each method as the percentage of the functionally equivalent reaction sets it identifies in top-k scoring reaction sets.

We compare the average recall values of different similarity measures when we use 100 perturbed instances of the energy metabolism of *E.coli*. This metabolism is a combination of four different pathways in the KEGG database, namely reductive carboxylate cycle, methane metabolism, nitrogen metabolism and sulfur metabolism. Overall energy metabolism of *E.coli* has 47 reactions, 87 compounds and 37 EFMs. The number of connected reaction sets of size at most two for this pathway is 130. Figure 5 plots the average recall values for four different similarity measures. The one with highest recall value for all different k is the functional similarity we propose in this paper. To see it in terms of numbers, when we consider top 20% of all the reaction sets (i.e., top 26 among 130) our similarity measure achieves $\approx$ 76% recall whereas the recall of the second best method (i.e., reaction participation similarity) stays at 57%. Also, our functional similarity reaches 95% recall at top 37% of all the reaction sets while the other methods can only achieve this recall at 64%, 77% and 90%. *The results of this experiment suggest that the functional similarity is a more accurate measure compared to existing ones in order to identify reactions sets that play similar biological roles in different pathways.*

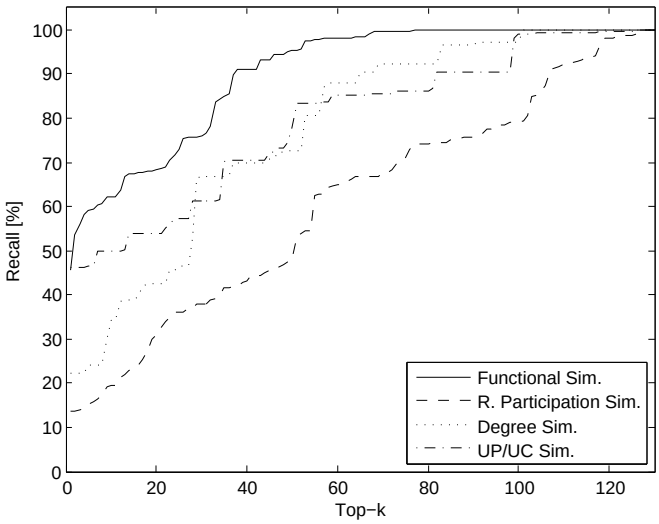

Figure 5: *Comparison of different similarity measures for identifying functionally similar reaction sets.* We combine two connected reactions with probability 0.05 to introduce functionally equivalent but perturbed instances of *the energy metabolism* of *E.coli*. We measure *recall* as the percentage of identifying functionally equivalent parts of the original and 100 perturbed pathways within the top-k most similar reaction sets.

5.2 Prediction of Essential Reaction Sets

An important and well-studied problem in biology is to find the essential components of a metabolism. Often the essentiality of a gene of an organism is determined by looking at its knockout phenotype. The essentiality can take binary values (i.e., essential or non-essential) as well as categorical or continuous values. This notion of essentiality can also be extended to other components of the metabolism such as reactions, enzymes and compounds.

Here, we utilize our characterization of the impact of a reaction set to determine its essentiality for a metabolic pathway. Our intuition is that among two reaction sets of a pathway, the one with the bigger impact is more essential compared to the other. In order to test our hypothesis, we use different pathways of *E.coli* that contain essential reactions for the organism. We construct the set of all essential reactions using the essential genes listed in the online database PEC (Profiling of *E.coli* Chromosome) [10]. We extract 341 essential reactions in *E.coli* by using 302 essential genes listed in PEC. Table 1 lists eight metabolic pathways that contain different fractions of essential reactions. For each one of these pathways, we first enumerate all its connected reaction sets of size at most three. The term *connected* indicates that each reaction in the set is a neighbor of at least one other reaction in that set. Then, for each connected reaction set R, we compute its essentiality in $\mathcal{P}$ as the fraction of the steady states that are unreachable after inhibiting the reactions in R. Formally, we compute the essentiality by using our functional similarity as follows:

Essentiality(R, $\mathcal{P}$) = Sim(R, $\mathcal{P} \mid \mathcal{P}$, $\mathcal{P}$) = Sim(R, $\mathcal{P} \mid \mathcal{P}$) We calculate the essentiality as the functional similarity of R and the set of all reactions of $\mathcal{P}$ given the pathway $\mathcal{P}$.

Table 1: ***Essential reactions in eight different metabolic pathways of E.coli.*** **[1] KEGG identifier of the pathway. [2]Number of essential reactions of the pathway according to PEC classification [10]. [3]Number of reactions of the pathway. [4] Probability of obtaining (randomly) at least as many essential reactions as our method predicts in top 10% of all connected reaction sets with size at most three.**

P. Id[1]	*Pathway Name*	*#E.*[2]	*#R.*[3]	*p-value*[4]
00240	Pyrimidine met.	10	69	7.99E-15
00860	Porphyrin and chl. met.	9	21	3.02E-06
00564	Glycerophospholipid met.	7	20	1.24E-05
00010	Glycolysis/Gluconeogenesis	4	26	6.71E-03
00670	One carbon pool by folate	6	16	1.23E-02
00540	Lipopolysaccharide bio.	15	19	4.86E-02
00760	Nicotinate metabolism	4	16	2.41E-01
00300	Lysine biosynthesis	8	15	2.45E-01

After computing essentiality for each reaction set in this manner, we sort them in decreasing order according to their essentiality scores. We calculate the statistical significance of our ranking by calculating its p-value. More specifically, let p be the probability that a randomly selected reaction is essential and t, n denote the number of essential reactions and all reactions in top 10% of connected reaction sets according to our ranking respectively. The p-value of our prediction for this pathway is

$$p\text{-}value(p,t,n) = \sum_{i=t}^{n} \binom{n}{i} p^i (1-p)^{n-i}$$

We report p-values for the eight pathways of *E.coli* listed in Table 1. Consider the Pyrimidine metabolism which is the first row in this table. The probability of a reaction being essential for this pathway is $p = \frac{10}{69} = 0.145$. Using this probability, the expected value of the number of appearances of essential reactions in top 10% of connected reaction sets for this pathway is 14.21 whereas according to our ranking we observe 111 appearances of essential reactions in this top 10%. This translates to a p-value in the order of E-15 which suggests strong statistical significance. All the p-values of the first six rows of Table 1 are also statistically significant (i.e., p-value< 0.05). For the last two pathways the p-value is greater than 0.05, however our method still reports more than the expected number of essential reactions.

Next, we take a closer look at how the statistical significance of our essentiality score changes when we consider different percentages of the highest scoring reaction sets. For this purpose, we use the Pyrimidine metabolism of *E.coli* and compute Z-scores for different percentages of k. Figure 6 plots the Z-scores for our method. We observe that Z-score reaches significant values even when we consider a small percentage of the top-k results. We achieve Z-score of 10 for top-10% which implies that our result is 10 standard deviations away from the random result and hence is statistically significant. Z-score reaches its peak at top-26.4%, which is shown by the dashed line in Figure 6. At this point Z-score is 18.24 and of all the reaction sets up to that point, approximately 88% $(228/259)$ contain at least one essential reaction. Considering that only 10 out of 69 reactions are essential, this 88% shows that our scoring scheme can successfully extract reaction sets with essential reactions by assigning them larger scores compared to the other sets. *The results of this section indicates that the functional similarity measure we propose can be extended to define an essentiality score that can accurately identify essential reactions and reaction sets of metabolic pathways.*

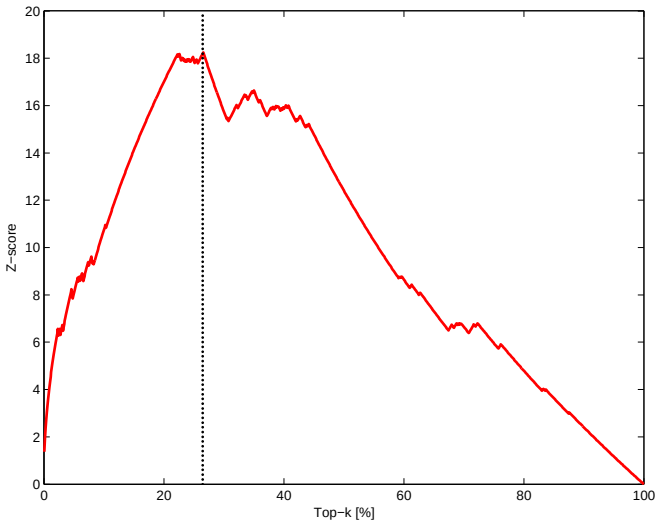

Figure 6: ***Statistical significance of our essentiality prediction for Pyrimidine metabolism of*** **E.coli.**

6. CONCLUSION

Understanding the functional role of a component set of a metabolic pathway has been an important problem in molecular biology. Often homological and topological features of the pathways have been used for this purpose. However, the biological functions of two reaction sets can be equivalent even when their size, connectivity, intermediate products and catalyzing enzymes are different. Therefore, neither the topological features (e.g., centrality) nor the homological similarities (e.g, enzymes similarity, compound similarity) can provide sufficient information for identifying such functional similarities.

In this paper, we developed a systematic way to characterize and compute the functional similarity between two given reaction subsets in metabolic pathways. We built a mathematical model that explains the impact of a reaction set in terms of the set of possible steady states of its pathway. Specifically, our model computes the impact as the portion of the flux cone of the original pathway that cannot be achieved without the reactions in the set that we consider. Using this model, we characterized the functional similarity of two reaction sets from potentially different pathways. We achieved this by first lifting the flux cones of the pathways to a higher dimensional flux space that is the union of the flux spaces of these pathways. We, then, represented the functional similarity of two reaction subsets as the ratio of the volume of the intersection of the regions corresponding to their impacts on their pathways to the volume of the union of these regions. We developed a novel method that computes this ratio as follows. It first computes the Elementary Flux Modes (EFMs) of the pathway with and without the given reaction set. It, then, transforms the polyhedral cones into polytopes by taking their intersections with a hyperplane in the positive quadrant of the new flux space. Lastly, it calculates the intersection of these polytopes as the intersection of their minimum enclosing balls (MEB). Our experiments on real metabolic pathways demonstrated that our method can identify biological similarities accurately. Moreover, we observed that our definition of impact

can provide biologically and statistically significant predictions of essential reaction sets.

Characterizing the function of a set of components (e.g., reactions) mathematically has great value in numerous applications of computational biology. Thus, we believe that the ideas developed in this paper has the great potential to lay foundations to many advances in understanding and comparing complex biological networks better.

7. ACKNOWLEDGMENTS

This work was supported partially by NSF under grants CCF-0829867 and IIS-0845439.

8. REFERENCES

[1] F. Ay and T. Kahveci. SubMAP: Aligning metabolic pathways with subnetwork mappings. In *International Conference on Research in Computational Molecular Biology (RECOMB)*, To appear 2010.

[2] F. Ay, T. Kahveci, and V. Crecy-Lagard. A fast and accurate algorithm for comparative analysis of metabolic pathways. *Journal of Bioinformatics and Computational Biology (JBCB)*, 7(3):389–428, 2009.

[3] A. Cakmak and G. Ozsoyoglu. Mining biological networks for unknown pathways. *Bioinformatics*, 23(20):2775–83, 2007.

[4] M. Campillos, M. Kuhn, A. C. Gavin, L. J. Jensen, and P. Bork. Drug Target Identification Using Side-Effect Similarity. *Science*, 321(5886):263–6, 2008.

[5] C. Conradi, D. Flockerzi, J. Raisch, and J. Stelling. Subnetwork analysis reveals dynamic features of complex (bio)chemical networks. *Proceedings of the National Academy of Sciences (PNAS)*, 104(49):19175–80, 2007.

[6] K. Fischer, B. Gartner, and M. Kutz. Fast Smallest-Enclosing-Ball Computation in High Dimensions. In *11th Annual European Symposium on Algorithms (ESA)*, pages 630–41, 2003.

[7] C. Francke, R. Siezen, and B. Teusink. Reconstructing the metabolic network of a bacterium from its genome. *Trends in Microbiology*, 13(11):550–8, 2005.

[8] M. Heymans and A. Singh. Deriving phylogenetic trees from the similarity analysis of metabolic pathways. *Bioinformatics*, 19:138–46, 2003.

[9] R. Jothi, T. M. Przytycka, and L. Aravind. Discovering functional linkages and uncharacterized cellular pathways using phylogenetic profile comparisons: A comprehensive assessment. *BMC Bioinformatics*, 8:173, 2007.

[10] J. Kato and M. Hashimoto. Construction of consecutive deletions of the Escherichia coli chromosome. *Molecular Systems Biology*, 3:132, 2007.

[11] S. Klamt and J. Stelling. Two approaches for metabolic pathway analysis? *Trends in Biotechnology*, 21(2):64–9, 2003.

[12] A. Larhlimi and A. Bockmayr. A new constraint-based description of the steady-state flux cone of metabolic networks. *Discrete Applied Mathematics*, 157:2257–66, 2009.

[13] L. Lynd, C. Wyman, and T. Gerngross. Biocommodity Engineering. *Biotechnology Progress*, 15(5):777–93, 1999.

[14] H. W. Ma and A. P. Zeng. The connectivity structure, giant strong component and centrality of metabolic networks. *Bioinformatics*, 19(11):1423–30, 2003.

[15] A. A. Margolin, K. Wang, W. K. Lim, M. Kustagi, I. Nemenman, and A. Califano. Reverse engineering cellular networks. *Nature Protocols*, 1(2):662–71, 2006.

[16] J. A. Papin, N. D. Price, and B. O. Palsson. Extreme pathway lengths and reaction participation in genome-scale metabolic networks. *Genome Research*, 12(12):1889–900, 2002.

[17] A. Samal, S. Singh, V. Giri, S. Krishna, N. Raghuram, and S. Jain. Low degree metabolites explain essential reactions and enhance modularity in biological networks. *BMC Bioinformatics*, 7:118, 2006.

[18] S. Schuster and C. Hilgetag. On elementary flux modes in biochemical reaction systems at steady state. *Journal of Biological Systems*, 2:165–82, 1994.

[19] S. Schuster, C. Hilgetag, J. Woods, and D. Fell. Reaction routes in biochemical reaction systems: algebraic properties, validated calculation procedure and example from nucleotide metabolism. *Journal of Mathematical Biology*, 45(2):153–81, 2002.

[20] S. Schuster, T.Dandekar, and D. A. Fell. Detection of elementary flux modes in biochemical networks: a promising tool for pathway analysis and metabolic engineering. *Trends in Biotechnology*, 17(2):53–60, 1999.

[21] D. Segre, D. Vitkup, and G. M. Church. Analysis of optimality in natural and perturbed metabolic networks. *Proceedings of the National Academy of Sciences (PNAS)*, 99(23):15112–7, 2002.

[22] P. Sridhar, T. Kahveci, and S. Ranka. An iterative algorithm for metabolic network-based drug target identification. In *Pacific Symposium on Biocomputing (PSB)*, volume 12, pages 88–99, 2007.

[23] J. Stelling, S. Klamt, K. Bettenbrock, S. Schuster, and E. D. Gilles. Metabolic network structure determines key aspects of functionality and regulation. *Nature*, 420:190–193, 2002.

[24] G. Stephanopoulos. Metabolic engineering. *Current Opinions in Biotechnology*, 5(2):196–200, 1994.

[25] B. L. Steven and B. O. Palsson. Expa: a program for calculating extreme pathways in biochemical reaction networks. *Bioinformatics*, 21(8):1739–40, 2005.

[26] A. Tatsuya, M. Satoru, and K. Satoru. Identification of genetic networks from a small number of gene expression patterns under the Boolean network model. In *Pacific Symposium on Biocomputing (PSB)*, volume 4, pages 17–28, 1999.

[27] M. Terzer and J. Stelling. Large scale computation of elementary flux modes with bit pattern trees. *Bioinformatics*, 24(19):2229–35, 2008.

[28] C. Trinh, A. Wlaschin, and F. Srienc. Elementary mode analysis: a useful metabolic pathway analysis tool for characterizing cellular metabolism. *Applied Microbiology and Biotechnology*, 81(5):813–26, 2009.

[29] S. Wong, L. Zhang, A. Tong, Z. Li, D. Goldberg, O. King, G. Lesage, M. Vidal, B. Andrews, H. Bussey, C. Boone, and F. Roth. Combining biological networks to predict genetic interactions. *Proceedings of the National Academy of Sciences (PNAS)*, 101(44):15682–7, 2004.

[30] X. Wu, L. Zhu, J. Guo, D.-Y. Zhang, and K. Lin. Prediction of yeast protein-protein interaction network: insights from the Gene Ontology and annotations. *Nucleic Acids Research*, 34(7):2137–50, 2006.