Molecular BioSystems

Cite this: *Mol. BioSyst.*, 2012, **8**, 843–850

www.rsc.org/molecularbiosystems

**PAPER**

# Uncovering the hidden geometry behind metabolic networks†

**M. Ángeles Serrano,*[a] Marián Boguñá[b] and Francesc Sagués[a]**

***Received 25th July 2011, Accepted 30th November 2011***
**DOI: 10.1039/c2mb05306c**

Metabolism is a fascinating cell machinery underlying life and disease and genome-scale reconstructions provide us with a captivating view of its complexity. However, deciphering the relationship between metabolic structure and function remains a major challenge. In particular, turning observed structural regularities into organizing principles underlying systemic functions is a crucial task that can be significantly addressed after endowing complex network representations of metabolism with the notion of geometric distance. Here, we design a cartographic map of metabolic networks by embedding them into a simple geometry that provides a natural explanation for their observed network topology and that codifies node proximity as a measure of hidden structural similarities. We assume a simple and general connectivity law that gives more probability of interaction to metabolite/reaction pairs which are closer in the hidden space. Remarkably, we find an astonishing congruency between the architecture of *E. coli* and human cell metabolisms and the underlying geometry. In addition, the formalism unveils a backbone-like structure of connected biochemical pathways on the basis of a quantitative cross-talk. Pathways thus acquire a new perspective which challenges their classical view as self-contained functional units.



## Introduction

Cells are self-organized entities that carryout specialized tasks at different interrelated omic-levels[1] involving different actors, from codifying genes to energy-carriers or constitutive metabolites. A key towards understanding this complex architecture at a systems level is provided by reliable genome-wide reconstructions of the set of biochemical reactions that underlie the functional cell machinery.[2] Such reconstructions can be analyzed using tools and techniques from complex networks theory,[3–5] a discipline that is being used in the characterization of biological, chemical, infrastructural, technological or social-based systems of complex relationships.[6,7] More precisely, nodes in metabolic networks account for either metabolites or reactions, while links represent the interactions among them. Apart from providing a large-scale organizational picture, these network-based representations have permitted to analyze sensible issues in cellular metabolism, like flux balances,[8,9] regulation,[10] robustness,[11] or reaction reliability.[12]

The advantage of using network-based representations, in whatever context we employ them, may be arguably questioned by the fact that complex networks are customarily modeled as pure topological constructions lacking a true geometric measure of separation among nodes. This is aggravated by the fact that complex networks have the small-world property,[13] meaning that every pair of nodes in the system are very close in topological distance. This is an important and obvious degeneracy if we think in terms of optimizing routing or transportation strategies in man-engineered networks, but can be equally crucial when referring to the description of the metabolic functioning at a single cell level. As a matter of fact, the related attempt of separating nodes into communities, that has been already pursued in different contexts[14] and, in particular, applied to metabolic networks,[15] has proven to be an extremely difficult task. Classical community detection approaches turn out to be *a posteriori* classification methods, and do not provide insights into any potential connectivity law underlying the observed topology. These questions could be significantly addressed by quantifying the abstract concept of node proximity in terms of a metric distance which could be combined into a simple and general probabilistic connectivity law. Such a biochemical connectivity law, relying on metric distances, may provide a simple explanation of the large-scale topological structure observed in metabolism,[16] and it can also be used, like in this work, to revisit the concept of biochemical pathways.

In this paper, we uncover the hidden geometry of the *E. coli* and human metabolisms and find that their network topologies obey an extremely simple and powerful—metric-based—probabilistic connectivity law. In particular, given a pair metabolite/reaction separated by a geometric distance $d_{mr}$ in the underlying metric space, the probability of existence of a connection between them is here shown to be a decreasing

[a] *Departament de Química Física, Universitat de Barcelona, Barcelona, Spain. E-mail: marian.serrano@ub.edu*
[b] *Departament de Física Fonamental, Universitat de Barcelona, Barcelona, Spain*
† Electronic supplementary information (ESI) available. See DOI: 10.1039/c2mb05306c

function of the effective distance $d_{\text{eff}} \equiv d_{mr}/(k_r k_m)$, where degrees $k_m$ and $k_r$ count the number of their respective neighboring nodes. The geometric distance $d_{mr}$—a measure of structural affinity between metabolites and reactions—is in this way modulated by the product of degrees of the two involved nodes, so that the degree heterogeneity observed in the metabolic network is properly taken into account.[17] Naturally, a key ingredient in our approach concerns the suitable geometry substantiating this distance. We find that a simple one dimensional closed Euclidean space, *i.e.* a circle, when combined with the network degree heterogeneity is enough to capture the global organization of the network. Using statistical inference techniques, we find angle-based coordinates in this space for the full set of metabolites and reactions, which expose the extraordinary congruency of our model.

As a direct application of our model, we compare the results of our embedding with the standard classification of reactions in terms of biochemical pathways. Such a reaction-aggregated analysis reveals rather disparate trends when pathways are characterized in terms of the circle-based localizations of their constituent reactions. Some specific pathways appear concentrated over narrow sectors of polar angles, while more transversal ones are widespread over the circle. This points to a diversity of pathway topologies, with some of them displaying groups of densely interconnected reactions while some others evidencing a much more weakly connected internal structure. Moreover, pathways themselves admit to be linked using the discovered connectivity law. This strategy reveals different levels of cross-talk between pathways, leading to a coarse-grained view of metabolic networks or, in other words, to the build-up of networks of pathways. Such a higher level in the hierarchical organization of metabolic networks advises against the study of pathways as autonomous subsystems and should permit us to calibrate more accurately how a pathway-localized perturbation spreads over the entire network.

## Results

### Embedding algorithm and validation

A simple abstraction of a given metabolism is given by its bipartite network representation. This amounts to consider metabolites and reactions as belonging to different subsets of nodes, with metabolites (irrespectively considered as reactants and products) linked to all reactions they take part in, thus avoiding connections between nodes of the same kind, see Fig. 1a. The first step towards mapping this network consists in defining a geometric model that can advantageously represent it. The simplest metric space that can globally embed a network is a circle of radius $R$. This is the simplest choice such that there are no *a priori* preferred locations and, therefore, any inhomogeneity in the final angular distribution is dictated by the network structure itself. Nodes, in our case metabolites and reactions separately, are distributed on it according to specific angular coordinates to be determined. The whole strategy to find these coordinates rests on a precise definition of the interactions between nodes in terms of their angular separation in the circle. We prescribe a connection probability between a reaction $r$ and a metabolite $m$, with respective bipartite degrees $k_r$ and $k_m$ and separated by a distance $d_{mr}$ on the circle ($d_{mr} = R\Delta\theta_{mr}$, $\Delta\theta_{mr}$ being the angular separation between the metabolite and reaction) to be a decreasing function of such distance rescaled by the product of node degrees,[18]

$$\text{Prob}\{m \text{ is connected to } r\} \equiv p\left(\frac{d_{mr}}{k_m k_r}\right). \qquad (1)$$

It is worth stressing that this is the central and unique law underlying the whole formalism. Note that this choice is particularly suggestive since by identifying the node degree

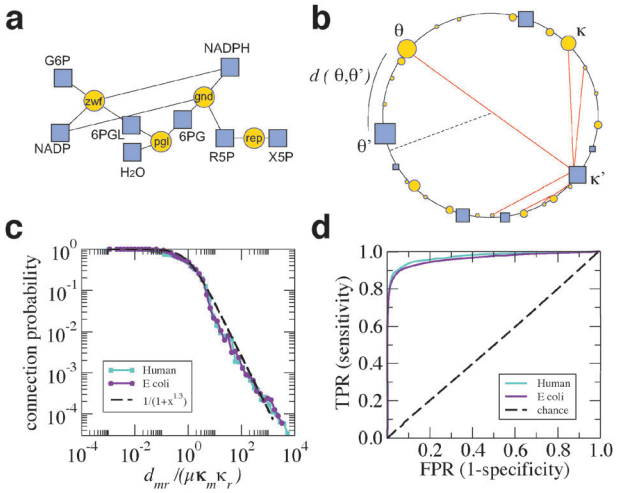


**Fig. 1** Model and empirical validation. (a) Bipartite network representation of four coupled stoichiometric equations in the pentose-phosphate pathway of *E. coli*. Reaction acronyms stand for the catalyzing enzyme: *zwf*, glucose-6-phosphate dehydrogenase [EC 1.1.1. 49]; *pgl*, 6- phospho-gluconolactonase [EC 3. 1. 1.31]; *gnd*, 6- phosphogluconate dehydrogenase [EC 1.1.1. 43]; *rpe*, ribulose-phosphate 3-epimerase [EC 5. 1.3. 1]. Note that connections (black lines) are always between reactions (yellow circles) and metabolites (blue squares), metabolites or reactions are never connected among themselves. (b) A sketch of the $\mathbb{S}^1 \times \mathbb{S}^1$ model. Nodes are randomly distributed in the circle and given expected degrees, symbolically represented by the sizes of the nodes. The distance between two nodes is computed as the length of the arc separating the nodes. Due to the peculiar rescaling of distances by degrees in eqn (1), a node can connect not only to nearby nodes but also to far apart nodes with large degrees. (c) The plot shows a comparison between the empirical connection probability for the *E. coli* and human metabolisms and the theoretical one given in eqn (2). The empirical connection probability is computed as the fraction between the number of actual connections at effective distance $d_{mr}/\mu k_m k_r$ and the total number of pairs at the same effective distance. (d) The Receiver Operating Characteristic (ROC) curve computed for our model for the *E. coli* and human metabolisms is shown. To calculate the ROC curves, we rank (from highest to lowest) the connection probabilities given by the model for all possible pairs metabolite/reaction (either present or absent) using the previously inferred coordinates. We then define a threshold probability that allows us to discriminate between positive interactions (those above the threshold) from negative ones (those below the threshold) and to compute the fraction of true positive connections (True Positive Rate, TPR) and that of false positive connections (False Positive Rate, FPR), with the understanding that a true positive connection is an observed link above the threshold, while a false positive is a non-existing one above the threshold (note that the terminology positive and negative connections has nothing to do here with a potential positive or negative biological effect). Each threshold value corresponds to one point of the ROC curve. The threshold is scanned throughout the whole range of probabilities to produce the complete ROC curve.

 

as a measure of its mass, this interaction mimics the Newtonian form of gravitational interaction. More precisely, the explicit form for the above interaction reads

$$p\left(\frac{d_{mr}}{k_m k_r}\right) = \frac{1}{1 + (d_{mr}/\mu k_m k_r)^{\beta}}. \qquad (2)$$

This particular prescription combines, in a simple way, the classical network topological concept of node degrees with the newly introduced notion of geometric distance. All in all, this functional form expresses an intuitive view, *i.e.* closer nodes in the metric space are more likely to be linked, while nodes with higher degrees sustain farther reaching connections regardless of their distances. Fig. 1b shows a visual sketch summarizing the basic trends of the bipartite formalism just outlined. We refer to it with the notation $\mathbb{S}^1 \times \mathbb{S}^1$, see *Methods*. Besides, this model gives rise to an ensemble of graphs that are maximally random given their specific constraints.[19,20] Finally, parameters $\mu$ and $\beta$ are consistently determined to reproduce the statistical properties of the original network. Parameter $\mu$ fixes the total number of edges, whereas $\beta$ controls clustering, *i.e.*, a measure of short-range loops, see ESI.†

To infer the angle-based coordinates for metabolites and reactions in the ring we use a two-step procedure. Starting from the original bipartite network, we first perform a one-mode projection over the set of metabolites by connecting two metabolites whenever they participate in the same reaction. We then circle-embed such a unipartite metabolites network applying the unipartite version of the formalism as described earlier.[21] Finally, using this partial allocation as an initial fixed template, we complete the embedding of the reactions by invoking a maximum likelihood inference strategy (the detailed description of the embedding algorithm and the coordinates of metabolites and reactions are fully reported in ESI†).

We apply our formalism to the iAF1260 version of the K12 MG1655 strain of *E. coli* metabolism[22] and to human cell metabolism,[23] both provided in the BiGG database,[24,25] see ESI.† Before presenting the embedding for these metabolic networks, we comment on the validation of the proposed mapping procedure. We first perform a direct calibration which amounts to compare the set of observed metabolite–reaction connection probabilities in the original reconstructions with the theoretical connection probability given by eqn (2). Explicit results are presented in Fig. 1c, both for *E. coli* and human metabolisms. Besides the striking agreement between observed and predicted connections, it is worth noting that the two analyzed networks are perfectly represented with the same $\beta$ exponent fitted to a value $\beta = 1.3$. We also check the discrimination power of our algorithm by computing the Receiver Operating Characteristics (ROC) curve of our model,[26] which compares the true positive rate (TPR) *vs.* the false positive rate (FPR) and informs us about how good is our method at correctly discerning real links. Results are shown in Fig. 1d. When representing the TPR in front of the FPR, a totally random guess would result in a straight line along the diagonal. In contrast, the ROC curve of our model lies far above the diagonal, which indicates a remarkable discrimination power. A convenient summary statistic can be defined as the area under the ROC curve (AUC statistic), which represents the probability that a randomly chosen observed link in the network has a higher probability of existence according to the model than a randomly chosen non-existing one. This statistic ranges in the interval [0.5,1], being AUC = 0.5 a random prediction and AUC = 1 a perfect prediction. In our case, values are AUC = 0.96 for *E. coli* and AUC = 0.97 for human metabolism. Both validation tests confirm that our model adjusts nearly perfectly to the real data.

Fig. 2 shows the embedding representation of the *E. coli* metabolism (the mapping of the human metabolism is provided in the ESI†). For the sake of clarity, metabolites are displaced towards the center of the circle by an amount proportional to their degree so that hub metabolites are close to the center of the disk whereas low degree ones are placed in the periphery. The distribution over the circle is far from being uniform as it could be naively expected. Indeed, this is a distinctive signature of the delicate structural organization of metabolic networks. In particular, different levels of aggregation are readily visible, in as much as human settlements that are unevenly distributed in population maps. Simultaneously with densely occupied areas, empty regions are visible and appear irregularly punctuated with occasional metabolite–reaction associations. As a whole, this landscape is an indication of some hierarchical trends existing in the analyzed networks and prompts us to look for eventual higher organizational levels. In this regard, we revise the biochemical concept of pathways, classically understood as chains of step-by-step reactions which transform a principal chemical into another either for immediate use, to propagate metabolic fluxes or for cell storage. In Fig. 2, we identify pathways in the circle by plotting their names at the average angular position of all their constitutive reactions.

## Pathway localization

In Fig. 3 and 4, we propose two complementary representations of the metabolic pathways of *E. coli* as they appear annotated in the BiGG database. In Fig. 3, we show the angular distribution on the ring of the whole list of pathways (up to 33, plus an *unassigned* category of reactions not represented in the figure), evaluated from the circle-based embedding of the reactions they involve. We recognize rather disparate spectra of angular distributions. Strongly localized pathways, *e.g.* the Folate pathway or Oxidative Phosphorylation, coexist with more distributed ones. The latter can adopt either an angular distribution with two or more maxima, *e.g.* the Histidine and Glycolysis pathways, respectively, or can even transversally spread over the ring closer to a homogeneous distribution. The Alternate Carbon, the Transport Inner Membrane, or the Cofactor and Prosthetic Group pathways are representative examples of the latter category (see Table S1 in ESI† for further details). Our method is, therefore, able to discriminate concentrated pathways, consistent with the classical view of modular subsystems, from others which are indeed formed of subunits, and even from those finally responsible for producing or consuming metabolites in turn extensively used by many other pathways.

The embedding of reactions and metabolites in the circle can also be used to aggregate pathways into broader categories. To do so, the embedding circle is first divided into eleven different

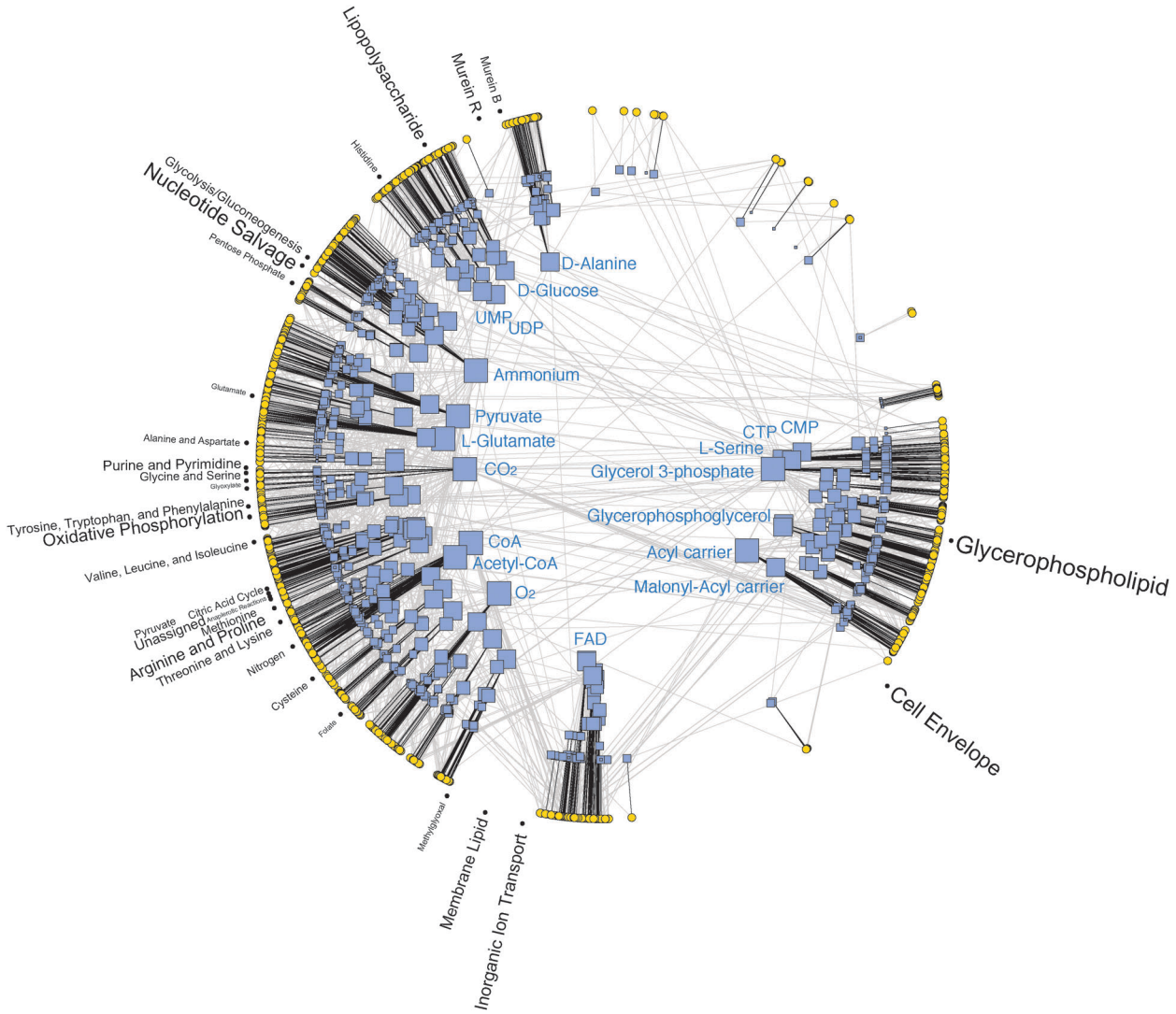


**Fig. 2** Global geometric map of *E. coli*'s metabolism. Angular distribution of reactions and metabolites inferred by the method. Yellow circles represent reactions whereas blue squares are metabolites. For each metabolite, the symbol size is proportional to the logarithm of the degree and radially placed according to the expression $r = R - 2\ln k_m$. Black (grey) connections are those that according to the model have a probability of existence larger (smaller) than 0.5. The names of the different pathways, radially-written, are located at the average angular position of all the reactions belonging to a given pathway, and the font size is proportional to the logarithm of the number of reactions in the pathway. Note that we do not represent transversal pathways and that some pathways seem to be located in empty regions (*e.g.* Inorganic Ion Transport). This is due to the fact that some pathways display angular distributions with two or more maxima so that the average appears in between the peaks, see Fig. 3 and Table S1 in ESI.†

angular sectors delimited by void regions in the ranked distribution of reaction angles, see Fig. 4a and ESI† for a precise definition of these angular sectors. These sectors stand as potential biochemical modules defined by chemical affinity (distance in the metric space). The pathway concentration, *i.e.* the fraction of reactions of that pathway in each sector, is shown in Fig. 4b–l. Clearly, there are sectors monopolized by one or at most two pathways—*e.g.*, Murein in Sector 10, Fig. 4k—whereas other sectors are largely shared by many pathways—*e.g.*, different Amino acid-based pathways in Sector 7, Fig. 4h. In all cases, the higher concentrations in each sector mostly correspond to pathways in related functional categories. Sector 1 and Sector 2 (Fig. 4b and c) aggregate pathways related to Cell Membrane metabolism. Biochemical reactions in Sectors 3–5 (Fig. 4d–f) focus on the processing of methyl groups, with Sector 3 aggregating part of the related sulfur-containing amino acid pathways Cysteine and Methionine, Sector 4 dominated by the Folate pathway (folate is for instance necessary for the regeneration of methionine), and Sector 5 basically populated by reactions in the Nitrogen and Methylglyoxal pathways, both producing metabolites upstream of central metabolism. Sector 6 and Sector 7 (Fig. 4g and h) concentrate on central metabolism, with Sector 6 including Energy and part of the Nucleotide metabolism and Sector 7 including Amino acids metabolism. Sector 8 (Fig. 4i) condenses the remaining Nucleotide metabolism. Finally, Sector 9, Sector 10, and Sector 11 (Fig. 4j–l) account for Glycan metabolism, with Sector 9 mixing basically mono and polysaccharide related pathways, Sector 10 comprising pathways related to murein, a polymer that forms the cell wall, and Sector 11 containing the remaining fraction of the Methylglyoxal pathway (producing pyruvate redirected to cell wall biosynthesis) and an important

 

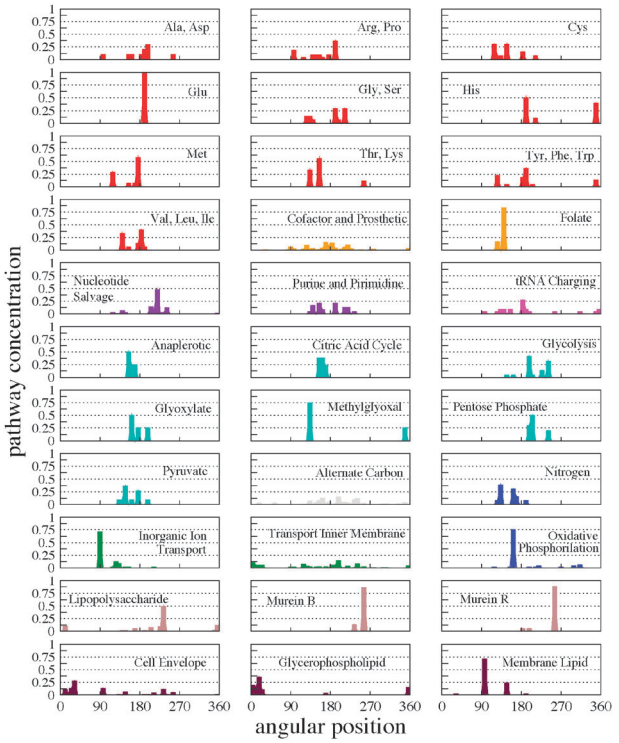


**Fig. 3** Angular distribution of biological pathways in *E. coli*. The whole angular domain [0°,360°] is divided into 50 bins of 7.2° each and for each bin we compute the fraction of reactions of the pathway in it. Each pathway is shown in a different graph. Different colors indicate different general metabolic classes: red for Amino acids metabolism (numbering the graphs from left to right and from top to bottom, 1–10), orange for metabolism of Cofactors and Vitamins (11 and 12), violet for Nucleotide metabolism (13 and 14), magenta for tRNA charging (15), turquoise for Carbohydrate metabolism (16–22), grey for Alternate Carbon metabolism (23), blue for Energy metabolism (24,27), green for Transport pathways (25 and 26), brown for Glycan metabolism (28–30), and maroon for Lipid metabolism (31–33). Pathway names have been abbreviated in standard forms whenever possible, see ESI.†

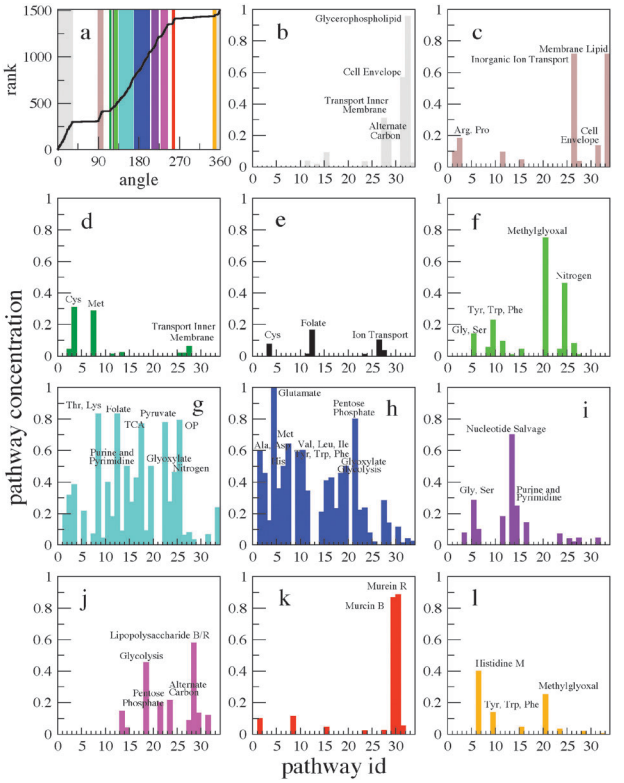


**Fig. 4** Sector modules for *E. coli* metabolism. Reactions in related functional categories are observed to aggregate in specific regions of the circle. The whole angular domain is divided into eleven different angular sectors delimited by void regions in the ranked distribution of reaction angles. This distribution and the angular sectors (each in a different color) are given in the left upper graph of the panel. The remaining graphs show the pathway concentration, the fraction of reactions of that pathway, in each sector. The higher concentrations in each sector mostly correspond to pathways in related functional categories. S1 and S2 (plots b and c) aggregate pathways related to Cell Membrane metabolism; Sectors 3–5 (plots d–f) are focused on the processing of methyl groups, with Sector 3 aggregating part of the sulfur-containing amino acid pathways Cysteine and Methionine, Sector 4 dominated by the Folate pathway (folate is for instance necessary for the regeneration of methionine), and S5 populated by reactions in the Nitrogen and Methylglyoxal pathways, both producing metabolites upstream of central metabolism; S6 and S7 (plots g and h) concentrate on central metabolism, with S6 including Energy and part of the Nucleotide metabolism and S7 including Amino acids metabolism; S8 (plot i) condenses the remaining Nucleotide metabolism; finally, S9–S11 (plots j–l) account for Glycan metabolism, with S9 mixing mono and polysaccharide related pathways, S10 comprising pathways related to murein, a polymer that forms the cell wall, and S11 containing the remaining fraction of the Methylglyoxal pathway and an important part of the metabolism of Histidine, both with a role in the cell wall biosynthesis. Pathway names have been abbreviated in standard forms whenever necessary, see ESI.†

part of the metabolism of Histidine (also with a regulatory role in the same function).

The corresponding representations for human metabolism are shown in the ESI†, Fig. S4 and S5. The number of pathways is considerably larger but common features to *E. coli* pathway localization patterns are evidenced in qualitative terms. Pathways can be divided again into different categories according to their angular concentration, with the difference that the general level of pathway localization in human metabolism is higher than in *E. coli*. The average angular concentration of pathways in human metabolism is 0.82, as compared to 0.79 in *E. coli* (see *Methods*) and the average size of maximum peaks in the pathways angular distributions is 0.36 for *E. coli* while for human metabolism it is 0.50. However, the higher level of localization seems to coexist with a higher entanglement of the different families of metabolic reactions, *i.e.* carbon metabolism, lipid metabolism, *etc*. Another observation is that transversal pathways in *E. coli*, like Cofactor and Prosthetic group or Transport, are split into a number of more specialized pathways in human metabolism and, in fact, the category of transversal pathways itself, as defined in *E. coli*, is here minimally represented.

### Cross-talk between pathways

Sectors, as proposed in the previous section, stand as potential biochemical modules defined by chemical affinity (distance in the metric space). However, classical pathways have been considered until now to be meaningful functional groups.

Therefore, it is natural to reanalyze classical biochemical pathways at the new light of our embedding. One way is to look at the angular distribution of their reactions in the circle, as it is done in Fig. 3. Another way is to take advantage of the estimated connection probability derived from the embedding for every pair reaction–metabolite to compute the strength of the interaction between every pair of pathways. This information can be used to build a higher hierarchical level in the architecture of the metabolic network: the network of pathways.

In this new network, pathways are the nodes whereas the strength of the interaction between a pair of pathways is computed on the basis of the corresponding lists of reactions in each pathway and the set of metabolites shared by both. When the set of overlapping metabolites is not empty, the connection probabilities for the observed links between pathway reactions and common metabolites are summed to give an absolute measure of the strength of the interaction between the pair of pathways, see ESI.† Overlaps between pathway pairs assemble a higher order weighted network where pathways are nodes and links display heterogeneous intensities. However, the resulting network is very dense and needs to be conveniently filtered in order to provide meaningful information about the system. In *E. coli*, 460 out of a potential total of 561 pathway pairs overlap while for human cells 1689 pathway pairs out of 4278 have common metabolites. In practice we use a disparity-based threshold[27] (see *Methods*) that discards links whose intensities are compatible with random fluctuations at some specific significance level. As a result these pathway-based networks provide metabolic backbones, *i.e.*, subnetworks of pathways which display the statistically relevant interactions.

As an illustration of the power of the metabolic backbone concept, panels in Fig. 5 reproduce backbones for *E. coli* and human metabolisms. We selected those with the closest confidence level to the standard values 0.05 and 0.01, respectively, that optimize the trade-off between maximum number of pathways and weight *versus* minimum number of interactions in the filtered network. Interestingly, metabolic backbones offer a perspective that reveals functional constraints. Both for *E. coli* and human metabolism, star-like patterns are particularly neat. In *E. coli*, transversal pathways act as hub-like structures that interconnect different number of specific and more localized pathways, usually belonging to the same metabolic family. For instance, the Cofactor and Prosthetic Group Biosynthesis pathway connects many of the amino acid pathways to energy or nucleotide metabolism, and Alternate Carbon acts as the main intermediary of many Carbohydrate pathways with the rest of the backbone. Analogously, some pathways in the metabolic backbone of the human cell, like Folate or Fatty Acid Oxidation or Keratan Sulfate Biosynthesis, play a relevant role in providing systems' level connectivity to the network and connect a number of other specific pathways.

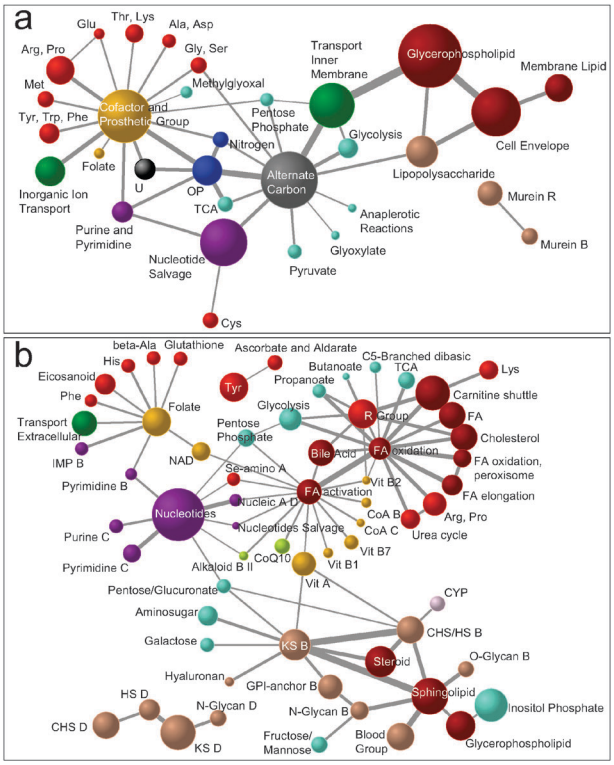


**Fig. 5** Metabolic backbones displaying pathway's cross-talks inferred from the model. (a) Metabolic backbone for *E. coli* metabolism at the 0.064 confidence level with 30% of the original total weight, 91% of the original number of pathways, and 9% of the original links. (b) Metabolic backbone for human cells at the 0.022 confidence level, with 20% of the original total weight, 69% of the original number of pathways, and 5% of the original links. Different colors indicate different metabolic families. The area of a circle representing a pathway is proportional to its size in number of reactions. The weights in the connections are proportional to the intensity of the cross-talk between the pathways. Pathway names have been abbreviated in standard forms when necessary, see ESI.†

## Discussion

From a broad perspective, a cartographic representation of complex networks[15] supposes to map the positions of nodes in an underlying geometric space and shares some fundamental problems with traditional geographical cartography on what concerns techniques, generalizations or design: how to represent the topology of the mapped network on the metric space, which characteristics of the network are not relevant to the map's purpose and can be eliminated, how to reduce the complexity of the characteristics that will be mapped, *etc*. Despite the difficulties, cartographic maps based on geometrical spaces are crucial to identify dominant nodes, to understand how different subparts of the system, like pathways in our context, relate to each other, to back up more accurate methods of prediction of missing and spurious interactions,[28,29] or to find optimal transport routes.

In our metabolic maps, the astonishing congruency between the architecture of metabolic networks and the underlying geometry is supported by a biochemical interaction law that, irrespective of the studied organisms, of the nature and complexity of the reactions they account for, or of the different structural label of the metabolites they involve, seems to comply with a simple Newtonian-like form and allows us to make predictions about the probabilities of interaction among sets of metabolites forming reactions. Specifically, the sum of the probabilities running over all the metabolites participating in a certain biochemical reaction can be interpreted as a topological version of the well-known concept of reaction-based affinity, and each summand could thus be

 

identified with the chemical potential of that particular metabolite in relation to its chemical partners in the particular reaction. Our results point to a systems level definition of chemical affinity in terms of network-based probabilities of interaction which depend on the distances in the underlying geometric space and on intrinsic properties of nodes which convert some of them in hubs.

Such probabilistic network-based chemical affinities allow us to recover the established biochemical organization of pathways as connected metabolic families, but at the same time raise new questions claiming for the need of rethinking its classical definition as self-contained units. We find that different pathways may have disparate internal structures, some of them being more modular and conforming better to the classical definition, while subunits pointing to differentiated functionalities can be distinguished in others. We have also unveiled a higher level of systems' level interactions represented by metabolic backbones, defined on the basis of a quantitative cross-talk between pathways. This particular idea advises us against the use of very specific biochemical protocols aimed to single-out particular pathways as they might be prone to underestimate the delicate connections that underlay the net and secure its proper functioning. Such metabolic features are common to human cells and *E. coli*. However, a comparative study shows that pathways in human metabolism are in general more modular and display less overlap of common metabolites with other pathways. At the same time the different human metabolic families are more entangled and sectors are more difficult to characterize, a possible signature of a higher functional complexity or merely a side effect of the kind of reconstruction that mixes in a single network reactions that occur in diversely differentiated cells.

Summarizing, in this work we provide cartographic maps of two representative metabolisms that capture their specific complexities, explaining many of their system properties, and provide a new perspective on the definition, cross-talk, and hierarchical organization of biochemical pathways. These maps, embedded in a simple geometric space, rely on a probabilistic biochemical connectivity law which emerges from the different physico-chemical forces acting at a molecular level and that naturally conveys a higher interaction likelihood to elements which are closer in the underlying space. Similar maps for other biological networks are expected to be equally congruent and to help transforming data into knowledge and knowledge into understanding, paving the way for new discoveries in systems biology prediction and control.

### Funding

This work was supported by MICINN Projects No. FIS2010-21781-C02-02, FIS2006-03525, and BFU2010-21847-C02-02; Generalitat de Catalunya grants No. 2009SGR838 and 2009SGR1055; the Ramón y Cajal program of the Spanish Ministry of Science; and the ICREA Academia prize 2010 funded by the Generalitat de Catalunya.

## Methods

### Hidden metric spaces and the $\mathbb{S}^1 \times \mathbb{S}^1$ model

The $\mathbb{S}^1 \times \mathbb{S}^1$ model can be used as a network generator as follows:

(1) $N_m$ metabolites and $N_r$ reactions are homogeneously distributed in a circle of radius $R$. The densities of metabolites and reactions in the circle are $\delta_m = N_m/2\pi R$ and $\delta_r = N_r/2\pi R$, taken independent of the network size. Without loss of generality, one of them can be set to 1.

(2) Metabolites and reactions are assigned expected degrees $k_m$ and $k_r$, drawn from the probability densities $\rho_m(k_m)$ and $\rho_r(k_r)$, respectively. To model metabolic networks, we use $\rho_m(k_m) \approx k_m^{-\gamma}$ and $\rho_r(k_r) = \delta(k_r - \langle k_r \rangle)$. Our choice of a power law degree distribution of metabolites is motivated by previous studies[17] and by a direct measurement of this distribution in our database, as shown in Fig. S1 (ESI†).

(3) Each possible pair metabolite/reaction—with degrees $k_m$ and $k_r$ and located at angular positions $\theta_m$ and $\theta_r$—is visited once and a link is created with probability

$$p(k_m, \theta_m; k_r, \theta_r) = p\left(\frac{d_{mr}}{\mu k_m k_r}\right), \qquad (3)$$

where $d_{mr} = R\Delta\theta_{mr}$ ($\Delta\theta_{mr}$ is the angular separation) is the distance metabolite/reaction in the circle. Function $p$ can be, *a priori*, any integrable function. However, the choice $p(x) = (1 + x^{\beta})^{-1}$ generates maximally random networks given the constraints of the model.

See ESI† for extended details on the $\mathbb{S}^1 \times \mathbb{S}^1$ model.

### Inverse problem

Given a complex network representation, the inverse problem of embedding the network in the hidden metric space amounts to find the optimal position of every node in that underlying geometry. The optimal coordinates would ensure that, given the specific form of the connection probability in eqn (2), the model has a maximum probability to reproduce the observed topology. In general terms, the embedding is resolved using statistical inference techniques, basically a maximum likelihood estimation in combination with a Monte Carlo method and a Metropolis–Hasting rule to explore and select possible configurations in the underlying space.[30] More precisely, the likelihood functional is defined as

$$L \equiv \prod_{m=1}^{N_m} \prod_{r=1}^{N_r} \left[p\left(\frac{d_{mr}}{\mu k_m k_r}\right)\right]^{a_{mr}} \left[1 - p\left(\frac{d_{mr}}{\mu k_m k_r}\right)\right]^{1-a_{mr}} \qquad (4)$$

where $a_{mr}$ is the bipartite adjacency matrix of the network, defined as $a_{mr} = 1$ if metabolite $m$ participates in reaction r and zero otherwise. The bipartite nature of metabolic networks together with the fact that reactions and metabolites have disparate degree distributions precludes to perform the mapping in a single-step. Rather the embedding into the $\mathbb{S}^1 \times \mathbb{S}^1$ space runs in two phases: first the one-mode projection of the metabolic subnetwork is embedded into a $\mathbb{S}^1$ space following the numerical optimization procedures described in ref. 21, and second the inferred angular coordinates of metabolites are used as an input to adjust the position of each individual reaction in the circle. See ESI† for a more complete description of the $\mathbb{S}^1 \times \mathbb{S}^1$ embedding algorithm.

### The disparity filter

To extract the metabolic backbone of cross-talks between pathways we apply the multi-scale disparity filter defined in ref. 27.

The disparity filter exploits local heterogeneity and correlations among weights in complex weighted network representations to extract the network backbone by considering the relevant edges at all the scales present in the system. It ensures that small nodes in terms of strength ($s_i = \Sigma_{j \equiv i\text{–neighbors}} w_{ji}$, sum of incident weights to node $i$) are not neglected and that the backbone remains connected and does not disaggregate into separate clusters. The methodology preserves interactions with a statistically significant intensity for at least one of the two nodes the edge is incident to. To decide whether a connection is relevant, the filter compares against a null hypothesis which assumes that the local weights associated to a node are uniformly distributed at random. In this way one discounts intensities that could be explained by random fluctuations. The disparity filter produces better results in terms of preserving the maximum number of nodes and weights in the backbone with the minimum number of links as compared to a global threshold filter that selects all the links with weights above a certain value, see ESI†, Fig. S3.

### Average angular position and concentration of pathways

To find the average angular position of a given pathway and a measure of its angular concentration (or dispersion), we use the following method. Each reaction $i$ of a given pathway (with $i = 1,\ldots,N_p$ reactions in it) is assigned a normalized vector $\vec{r}_i$ pointing to the position of the reaction in a circle with radius 1 using as an angular coordinate the one inferred by our method. The average angular position of the pathway is then defined as the angular coordinate of the average vector $\langle \vec{r} \rangle \equiv \sum_{i=1}^{N_p} \vec{r}_i / N_p$. We use this method to plot the names of the different pathways in Fig. 2. The length of the average vector $|\langle \vec{r} \rangle|$ is a measure of the angular concentration of the reactions. A value $|\langle \vec{r} \rangle| = 1$ means that all reactions in the pathway have the same angular coordinates whereas $|\langle \vec{r} \rangle| = 0$ indicates a perfect homogeneous distribution over the circle.

## References


1 B. Palsson and K. Zengler, *Nat. Chem. Biol.*, 2010, **6**, 787789.
2 B. O. Palsson, *Systems Biology: Properties of Reconstructed Networks*, Cambridge University Press, Cambridge, 2006.
3 S. N. Dorogovtsev and J. F. F. Mendes, *Adv. Phys.*, 2002, **51**, 1079–1187.
4 R. Albert and A.-L. Barabási, *Rev. Mod. Phys.*, 2002, **74**, 47–97.
5 M. E. J. Newman, *SIAM Rev.*, 2003, **45**, 167–256.
6 S. N. Dorogovtsev and J. F. F. Mendes, *Evolution of networks: From biological nets to the Internet and WWW*, Oxford University Press, Oxford, 2003.
7 M. E. J. Newman, *Networks: An introduction*, Oxford University Press, 2010.
8 J. S. Edwards, R. U. Ibarra and B. O. Palsson, *Nat. Biotechnol.*, 2001, **19**, 125–130.
9 E. Almaas, B. Kovács, T. Vicsek, Z. N. Oltvai and A.-L. Barabási, *Nature*, 2004, **427**, 839–843.
10 J. Stelling, S. Klamt, K. Bettenbrock, S. Schuster and E. D. Gilles, *Nature*, 2002, **420**, 190–193.
11 A. G. Smart, L. A. N. Amaral and J. Ottino, *Proc. Natl. Acad. Sci. U. S. A.*, 2008, **105**, 13223–13228.
12 M. A. Serrano and F. Sagués, *BMC Syst. Biol.*, 2011, **5**, 76.
13 D. J. Watts and S. H. Strogatz, *Nature*, 1998, **393**, 440–442.
14 S. Fortunato, *Phys. Rep.*, 2010, **486**, 75–174.
15 R. Guimerà and L. A. N. Amaral, *Nature*, 2005, **433**, 895–900.
16 H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai and A.-L. Barabási, *Nature*, 2000, **407**, 651–654.
17 E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai and A.-L. Barabási, *Science*, 2002, **297**, 1551–1555.
18 M. A. Serrano, D. Krioukov and M. Boguñá, *Phys. Rev. Lett.*, 2008, **100**, 078701.
19 D. Garlaschelli and M. I. Loffredo, *Phys. Rev. E*, 2008, **78**, 015101.
20 D. Krioukov, F. Papadopoulos, M. Kitsak, A. Vahdat and M. Boguñá, *Phys. Rev. E*, 2010, **82**, 036106.
21 M. Boguñá, F. Papadopoulos and D. Krioukov, *Nat. Commun.*, 2010, **1**, 62.
22 A. M. Feist, C. S. Henry, J. L. Reed, M. Krummenacker, A. R. Joyce, P. D. Karp, L. J. Broadbelt, V. Hatzimanikatis and B. O. Palsson, *Mol. Syst. Biol.*, 2007, **3**, 121.
23 N. C. Duarte, S. A. Becker, N. Jamshidi, I. Thiele, M. L. Mo, T. D. Vo, R. Srivas and B. O. Palsson, *Proc. Natl. Acad. Sci. U. S. A.*, 2007, **104**, 1777–1782.
24 J. Schellenberger, J. O. Park, T. C. Conrad and B. O. Palsson, *BMC Bioinf.*, 2010, **11**, 213.
25 BiGG database, http://bigg.ucsd.edu/.
26 T. Fawcett, *Pattern Recogn. Lett.*, 2006, **27**, 861–874.
27 M. A. Serrano, M. Boguñá and A. Vespignani, *Proc. Natl. Acad. Sci. U. S. A.*, 2009, **106**, 6483–6488.
28 A. Clauset, C. Moore and M. Newman, *Nature*, 2008, **453**, 98–101.
29 R. Guimerà and M. Sales-Pardo, *Proc. Natl. Acad. Sci. U. S. A.*, 2009, **106**, 22073–22078.
30 M. E. J. Newman and G. T. Barkema, *Monte Carlo Methods in Statistical Physics*, Oxford University Press, Oxford, UK, 1999.


 