Max-Planck-Institut
für Mathematik
in den Naturwissenschaften
Leipzig

Flux-based classification of reactions reveals a functional bow-tie organization of complex metabolic networks

by

Shalini Singh, Areejit Samal, Varun Giri, Sandeep Krishna,
Nandula Raghuram, and Sanjay Jain

Preprint no.: 43 2013

Flux-based classification of reactions reveals a functional bow-tie organization of complex metabolic networks

Shalini Singh[1,2], Areejit Samal[1,3,4], Varun Giri[1], Sandeep Krishna[5], Nandula Raghuram[6], and Sanjay Jain[1,7,8*]
[1]*Department of Physics and Astrophysics, University of Delhi, Delhi 110007, India*
[2] *Department of Genetics, University of Delhi, South Campus, New Delhi, India*
[3]*Max Planck Institute for Mathematics in the Sciences, Inselstrasse 22, D-04103 Leipzig, Germany*
[4]*Laboratoire de Physique Théorique et Modèles Statistiques,*
CNRS and Univ Paris-Sud, UMR 8626, F-91405 Orsay, France
[5]*National Centre for Biological Sciences, UAS-GKVK Campus, Bangalore 560065, India*
[6]*School of Biotechnology, GGS Indraprastha University, Dwarka, New Delhi 110078, India*
[7] *Jawaharlal Nehru Centre for Advanced Scientific Research, Bangalore 560064, India and*
[8]*Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501, USA*

Unraveling the structure of complex biological networks and relating it to their functional role is an important task in systems biology. Here we attempt to characterize the functional organization of the large-scale metabolic networks of three microorganisms. We apply flux balance analysis to study the optimal growth states of these organisms in different environments. By investigating the differential usage of reactions across flux patterns for different environments, we observe a striking bimodal distribution in the activity of reactions. Motivated by this, we propose a simple algorithm to decompose the metabolic network into three sub-networks. It turns out that our reaction classifier which is blind to the biochemical role of pathways leads to three functionally relevant sub-networks that correspond to input, output and intermediate parts of the metabolic network with distinct structural characteristics. Our decomposition method unveils a functional bow-tie organization of metabolic networks that is different from the bow-tie structure determined by graph-theoretic methods that do not incorporate functionality.

PACS numbers: 82.39.Rt 87.18.Vf 87.18.-h

I. INTRODUCTION

Biological systems provide many examples of the intricate relationship between the structure and functionality of complex networks [1–7]. Cellular metabolism is a complex biochemical network of several hundred metabolites that are processed and interconverted by enzyme-catalyzed reactions [8–13]. Metabolic networks have a dynamic flexibility that enables organisms to survive under diverse environmental conditions. A key goal of systems biology is to unveil the functional organization of metabolic networks explaining their system-level response to different environments. To this end, we have attempted to decompose metabolic networks into functionally relevant sub-networks. Flux balance analysis (FBA) has been widely used to harness the knowledge of large-scale metabolic networks and investigate genotype-phenotype relationships [14–16]. FBA has been successful in predicting the growth and deletion phenotypes of organisms [17–19]. Reaction fluxes carry information about the flows on metabolic networks and, as such, describe the functional use of the network by the organism. In this paper, we have used this information to decompose the network into functionally relevant sub-networks.

The paper is organized as follows: In section II we describe the modelling framework in which we study metabolic networks. In section III we discuss the classification of active reactions in metabolic networks into three categories by an algorithm that is blind to their biochemical roles. Section IV shows that the three categories are functionally relevant for the organism. In section V we compare the bow-tie architecture obtained by our functional classification of reactions with that obtained by graph-theoretic methods that do not employ functional information. In the last section we conclude with a summary.

II. THE MODELLING FRAMEWORK

A. Flux balance analysis (FBA)

Flux balance analysis (FBA) is a computational approach widely used to analyze the capabilities of genome-scale metabolic networks [14–16]. The stoichiometric matrix **S** encapsulates the stoichiometric coefficients of different metabolites involved in various reactions of the metabolic network. The stoichiometric matrix $\mathbf{S} = (S_{pj})$ has dimensions $P \times N$, where P denotes the number of metabolites and N denotes the number of reactions in the metabolic network. S_{pj} is the number of molecules of the metabolite p produced in reaction j (if metabolite p is consumed in reaction j, S_{pj} is negative). The stoichiometric matrix for a hypothetical reaction network is shown in Fig. 1. FBA primarily uses structural information of the metabolic network contained in the matrix **S** to predict the possible steady state flux distribution of all reactions and the maximum growth rate of an organism. In any metabolic steady state, the metabolites achieve a dynamic mass balance wherein the vector

[*] jain@physics.du.ac.in

TABLE I. Comparison of the three metabolic networks: *E. coli*, *S. cerevisiae* and *S. aureus*.

Property	*E. coli*	*S. cerevisiae*	*S. aureus*
Number of metabolites	761	1061	648
Number of reactions in the model	931	1149	641
Number of one-sided reactions in the equivalent network	1167	1576	863
Number of external metabolites	143	116	84
Number of organic external metabolites (carbon sources)	131	107	68
Number of biomass metabolites	49	42	56
Number of feasible minimal environments	89	43	27
Number of active reactions	585	482	418
Number of reactions in category I	185	89	84
Number of reactions in category IIa	147	117	194
Number of reactions in category IIb	42	46	28
Number of reactions in category III	211	230	112

$\mathbf{v}$ of fluxes through the reactions satisfies the following equation representing the stoichiometric and mass balance constraints:

$$\mathbf{S}.\mathbf{v} = 0. \tag{1}$$

Equation 1 is an under-determined linear system of equations relating various reaction fluxes in genome-scale metabolic networks leading to a large solution space of allowable fluxes. The space of allowable solutions can be reduced by incorporating thermodynamic and enzyme capacity constraints. To obtain a particular solution, linear programming is used to find a set of flux values - a particular flux vector $\mathbf{v}$ - that maximizes a biologically relevant linear objective function Z. The linear programming formulation of the FBA problem can be written as:

$$\max Z \;=\; \max \, \{\mathbf{c}^{\mathbf{T}}\mathbf{v} | \mathbf{S}.\mathbf{v} = 0, \mathbf{a} \leq \mathbf{v} \leq \mathbf{b}\}, \tag{2}$$

where vectors $\mathbf{a}$ and $\mathbf{b}$ contain the lower and upper bounds of different fluxes in $\mathbf{v}$ and the vector $\mathbf{c}$ corresponds to the coefficients of the objective function Z. In FBA, the objective function Z is usually taken to be the growth rate of the organism. The environment, or medium, is defined in this approach by the components of $\mathbf{a}$ and $\mathbf{b}$ corresponding to the transport reactions, which determine, in particular, the set of metabolites whose uptake is allowed.

B. Large-scale metabolic networks

In this work, we have analyzed the large-scale metabolic networks of three microorganisms: *Escherichia coli* (version iJR904 [20]), *Saccharomyces cerevisiae* (version iND750 [21]) and *Staphylococcus aureus* (version iSB619 [22]). Table I gives the number of metabolites and reactions in the metabolic networks of these three organisms. The metabolic networks contain internal and transport reactions. Internal reactions occur within the cell boundary. Transport reactions represent processes involving import or export of metabolites across the cell boundary. Each model also contains a pseudo biomass reaction that simulates the drain of various biomass precursor metabolites for growth in the specific organism. Starting from the published metabolic network, we obtain an equivalent reaction network as follows: Every reversible reaction in the network is converted into two one-sided (irreversible) reactions so that all reaction fluxes in the equivalent system are non-negative. A few reactions appear in duplicate in these networks, and only a single copy of each reaction is kept in the equivalent network. The equivalent metabolic network is a reaction set consisting of N unique one-sided reactions where N is 1167, 1576 and 863 for *E. coli*, *S. cerevisiae* and *S. aureus*, respectively (cf. Table I).

C. Feasible minimal environments and associated flux vectors

In this work, we have considered ‘minimal’ aerobic environments – minimal in the sense that each environment contains a single organic external metabolite that is the sole source of carbon, and single inorganic sources for each of the elements nitrogen, phosphorus, sulphur, oxygen, sodium, potassium and iron, apart from hydrogen ions and water. Aerobic means that molecular oxygen is available in the external medium. Furthermore the minimal environments differ from each other solely in their organic carbon source; the set of inorganic sources is the same for all the minimal environments considered here for any given organism. Thus the number of environments we consider for each organism coincides with the number of organic external metabolites (carbon sources) in its metabolic network (cf. Table I). We further assume that each environment contains a limited amount of the organic carbon source and unlimited amounts of the inorganic metabolites, namely, ammonia (source of nitrogen), pyrophosphate (source of phosphorus), sulphate (source of sulphur), molecular oxygen, ions of sodium, potassium, iron and hydrogen, and water molecules. From this set of minimal environments, we used FBA to determine

Reaction Network

R1: $2A + B \rightarrow C + 3D$

R2: $A + 3B \rightarrow C + E$

R3: $A \rightarrow 2B + D + E$

R4: $4B \rightarrow D + A$

R5: $D + 2B \rightarrow C + 2E$

R6: $C + 4E \rightarrow 3B + D$

Stoichiometric Matrix

$$\begin{array}{c|cccccc|}
 & R1 & R2 & R3 & R4 & R5 & R6 \\
\hline
A & -2 & -1 & -1 & 1 & 0 & 0 \\
B & -1 & -3 & 2 & -4 & -2 & 3 \\
C & 1 & 1 & 0 & 0 & 1 & -1 \\
D & 3 & 0 & 1 & 1 & -1 & 1 \\
E & 0 & 1 & 1 & 0 & 2 & -4
\end{array}$$

FIG. 1. **Example of stoichiometric matrix for a hypothetical reaction network.** The hypothetical reaction network has 6 reactions involving 5 metabolites. The rows of the stoichiometric matrix correspond to various metabolites and the columns correspond to various reactions in the metabolic network.

the subset of minimal environments supporting growth in the metabolic networks of *E. coli*, *S. cerevisiae* and *S. aureus*. A minimal environment was termed as *feasible* if the growth rate predicted by FBA was found to be nonzero for that environment. The number M of feasible minimal environments in *E. coli*, *S. cerevisiae* and *S. aureus* was obtained to be 89, 43 and 27, respectively (cf. Table I) [23]. For each organism, and for each feasible minimal environment for that organism, we obtained an N-dimensional optimal flux vector $\mathbf{v}$ using FBA whose component v_j gives the flux of reaction j. For every organism this led to a set of M flux vectors corresponding to the M feasible minimal environments, which were stored in the form of a matrix $\mathbf{V}=(v_j^\alpha)$ of dimensions $N\times M$ where the rows (j=1,2,. . .,N) correspond to different reactions in network and columns (α=1,2,. . .,M) to different feasible minimal environments. v_j^α is defined as the flux of reaction j in the optimal flux vector $\mathbf{v}$ obtained for environment α.

D. Active reactions

A given reaction j is termed as *active* in an environment α if $v_j^\alpha>0$. The *activity* m of a reaction denotes the number of minimal environments in which the reaction is active. The activity m for a reaction ranges from 0 to M with M equal to 89, 43 and 27 for *E. coli*, *S. cerevisiae* and *S. aureus*, respectively. A reaction j is termed as active in an organism if $m\geq1$ (i.e., if it is active in at least one feasible minimal environment for that organism). The number of active reactions in *E. coli*, *S. cerevisiae* and *S. aureus* was obtained to be 585, 482 and 418, respectively (cf. Table I). This paper primarily focuses on decomposing this set of active reactions into functionally relevant sub-networks.

III. CLASSIFICATION OF ACTIVE REACTIONS

We ask the question: How does the activity of a reaction vary across different environments? To address this question, we determine the frequency distribution of the activity of reactions in an organism. Fig. 2 shows the histogram of the activity of reactions in the *E. coli* metabolic network. The distribution is bimodal. Most reactions in *E. coli* are either once-active (m=1) or always active (m=89); the number of reactions for any given intermediate activity m in the range $1<m<89$ is small. Thus, the largest number of active reactions in the metabolic network are used in either one environment or in all environments. The histograms of activity of reactions in *S. cerevisiae* and *S. aureus* also have a pattern similar to that in *E. coli* (cf. Fig. 2). The frequency distribution of activity of reactions in the three organisms suggests a natural classification of active reactions into three categories:

(a) Category I reactions or once-active reactions (m=1)

(b) Category II reactions or always active reactions (m=M)

(c) Category III reactions with intermediate activity ($1<m<M$)

A. Sub-classification based on correlation of reaction fluxes

Clustering of gene expression data using the correlation coefficient has been successful in predicting regulatory modules associated with a biological function across diverse conditions [24]. We used the correlation coefficient to identify sets of reactions whose fluxes are correlated across different environments. We used the set of M flux vectors corresponding to M feasible minimal

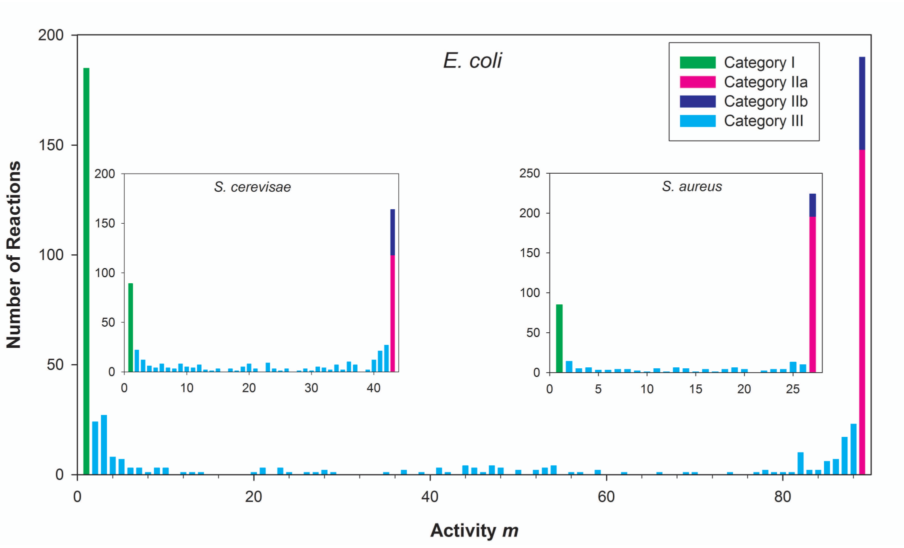

FIG. 2. (Color online) **The histogram of activity of reactions in the *E. coli* metabolic network.** The bars show the number of reactions that have an activity m where m ranges from 1 to 89 feasible minimal environments in the *E. coli* metabolic network. The green bar represents 185 category I reactions which are once-active. The pink bar represents 147 category IIa reactions (a subset of 189 always active category II reactions) that have fluxes perfectly correlated across environments. The deep blue bar represents 42 category IIb reactions that account for the remaining category II reactions. The light blue bars account for 211 category III reactions with intermediate activity. **Insets:** Histograms of activity of reactions in *S. cerevisiae* and *S. aureus*. The three categories of reactions in *S. cerevisiae* and *S. aureus* were defined in a manner similar to *E. coli*.

environments contained in the matrix $\mathbf{V} = (v_j^\alpha)$ to obtain the matrix $\mathbf{C} = (C_{jk})$ where C_{jk} is the correlation coefficient between two active reactions j and k and is given by:

$$C_{jk} = \frac{1}{M}\sum_{\alpha=1}^{M}\frac{v_j^\alpha v_k^\alpha}{\phi_j \phi_k}, \tag{3}$$

$$\text{where } \phi_j = \sqrt{\frac{1}{M}\sum_{\alpha=1}^{M} {v_j^\alpha}^2}.$$

If $C_{jk} = 1$ then reactions j and k are perfectly correlated with each other in the given set of environments. Perfect clusters in metabolic networks are maximal sets of reactions that are perfectly correlated to each other pairwise. Perfect clusters are similar to enzyme subsets [25, 26], correlated reaction sets [27, 28] or fully coupled sets [29] which have been used to detect modules in metabolic networks.

We use Eq. 3 to identify perfect clusters in metabolic networks of *E. coli*, *S. cerevisiae* and *S. aureus*. In particular, a large perfect cluster of 147 reactions was found in *E. coli* that is a subset of category II reactions. We refer to this subset of perfectly correlated reactions within category II as category IIa reactions. The remaining 42 category II reactions that are always active but not perfectly clustered with category IIa reactions are part of category IIb. Similarly, large perfect clusters of sizes 117 and 194 were found in category II reactions of *S. cerevisiae* and *S. aureus*, respectively. In Fig. 2, category IIa and IIb reactions are shown in pink and blue colours, respectively. We have shown elsewhere that perfect clusters are metabolic modules that can be explained by studying the connectivity of their constituent metabolites [23].

As mentioned earlier we obtained the flux vectors by maximizing the objective function Z that corresponds to the growth rate of the cell. In FBA cell growth stands for the production of all the 'biomass metabolites' in specified ratios that correspond to the composition of the average cell under consideration. The role of growth maximization is to obtain an explicit flux vector for each medium. While the magnitudes of the components of $\mathbf{v}$ obtained by maximization of the growth rate depend upon the precise ratios, the activity of a reaction, as defined above, depends not on the actual magnitude of the corresponding component of $\mathbf{v}$, but only on whether the magnitude is zero or nonzero. The latter does not depend upon the precise ratios of the biomass metabolites in the objective function, but only on the set of metabolites that are present in the objective function. Thus our classification results are quite robust to the perturbation of the ratios in the objective function, as long as the set of biomass metabolites is held fixed (details not shown).

Note that we have used a single optimal flux vector $\mathbf{v}$ obtained using FBA for each of the M feasible minimal environments to determine the activity of a reaction and the set of active reactions in the metabolic network of an organism. However, it is well known that there exist multiple flux vectors or alternate optimal solutions in most large-scale metabolic networks that maximize growth in a given environment [28, 30–32]. In principle, due to the presence of alternate optima, the set of active reactions can change depending on the choice of the flux vectors. In Appendix A, we show the robustness of our reaction

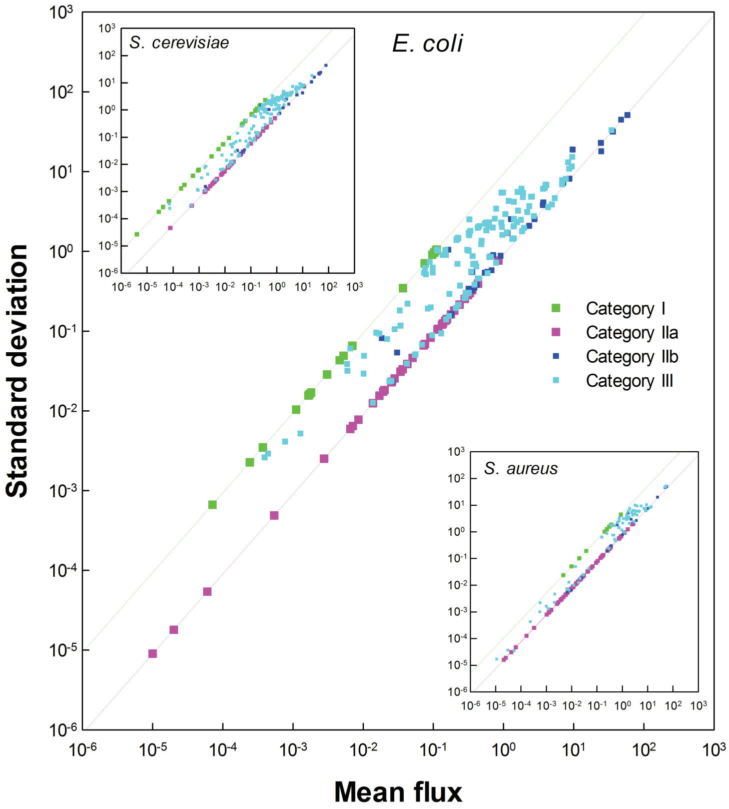

FIG. 3. (Color online) **Standard deviation versus mean flux of active reactions in the *E. coli* metabolic network.** The plot shows standard deviation σ versus mean flux $\langle v\rangle$ of the 585 active reactions in *E. coli* metabolic network across $M = 89$ feasible minimal environments on a logarithmic scale. The green, pink, dark blue and cyan dots represent category I, IIa, IIb and III reactions, respectively. The three categories of reactions show up quite distinctly (upper line, category I; lower line, category IIa; with category IIb and category III in between the two lines). The upper line is the expected curve $\sigma = (M-1)^{1/2}\langle v\rangle$ for category I reactions. The lower line is the expected curve $\sigma = b\langle v\rangle$ for perfectly correlated category IIa reactions with $b = 0.98 \pm 0.1$ obtained via best fit to the data. **Insets:** Scatter plots of σ versus $\langle v\rangle$ of active reactions in *S. cerevisiae* and *S. aureus* metabolic networks.

categories to the presence of alternate optima.

B. Scatter plot of standard deviation versus mean flux of reactions across environments discriminates between the three categories

For each active reaction, following Almaas *et al* [33], we have calculated the mean flux $\langle v\rangle$ and the standard deviation σ around this mean by averaging the flux of the reaction over M feasible minimal environments. Fig. 3 shows the scatter plot of σ versus $\langle v\rangle$ for active reactions in *E. coli*. It is evident that the distribution of points is different for the various categories we have defined. All category I points lie on the upper line, all category IIa points lie on the lower line, while category IIb and category III points lie largely in between the two lines. The upper line in Fig. 3 is the expected curve $\sigma = (M-1)^{1/2}\langle v\rangle$ for category I reactions and the lower line is the curve $\sigma = b\langle v\rangle$, where b is obtained via best fit of data for category IIa reactions. Appendix B gives the derivation of the relation between σ and $\langle v\rangle$ for category I and IIa reactions. Our classification of reactions into the three categories did not use the actual values of the fluxes of the reactions, but only the information about whether the flux was zero or nonzero in a particular medium. Fig. 3 uses information about the actual flux values. It shows that the different categories of reactions are distinct from each other by virtue of the statistical properties of their magnitudes as well.

IV. FUNCTIONAL RELEVANCE OF THE THREE CATEGORIES OF REACTIONS

Until now our classification of active reactions into the three categories was solely motivated by the activity of reactions in *E. coli*, *S. cerevisiae* and *S. aureus* with two very prominent peaks for once-active and always active reactions (cf. Fig. 2). However, we now show that our three categories I, II, and III obtained using a computational algorithm blind to the biochemical nature of pathways correspond to the input, output and intermediate sub-networks, respectively. Thus, each category of reac-

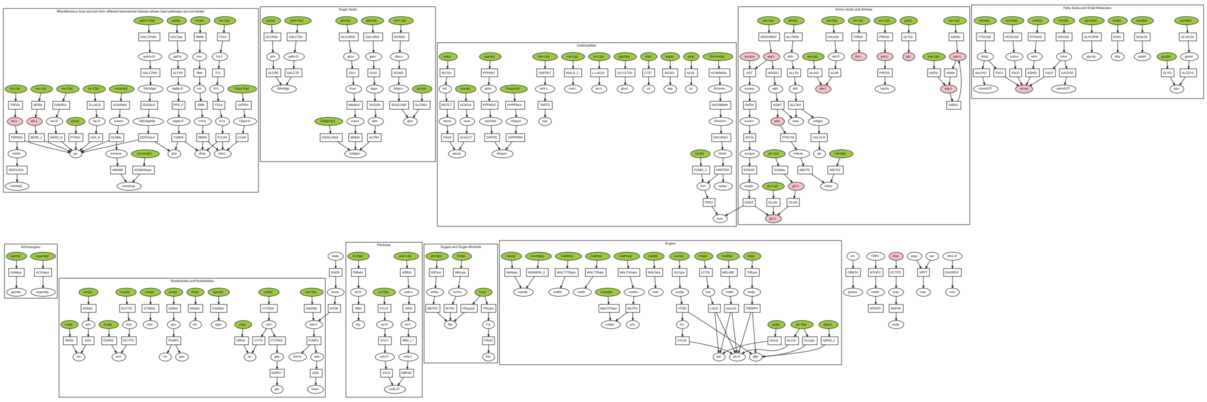

FIG. 4. (Color online) **Category I reactions in *E. coli.*** This figure shows the bipartite graph of 185 category I reactions in *E. coli*. Rectangles represent reactions and ovals metabolites. External nutrient metabolites (organic carbon sources) are depicted in green and biomass metabolites in pink. For convenience, we have chosen to omit the high degree currency metabolites (such as ATP) from the figure in order to reduce clutter and focus on the biochemically relevant transformation in each reaction. Abbreviation of metabolites and reactions are as in iJR904 model [20]. The figure was drawn using Graphviz software [34]. The high resolution electronic version of this figure can be zoomed in to read node labels and biochemical categories of boxes. We have classified the external metabolites and grouped together their input pathways in boxes based on biochemical similarity.

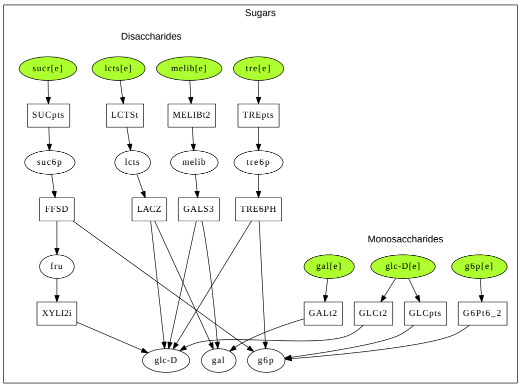

FIG. 5. (Color online) **A small portion of category I sub-network in *E. coli* showing sugar input pathways.** The figure shows category I reactions in the input pathways for external nutrient metabolites classified into the biochemical category 'Sugars'. Two kinds of sugars are shown here: monosaccharides and disaccharides. The input pathways for 7 external sugar metabolites fan-in downstream into 3 monosaccharide metabolites which occur at the boundary between category I and III sub-networks. Conventions are the same as in Figure 4.

tions is a sub-network with a distinct functional role in metabolism.

A. Category I: Fan-in of input pathways

Fig. 4 shows the sub-network of all 185 category I reactions in *E. coli.* The figure shows a number of essentially linear paths of one to about five reactions starting from an external nutrient metabolite, often converging to some other metabolite. These are the input pathways of those metabolites, typically starting from their transport reaction that brings them into the cell, and subsequent catabolic reactions that break them down into other metabolites. Input pathways of 86 out of the 89 external nutrient metabolites (carbon sources) characterizing different feasible minimal environments are contained in category I, thereby implying that category I essentially covers all the input pathways of metabolism. Similarly, we find that category I reactions in *S. cerevisiae* and *S. aureus* contain input pathways for most external nutrient metabolites characterizing different feasible minimal environments. Thus, category I essentially corresponds to input part of the metabolic network.

Fig. 5 shows a portion of category I reactions belonging to sugar input pathways in *E. coli* where several external sugar metabolites converge downstream into a few intermediate metabolites. Thus, the input pathways in category I exhibit the *fan-in* property whereby diverse external nutrient metabolites are first catabolized into a smaller set of intermediate metabolites before being drawn into the interior of the metabolic network. Usually the external nutrients whose input pathways converge to a common metabolite belong to the same biochemical class (cf. Figures 4 and 5). Fig. 4 contains a number of disconnected subgraphs each describing the input pathways of one or more biochemically similar metabolites; these disconnected paths get connected to the larger metabolic network via further downstream reactions that belong to other categories and are not shown in Fig. 4.

B. Category II: Output biosynthetic pathways

A key biological function of the metabolic network is to convert nutrient metabolites in the environment into biomass metabolites required for growth and maintenance of the cell. The biomass metabolites, which include all the amino acids, nucleotides, lipids and certain cofactors, may be considered to be the output of the metabolic network. Category II reactions are always-active and have a nonzero flux for any feasible minimal environment. We found that the category II sub-network has biosynthetic pathways for 30 out of the 49 biomass metabolites in *E. coli*. These pathways are typically the sole production pathways of those biomass metabolites in *E. coli* [23]. Thus, this sub-network is at the output end of the metabolism.

Of the 189 category II reactions in *E. coli*, 147 reactions belong to category IIa, whose fluxes are perfectly correlated across the different minimal environments. Fig. 6 shows the graph of the category IIa sub-network in *E. coli*, which is the single largest perfect cluster of reactions. The remaining 42 reactions in category II constitute the category IIb; these are always active but not perfectly correlated with category IIa reactions and with each other. Thus, the fluxes of category IIb reactions vary in a more complicated manner across minimal environments. Categories IIa and IIb exist with similar properties in the metabolic networks of the other two organisms (cf. Table I). In our previous work, we have shown that most of the category II reactions are essential for growth irrespective of the environment [23]. The set of category II reactions is a superset of reactions in the activity core found earlier by Almaas *et al* [35] which are reactions always used across minimal as well as rich environments.

C. Category III: Intermediate pathways between input and output

Fig. 7 shows the sub-network of category III reactions in *E. coli*, which are neither once-active nor always active; the activity of these reactions depends on the availability of nutrients in a more complicated manner. Category III reactions may be considered to constitute the intermediate part of the network. By comparing the structures of the three categories, it is evident that category III has a highly reticulate and complex architecture compared to categories I and II. There is a functional reason for the observed complexity in the category III sub-network. The biomass metabolites collectively contain several different types of chemical structures (moieties), and the *E. coli* metabolic network is capable of producing these biomass metabolites from different minimal environments, each containing a different (and single) carbon source. A typical external carbon source has one or a few moieties with different nutrients containing different subsets of moieties. Category I reactions transport the carbon sources into the cell and break it down into a small set of moieties. The function of category III reactions is to start with a small set of moieties and produce all the moieties required for biomass production. This requires a complex set of internal transformations and the exact set of transformations required depends on the nature of the input moieties. Thus, the activity of category III transforming reactions depends upon the biochemical nature of available nutrients in different minimal environments. We find that category III contains most of the reactions in central metabolism such as the citric acid cycle. A similar architecture of the category III sub-network was found in the metabolic networks of the other two organisms as well. Some of the biomass metabolites are produced in category III itself. For the other biomass metabolites category III produces precursors which are then taken up in the biosynthetic pathways of category II to produce the biomass metabolites.

V. COMPARISON OF FUNCTIONAL BOW-TIE DECOMPOSITION WITH GRAPH-THEORETIC BOW-TIE DECOMPOSITION

Ma and Zeng [11, 36] have used graph-theoretic measures to reveal a bow-tie architecture of metabolic networks similar to that seen in World Wide Web (WWW) [37], wherein the network can be decomposed into an in-component, out-component and a giant strong component. Given a directed graph, a strong component is a maximal subgraph such that for any pair of nodes i and j in the subgraph there exists a directed path from i to j and from j to i within the subgraph. In general, a directed graph can have many strong components, and the strong component with the largest number of nodes is designated as the giant strong component (GSC). The associated in-component consists of nodes which have access to GSC nodes via some directed path, but cannot be reached from any GSC node via a directed path. The out-component consists of nodes which can be reached from the GSC nodes via some directed path, but lack access to any GSC node via a directed path. A picture of the ideal graph-theoretic bow tie is shown in Fig. 8.

In this work, we have decomposed the metabolic network into three categories using a simple algorithm based on activity patterns of reactions across different minimal environments. Our categorization reveals a functional bow-tie architecture wherein the input pathways (category I reactions) fan into intermediate metabolism (category III reactions) which forms the knot of a bow-tie and from where the output pathways (category II reactions) for various biomass components fan out.

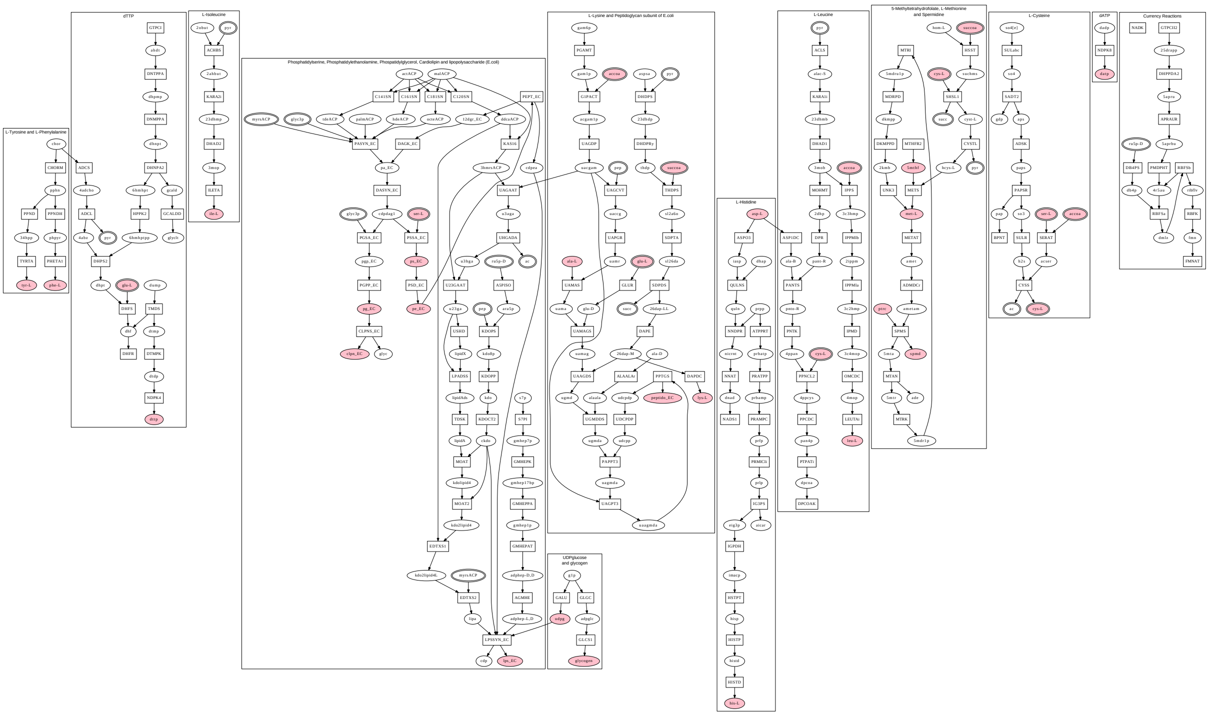
dTTP
L-Isoleucine
L-Tyrosine and L-Phenylalanine
Phosphatidylserine, Phosphatidylethanolamine, Phosphatidylglycerol, Cardiolipin and lipopolysaccharide (E.coli)
L-Lysine and Peptidoglycan subunit of E.coli
UDPglucose and glycogen
L-Histidine
L-Leucine
5-Methyltetrahydrofolate, L-Methionine and Spermidine
L-Cysteine
dATP
Currency Reactions

FIG. 6. (Color online) **Category IIa reactions in *E. coli.*** This figure shows the graph of 147 category IIa reactions in *E. coli* whose reaction fluxes are perfectly correlated across minimal environments. Conventions are the same as in Figure 4. The preponderance of biomass metabolites (pink ovals) in this figure signifies that these reactions are at the output end of the metabolic network. The reactions have been grouped together into boxes based on common biosynthetic pathways.

In our functional bow-tie decomposition, the three categories I, II and III of reactions discussed above broadly correspond to the in-component, out-component and GSC, respectively, of the graph-theoretic bow-tie decomposition by Ma and Zeng [11, 36]. However, the corresponding sets of reactions in the two decompositions differ in detail. For example, we find that the end products of several (and often long) chains of reactions in the category II sub-network are re-cycled resulting in feedback loops. Such feedback loops in the category II sub-network presumably minimize wastage and could be instrumental in producing the biomass metabolites in the desired ratios. An example of such a feedback loop in category II sub-network is the one involving metabolite 5mdr1p (which can be seen in the electronic version of Fig. 6 upon zooming). The biosynthetic pathways involved in such feedback loops appropriately belong to the output part of metabolism because they connect the precursor metabolites to the outputs. However, the graph-theoretic bow-tie decomposition would classify such category II reactions in feedback loops into the GSC. Thus, our functional bow-tie decomposition based on fluxes of reactions across different environments gives a better insight and is biochemically more realistic. The picture of the metabolic network our decomposition reveals is similar in spirit to the one envisioned by Csete and Doyle [12].

VI. DISCUSSION AND CONCLUSIONS

In this paper, we have performed flux balance analysis (FBA) for the metabolic networks of three microorganisms: *E. coli*, *S. cerevisiae* and *S. aureus* to obtain fluxes of reactions in the network under diverse environmental conditions. We have followed a purely algorithmic approach leveraging on the predicted fluxes of reactions across different minimal environments to decompose the metabolic network into functionally relevant sub-networks. We find that the activity of a reaction given by the number of minimal environments for which it has a nonzero flux is an important indicator of the functional role of a reaction. We have classified the reactions into three functional categories based on their activity. Category I contains once-active reactions which are used in only one minimal environment. Most reactions belonging to the category I sub-network are uptake pathways for external nutrients in feasible minimal environments, and the primary function of these reactions is to catabolize external nutrients into simpler metabolites which can be further processed by intermediary metabolism. Category II contains always active reactions which are used in all minimal environments. The category II sub-network is critical for the survival of the organism and accounts for the majority of the biosynthetic pathways for the production of the biomass metabolites at the output end of metabolic network. Category III contains reactions which are used in an intermediate number of min-

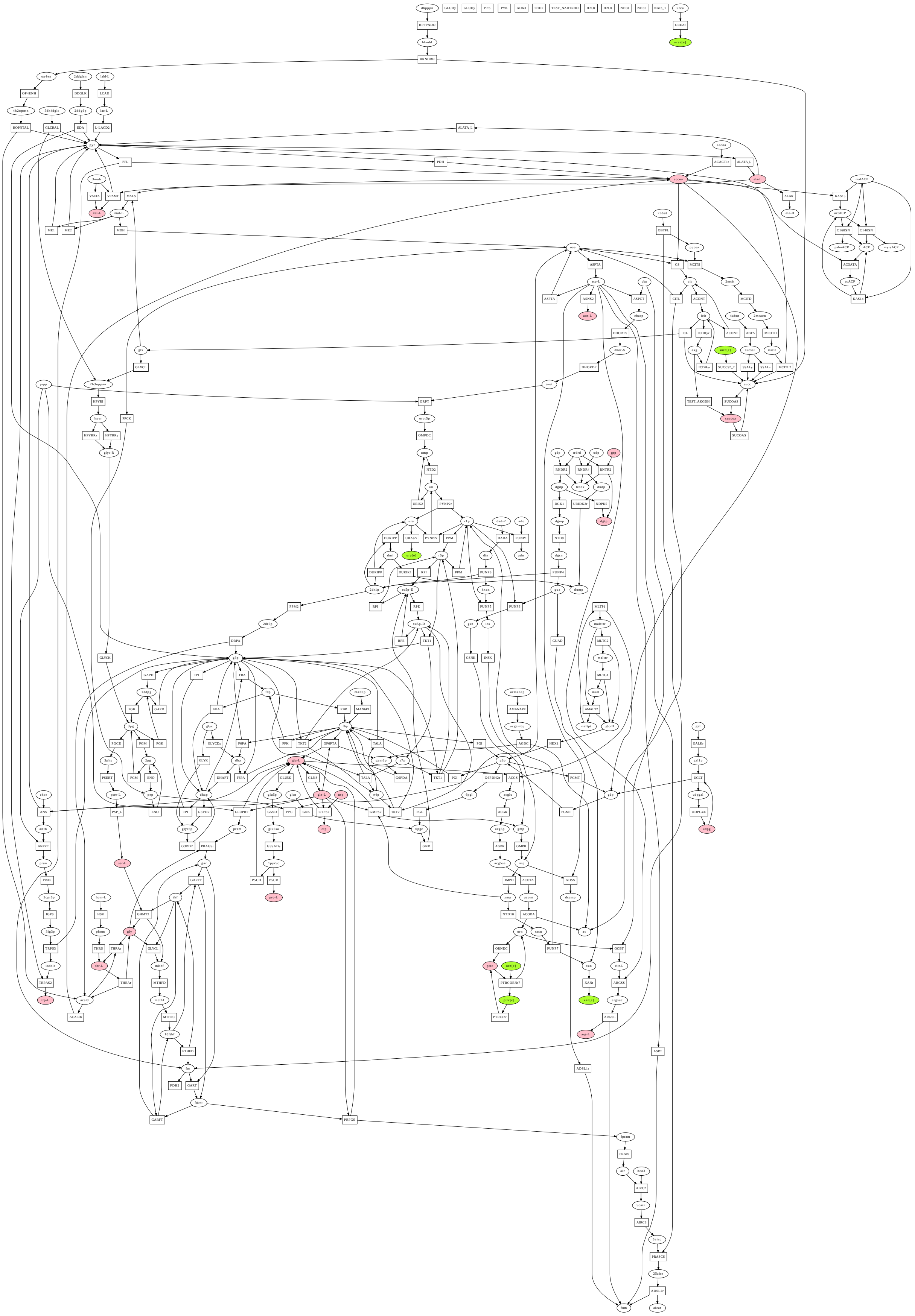

FIG. 7. (Color online) **Category III reactions in *E. coli.*** This figure shows the network of reactions that are active in two or more minimal environments considered, but not in all the environments. Conventions are the same as in Figure 3. Comparing this graph of category III reactions with category I and IIa reactions (cf. Figures 4 and 5), it is evident that category III sub-network has a highly reticulate structure with many loops.

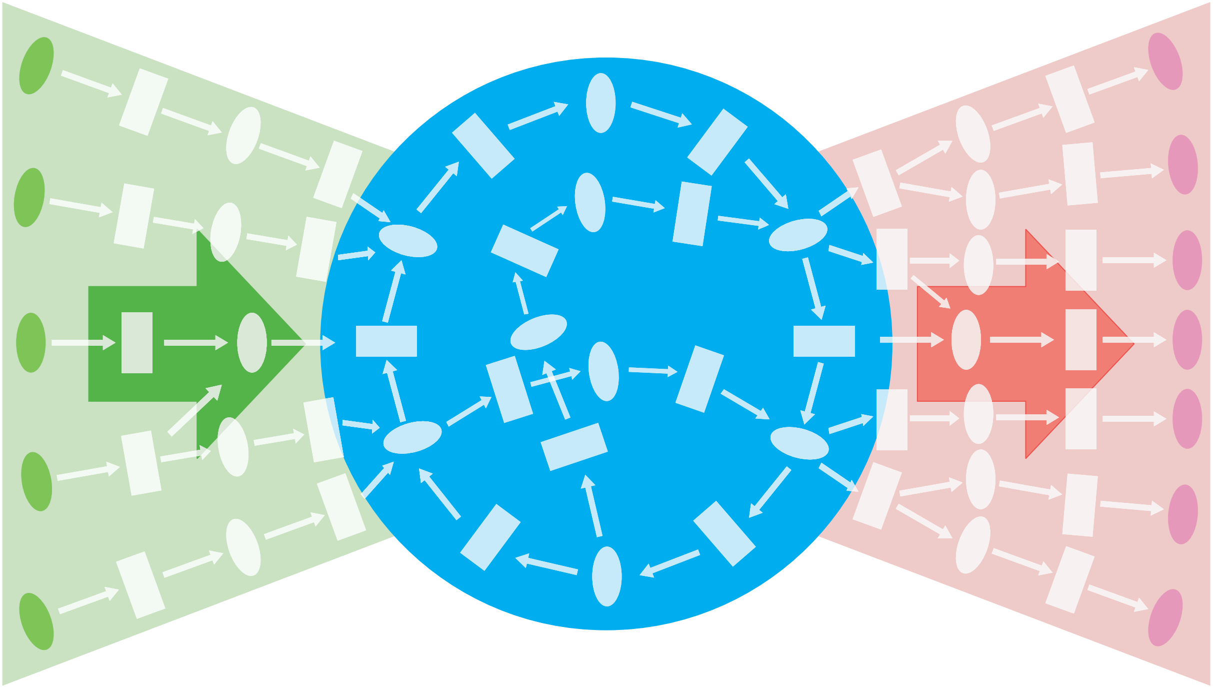

FIG. 8. (Color online) **The ideal graph-theoretic bow-tie for a directed bipartite graph.** The figure depicts the ideal bow-tie decomposition of a directed bipartite graph into three components: in, out and giant strong component corresponding to shaded regions green, pink and blue, respectively. Ovals represent objects (e.g., metabolites in the metabolic network) and rectangles processes (e.g., chemical reactions) that modify or combine objects to produce other objects. The figure shows pathways starting from the input nodes in the in component (green ovals in green region) and converging to an irreducible subgraph representing the giant strong component (blue region). Output paths fan out from the giant strong component and terminate in the output nodes in the out component (red ovals in pink region).

imal environments, and is responsible for generating the 'precursor' molecules that are eventually converted into biomass metabolites by Category II reactions. We find that while category I and II sub-networks are dominated by simple linear pathways, the structure of the category III sub-network is highly reticulate. In summary, our decomposition method for large-scale metabolic networks based on activity of reactions captures the proposed functional bow-tie organization by Csete and Doyle: the input pathways (category I reactions) for nutrients in the environment fan into intermediate metabolism (category III reactions) which forms the knot of bow-tie from where the output biosynthetic pathways (category II reactions) for biomass components fan out. Our results are valid for metabolic networks of three phylogenetically different organisms (two distinct prokaryotes and a eukaryote), which suggests that the observed functional bow-tie organization could be quite common in living systems.

Our functional classification of reactions uses an important additional piece of information that the purely graph-theoretic classification does not, namely, the list of the biomass metabolites that are the outputs of metabolism. The question arises as to whether the classification predicted by the graph-theoretic approach could be significantly improved by including this information (say, by somehow tagging the biomass metabolites in the graph). We think that this is unlikely. There does not seem to be any obvious method of utilizing this information in a purely topological analysis of the network. One might consider declaring these tagged metabolites to be present only at the output end of the network and thus exclude them (by hand) from the intermediate pathways. However, we note that while biosynthetic pathways of 30 of the biomass metabolites were found in category II reactions, several of the biomass metabolites were synthesized in the category III reactions. The latter metabolites such as alanine and valine are thus not only the outputs of metabolism, they also play an important role in the intermediate pathways required for the interconversion and synthesis of other metabolites. Thus a declaration such as the above would not be appropriate.

We remark that in the present work we have classified only the *reactions* of the metabolic network into three broad categories: input, output and intermediate. The classification of metabolites is more subtle and we intend to report on this in another contribution. While some metabolites participate in reactions belonging to only one of the three categories, several participate in reactions belonging to more than one category. The latter includes the currency metabolites such as ATP, ADP, NADP, NADPH, etc. It is important to note that our flux-based categorization of reactions does not involve the a priori exclusion of the high degree currency metabolites as was needed in the graph-theoretic bow-tie decomposition of the metabolic network [11, 36].

Cellular metabolism is only one of a large class of functional systems where inputs are transformed into outputs through 'reactions' or processes involving disintegrations, conversions, recombinations, etc. Other examples include any complex manufacturing facility, or even a production economy as a whole. Communication systems also share some of the features. The patterns of flows across the network as captured by the fluxes of the reactions carry important information about network architecture and functionality. The methods presented here could be useful in studying these patterns in fields other than cellular metabolism.

Appendix A: Robustness of categorization of reactions to alternate optimal solutions

In this work, flux balance analysis (FBA) was used to obtain a particular flux vector $\mathbf{v}$ or optimal solution that maximizes the objective function taken as the growth rate in a given minimal environment. However, for large-scale metabolic networks, there exist multiple flux vectors $\mathbf{v}$ or alternate optimal solutions that maximize growth in a given minimal environment, i.e., there are many flux vectors $\mathbf{v}$ with exactly the same value of the objective function but use different alternate pathways in the network [28, 30–32]. FBA finds one of many possible alternate optima for a given minimal environment that maximizes growth. In the main text, we have used a single optimal flux vector $\mathbf{v}$ for each of the M feasible minimal environments to determine the activity of a reaction and the set of active reactions in the metabolic network of an organism. Since, in principle, the activity of a reaction can change depending on the particular flux vector considered, we study the robustness of our categorization of reactions to the presence of alternate optima.

Flux variability analysis (FVA) [31] can be used to determine the set of reactions whose fluxes vary across alternate optima for a given minimal environment. Specifically, FVA determines the maximum and minimum flux value that each reaction can take across alternate optima for a given minimal environment. FVA involves the following steps:

(a) Determine using FBA the maximum value of the objective function Z or growth rate $v^{\alpha}_{biomass}$ in a given minimal environment α.

(b) Fix the flux of the biomass reaction equal to $v^{\alpha}_{biomass}$.

(c) Change the objective function Z to be the flux of a reaction j.

(d) Using linear programming determine the maximum flux value $v^{\alpha}_{j,max}$ of reaction j in the minimal environment α, constraining the biomass reaction to have a flux equal to $v^{\alpha}_{biomass}$.

(e) Using linear programming determine the minimum flux value $v^{\alpha}_{j,min}$ of reaction j in the minimal environment α, constraining the biomass reaction to have a flux equal to $v^{\alpha}_{biomass}$.

(f) The range $v^{\alpha}_{j,min}$ to $v^{\alpha}_{j,max}$ gives the variability of flux of reaction j across different alternate optima.

(g) The above steps c, d, e and f can be repeated for every reaction j in the metabolic network to determine the flux variability of each reaction across alternate optima for a given minimal environment α.

We have used FVA to determine $v^{\alpha}_{j,max}$ and $v^{\alpha}_{j,min}$ for each reaction j and for each feasible minimal environment α in the *E. coli* metabolic network. A reaction j is designated as *blocked* if $v^{\alpha}_{j,max}$=0 for all M feasible minimal environments [29, 38]. We found 329 blocked reactions in the *E. coli* metabolic network. The remaining 838 reactions, for which $v^{\alpha}_{j,max}$>0 for at least some environment α are designated as *potentially active* reactions. This set includes the 585 active reactions considered in the main text. We define a reaction j as *essential* for a given minimal environment α if $v^{\alpha}_{j,min}$>0. 484 reactions were found to be essential for some α in the *E. coli* metabolic network which are a subset of the 585 active reactions considered in the main text. We now classify these 484 reactions into the following three categories:

(a) Essential category I: Reactions which satisfy $v^{\alpha}_{j,min}$>0 for exactly one minimal environment. We found 162 reactions in the *E. coli* metabolic network to be in Essential category I. Of these, 153 reactions belong to category I of the main text.

(b) Essential category II: Reactions which satisfy $v^{\alpha}_{j,min}$>0 for all M minimal environments. We found 171 reactions in the *E. coli* metabolic network to be in Essential category II. All of these belong to category II of the main text.

(c) Essential category III: Reactions which satisfy $v^{\alpha}_{j,min}$>0 for m minimal environments where $1<m<M$. We found 151 reactions in the *E. coli* metabolic network to be in Essential category III. Of these, 145 belong to category III of the main text.

Thus we find that the classification discussed in the main text which uses a particular flux vector correctly predicts the essential category I, II or III of 469 out of the 484 essential reactions.

Appendix B: Relation between standard deviation σ and mean flux $\langle v\rangle$ for category I and category IIa reactions

In Fig. 3, we plot the standard deviation σ versus the mean flux $\langle v\rangle$ for active reactions in a metabolic network across its M feasible minimal environments. Here, we derive the relation between mean flux $\langle v\rangle$ and standard deviation σ for reactions in category I and category IIa shown as upper and lower lines, respectively, in Fig. 3.

1. Category I reactions

In a given organism any reaction belonging to category I has activity m=1, and is active for a single environment (say α_0). The mean flux $\langle v_j\rangle$ of a category I reaction j across M feasible environments is given by:

$$\begin{aligned}\langle v_j\rangle &= \frac{1}{M}\sum_{\alpha=1}^{M} v_j^\alpha \\ &= \frac{v_j^{\alpha_0}}{M},\end{aligned} \tag{B1}$$

where v_j^α is the flux of reaction j in the environment α ($\alpha = 1, 2, \ldots, M$). $v_j^{\alpha_0}$ is the flux of reaction j in the only feasible minimal environment α_0 where the reaction has nonzero value and in all other feasible minimal environments the flux of reaction j is 0.

Thus, the standard deviation σ_j for a category I reaction j is given by:

$$\begin{aligned}\sigma_j &= \sqrt{\frac{1}{M}\sum_{\alpha=1}^{M}(v_j^\alpha - \langle v_j\rangle)^2} \\ &= \sqrt{\frac{1}{M}[(M-1)\langle v_j\rangle^2 + (v_j^{\alpha_0} - \langle v_j\rangle)^2]} \\ &= \sqrt{M-1}\langle v_j\rangle,\end{aligned} \tag{B2}$$

where we have used the result in Eq. B1.

2. Category IIa reactions

The fluxes of reactions in category IIa are perfectly correlated with each other. This means that the fluxes of category IIa reactions are proportional to each other having the same proportionality constant for all minimal environments. Thus, for a minimal environment α, we can write the flux of category IIa reaction j as:

$$v_j^\alpha = c^\alpha v_j^0, \tag{B3}$$

where c^α is a constant for the minimal environment α and v_j^0 is some number. For any two reactions j and k in category IIa with fluxes correlated across minimal environments, we have:

$$\begin{aligned}\frac{v_j^\alpha}{v_k^\alpha} &= \frac{c^\alpha v_j^0}{c^\alpha v_k^0} \\ &= \frac{v_j^{\alpha'}}{v_k^{\alpha'}},\end{aligned} \tag{B4}$$

where α and α' are two different feasible minimal environments for the organism.

The mean flux of reaction j is:

$$\begin{aligned}\langle v_j\rangle &= \frac{1}{M}\sum_{\alpha=1}^{M} v_j^\alpha \\ &= v_j^0 \frac{1}{M}\sum_{\alpha=1}^{M} c^\alpha \\ &= v_j^0\langle c\rangle,\end{aligned} \tag{B5}$$

where $\langle c\rangle$ is the mean of c^α across the set of feasible minimal environments.

The standard deviation σ_j for category IIa reaction j is given by:

$$\begin{aligned}\sigma_j &= \sqrt{\frac{1}{M}\sum_{\alpha=1}^{M}(v_j^\alpha - \langle v_j\rangle)^2} \\ &= v_j^0\sqrt{\frac{1}{M}\sum_{\alpha=1}^{M}(c^\alpha - \langle c\rangle)^2} \\ &= v_j^0\sigma_c \\ &= \frac{\sigma_c\langle v_j\rangle}{\langle c\rangle} \\ &= b\langle v_j\rangle,\end{aligned} \tag{B6}$$

where we have used the result in Eq. B5.

ACKNOWLEDGMENTS

SS and VG acknowledge support from University Grants Commission (UGC), AS from Council for Scientific and Industrial Research (CSIR), and SJ from Department of Biotechnology (DBT), India.

[1] L. Hartwell, J. Hopfield, S. Leibler, and A. Murray, Nature **402**, C47 (1999).

[2] S. Bornholdt, H. Schuster, and J. Wiley, *Handbook of graphs and networks*, Vol. 2 (Wiley Online Library, 2003).

[3] A. Barabási and Z. Oltvai, Nature Reviews Genetics **5**, 101 (2004).

[4] A. Wagner, *Robustness and evolvability in living systems* (Princeton University Press Princeton, NJ:, 2005).

[5] K. Sneppen and G. Zocchi, *Physics in molecular biology* (Cambridge University Press, 2005).

[6] U. Alon, *An introduction to systems biology: design principles of biological circuits*, Vol. 10 (Chapman & Hall/CRC, 2006).

[7] K. Kaneko, *Life: An introduction to complex systems biology*, Vol. 171 (Springer Heidelberg, Germany:, 2006).

[8] R. Heinrich and S. Schuster, *The regulation of cellular systems*, Vol. 416 (Chapman & Hall New York, 1996).

[9] H. Jeong, B. Tombor, R. Albert, Z. Oltvai, and A. Barabási, Nature **407**, 651 (2000).

[10] A. Wagner and D. Fell, Proceedings of the Royal Society of London. Series B: Biological Sciences **268**, 1803 (2001).

[11] H. Ma and A. Zeng, Bioinformatics **19**, 1423 (2003).

[12] M. Csete and J. Doyle, Trends in Biotechnology **22**, 446 (2004).

[13] B. Palsson, *Systems biology: properties of reconstructed networks* (Cambridge University Press, 2006).

[14] N. Price, J. Reed, and B. Palsson, Nature Reviews Microbiology **2**, 886 (2004).

[15] A. Feist and B. Palsson, Nature biotechnology **26**, 659 (2008).

[16] M. Oberhardt, B. Palsson, and J. Papin, Molecular Systems Biology **5** (2009).

[17] J. Edwards, R. Ibarra, B. Palsson, *et al.*, Nature Biotechnology **19**, 125 (2001).

[18] R. Ibarra, J. Edwards, and B. Palsson, Nature **420**, 186 (2002).

[19] D. Segre, D. Vitkup, and G. Church, Proceedings of the National Academy of Sciences **99**, 15112 (2002).

[20] J. Reed, T. Vo, C. Schilling, B. Palsson, *et al.*, Genome Biol **4**, R54 (2003).

[21] N. Duarte, M. Herrgård, and B. Palsson, Genome Research **14**, 1298 (2004).

[22] S. Becker and B. Palsson, BMC Microbiology **5**, 8 (2005).

[23] A. Samal, S. Singh, V. Giri, S. Krishna, N. Raghuram, and S. Jain, BMC bioinformatics **7**, 118 (2006).

[24] M. Eisen, P. Spellman, P. Brown, and D. Botstein, Proceedings of the National Academy of Sciences **95**, 14863 (1998).

[25] T. Pfeiffer, F. Montero, S. Schuster, *et al.*, Bioinformatics **15**, 251 (1999).

[26] J. Stelling, S. Klamt, K. Bettenbrock, S. Schuster, E. Gilles, *et al.*, Nature **420**, 190 (2002).

[27] J. Papin, N. Price, and B. Palsson, Genome Research **12**, 1889 (2002).

[28] J. Reed and B. Palsson, Genome Research **14**, 1797 (2004).

[29] A. Burgard, E. Nikolaev, C. Schilling, and C. Maranas, Genome Research **14**, 301 (2004).

[30] S. Lee, C. Phalakornkule, M. Domach, and I. Grossmann, Computers & Chemical Engineering **24**, 711 (2000).

[31] R. Mahadevan, C. Schilling, *et al.*, Metabolic engineering **5**, 264 (2003).

[32] A. Samal, Systems and synthetic biology **2**, 83 (2008).

[33] E. Almaas, B. Kovacs, T. Vicsek, Z. Oltvai, and A. Barabási, Nature **427**, 839 (2004).

[34] J. Ellson, E. Gansner, L. Koutsofios, S. North, and G. Woodhull, in *Graph Drawing* (Springer, 2002) pp. 594–597.

[35] E. Almaas, Z. Oltvai, and A. Barabási, PLoS Computational Biology **1**, e68 (2005).

[36] H. Ma, X. Zhao, Y. Yuan, and A. Zeng, Bioinformatics **20**, 1870 (2004).

[37] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener, Computer networks **33**, 309 (2000).

[38] S. Schuster and R. Schuster, Journal of Mathematical Chemistry **6**, 17 (1991).