A Toolbox Model of Evolution of Metabolic Pathways on Networks of Arbitrary Topology

Tin Yau Pang[1,2], Sergei Maslov[1]*

1 Department of Condensed Matter Physics and Materials Science, Brookhaven National Laboratory, Upton, New York, United States of America, **2** Department of Physics and Astronomy, Stony Brook University, Stony Brook, New York, United States of America

Abstract

In prokaryotic genomes the number of transcriptional regulators is known to be proportional to the square of the total number of protein-coding genes. A toolbox model of evolution was recently proposed to explain this empirical scaling for metabolic enzymes and their regulators. According to its rules, the metabolic network of an organism evolves by horizontal transfer of pathways from other species. These pathways are part of a larger ''universal'' network formed by the union of all species-specific networks. It remained to be understood, however, how the topological properties of this universal network influence the scaling law of functional content of genomes in the toolbox model. Here we answer this question by first analyzing the scaling properties of the toolbox model on arbitrary tree-like universal networks. We prove that critical branching topology, in which the average number of upstream neighbors of a node is equal to one, is both necessary and sufficient for quadratic scaling. We further generalize the rules of the model to incorporate reactions with multiple substrates/products as well as branched and cyclic metabolic pathways. To achieve its metabolic tasks, the new model employs evolutionary optimized pathways with minimal number of reactions. Numerical simulations of this realistic model on the universal network of all reactions in the KEGG database produced approximately quadratic scaling between the number of regulated pathways and the size of the metabolic network. To quantify the geometrical structure of individual pathways, we investigated the relationship between their number of reactions, byproducts, intermediate, and feedback metabolites. Our results validate and explain the ubiquitous appearance of the quadratic scaling for a broad spectrum of topologies of underlying universal metabolic networks. They also demonstrate why, in spite of ''small-world'' topology, real-life metabolic networks are characterized by a broad distribution of pathway lengths and sizes of metabolic regulons in regulatory networks.

Citation: Pang TY, Maslov S (2011) A Toolbox Model of Evolution of Metabolic Pathways on Networks of Arbitrary Topology. PLoS Comput Biol 7(5): e1001137. doi:10.1371/journal.pcbi.1001137

Editor: Eugene I. Shakhnovich, Harvard University, United States of America

Received September 20, 2010; **Accepted** April 14, 2011; **Published** May 19, 2011

Funding: Work at Brookhaven National Laboratory was carried out under Contract No. DE-AC02-98CH10886, Division of Material Science, US Department of Energy. Part of this study was supported by the program ''Computational Biology and Bioinformatic Methods to Enable a Systems Biology Knowledgebase'' (DE-FOA-0000143) of the Office of Biological and Environmental Research, US Department of Energy. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: maslov@bnl.gov

Introduction

In prokaryotic genomes the number of transcriptional regulators is known to quadratically scale with the total number of protein-coding genes [1]. The toolbox model of co-evolution of metabolic and regulatory networks was recently proposed [2] to explain this scaling in parts of the genome responsible for metabolic functions. In this model prokaryotes acquire new metabolic capabilities by horizontal transfer of entire metabolic pathways from other organisms. One can conveniently think of these new pathways as coming from some ''universal network'' formed by the union of metabolic repertoires of all potential donor organisms. The essence of the toolbox argument [2] can be summarized as follows: as the non-regulatory part of the genome of an organism (its ''toolbox'') grows, it typically needs to acquire fewer and fewer extra new genes (''tools'') in a pathway offering it some new metabolic capability (e.g. the ability to utilize a new nutrient or synthesize a new metabolic product). As a consequence, the number of pathways and by extension the number of their transcriptional regulators grows faster than linearly with the number of non-regulatory genes in the genome. While this qualitative explanation is rather general and therefore does not depend on specific details such as topology of the universal network, the exact value of the exponent α connecting the number of transcription factors (equal to N_L- the number of pathways or leaves of the network) to the number of metabolites in the metabolic network of an organism N_M, as $N_L \sim N_M^{\alpha}$, is in general model-dependent. In [2] we mathematically derived the quadratic scaling ($\alpha = 2$) for the toolbox model with linear pathways on a fully connected graph in which any pair of metabolites can in principle be converted to each other in just one step via a single metabolic reaction. While this situation is obviously unrealistic from biological standpoint, before present study it remained the only mathematically treatable variant of the toolbox model. The universality of the exponent $\alpha = 2$ was then corroborated [2] by numerical simulations of the toolbox model with linearized pathways on the universal network formed by the union of all metabolic reactions in the KEGG database. While the agreement between the values of the exponent α in these two cases hinted at underlying general principles at work, the detailed understanding of these principles remained elusive.

Author Summary

It has been previously reported that in prokaryotic genomes the number of transcriptional regulators is proportional to the square of the total number of genes. We recently offered a general explanation of this empirical powerlaw scaling in terms of the "toolbox" model in which metabolic and regulatory networks co-evolve together. This evolution is driven by horizontal gene transfer of co-regulated metabolic pathways from other species. These pathways are part of a larger "universal" network formed by the union of all species-specific networks. In the present work we address the question of how topological properties of this universal network influence the power-law scaling of regulators in the toolbox model. We also generalize its rules to include reactions with multiple substrates and products, branched and cyclic metabolic pathways, and to account for optimality of metabolic pathways. The main conclusion of our analytical and numerical modeling efforts is that the quadratic scaling is the robust feature of the toolbox model in a broad range of universal network topologies. They also demonstrate why, in spite of "small-world" topology, real-life metabolic networks are characterized by a broad distribution of pathway lengths and sizes of regulons in regulatory networks.

The question we address in this study is how the topology of the universal network determines this scaling exponent. To answer this question we first consider and solve a more realistic (yet still mathematically treatable) case in which the universal metabolic network is a directed tree of arbitrary topology. While being closer to reality than previously solved [2] case of fully connected network, the toolbox model on a tree-like universal network still retains a number of simplifications such as strictly linear pathways and one substrate $\rightarrow$ one product reactions.

To make our approach even more realistic we propose and numerically study a completely new version of the toolbox model incorporating metabolic reactions with multiple substrates and products as well as branched and cyclic metabolic pathways. Furthermore, unlike random linear pathways on a universal network [2] that can be long and therefore suboptimal from an evolutionary standpoint, the new model uses evolutionarily optimized pathways with the smallest number of reactions from the KEGG database sufficient to achieve a given metabolic task.

Results

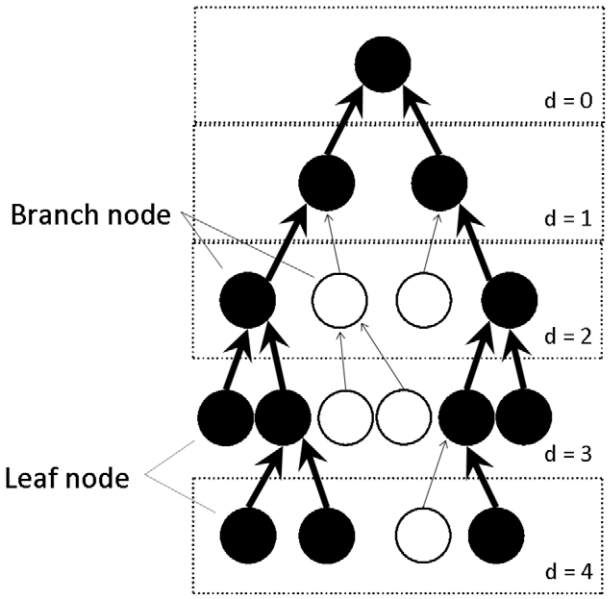
The toolbox model on a tree-like universal network: General mathematical description

We will first consider the case where the universal metabolic network is a directed tree. For simplicity in this section we will consider the case of catabolic pathways, while identical arguments (albeit with opposite direction of all reactions) apply to anabolic pathways. The root of the tree corresponds to the central metabolic core of the organism responsible for biomass production. Peripheral catabolic pathways (branches of the tree) convert external nutrients (leaves) to this core, while the internal nodes of the tree represent intermediate metabolites. Each of metabolites is characterized by its distance $0 \le d \le d_{max}$ from the root of the network. The universal network has $N_M^{(U)}(d)$ metabolites at distance d from the root that included $N_L^{(U)}(d)$ leaves (nutrients used in the first step of catabolic pathways) and $N_B^{(U)}(d)$ branching points corresponding to intermediate metabolites generated by more than one metabolic reaction at the next level (see Figure 1). An organism-specific network (filled circles and thick edges in Figure 1) at distance d from the root contains $N_M(d) \le N_M^{(U)}(d)$ metabolites composed of $N_L(d) \le N_L^{(U)}(d)$ leaves, $N_B(d) \le N_B^{(U)}(d)$ branching points, and $N_M(d) - N_L(d) - N_B(d)$ metabolites inside linear branches ("one reaction in-one reaction out") . For simplicity we assume that in the universal network (and thus also in any of its organism-specific subnetworks) no more than two reaction edges can combine at any node (metabolite), while the most general case of an arbitrary distribution of branching numbers can be treated in a very similar fashion.

The toolbox model specifies rules by which organism acquires new pathways in the course of its evolution. It consists of the following steps: 1) randomly pick a new nutrient metabolite (a leaf node of the universal network that currently does not belong to the metabolic network of the organism) 2) use the universal network to identify the unique linear pathway which connects the new nutrient to the root of the tree (the metabolic core) and finally 3) add the reactions and intermediate metabolites in the new pathway to the metabolic network of the organism (filled circles and thick edges in Figure 1). One needs to only add those enzymes that are not yet present in the "genome" of the organism. Graphically it means that the new branch of the universal network is extended until it first intersects the existing metabolic network of the organism.

Consider an organism capable of utilizing $N_L \le N_L^{(U)}$ nutrients represented by leaves in the universal network, where $N_L = \sum_{d=1}^{d_{\max}} N_L(d)$ and $N_L^{(U)} = \sum_{d=1}^{d_{\max}} N_L^{(U)}(d)$. Since we assume that each nutrient utilization pathway is controlled by a dedicated transcriptional regulator sensing its presence or absence in the environment (e.g. LacR for lactose), the corresponding regulatory network would also have N_L transcription factors (in the model we ignore transcription factors controlling non-metabolic functions). The non-regulatory part of the genome consists of $N_M = \sum_{d=1}^{d_{\max}} N_M(d)$ enzymes catalyzing metabolic reactions (strictly speaking N_M is the number of metabolites/nodes so that the number of enzymes/edges is $N_M - 1$). Quadratic scaling plots [1] shows the number of transcriptional regulators $N_R = N_L$ vs. the total number of genes in the genome (both regulatory and non-regulatory) $N_G = N_M - 1 + N_L$. However, since in all organism-specific networks $N_M \gg N_L$, the quadratic scaling between N_R and N_G is equivalent to $N_L \sim N_M^2$.

We further assume that due to random selection N_L nutrients are expected to be uniformly distributed among all d levels. Therefore, the expected number of leaves at a given level is given by $N_L(d) = \tau N_L^{(U)}(d)$ where the fraction $\tau = N_L \Big/ N_L^{(U)}$ is the same at all levels. On the other hand the fraction $\mu(d) = N_M(d) \Big/ N_M^{(U)}(d)$ varies from level to level. It usually tends to increase as one gets closer towards the root of the tree reaching 1 for $d = 0$ (the root node itself). To derive the equation for $\mu(d)$, one first notices that each of $N_M(d+1)$ metabolites at level $d+1$ is converted to another intermediate metabolite at level d. Due to merging of pathways at $N_B(d)$ branching points the number of unique intermediate metabolites at the level d is actually smaller: $N_M(d+1) - N_B(d)$. To calculate $N_B(d) \le N_B^{(U)}(d)$ one uses the fact that each of the two nodes downstream of a branching point in the universal network is present in the organism-specific network with probability $N_M(d+1) \Big/ N_M^{(U)}(d+1)$. The probability that they are both present is $\left(N_M(d+1) \Big/ N_M^{(U)}(d+1) \right)^2$ and thus the number of branching points at level d of the organism-specific metabolic network is $N_B(d) = \left(\dfrac{N_M(d+1)}{N_M^{(U)}(d+1)} \right)^2 N_B^{(U)}(d)$. The intermediate metabolites together with new nutrients



Universal		Organismal	
Expression	Value	Expression	Value
$N_M^{(U)}(3)$	6	$N_M(3)$	4
$N_B^{(U)}(3)$	2	$N_B(3)$	1
$N_L^{(U)}(3)$	4	$N_L(3)$	2

Figure 1. An example of organism-specific metabolic network and the corresponding universal network. The organism-specific metabolic network (filled circles and thick edges) is always a subset of the universal network (the entire tree). Nodes are divided into layers based on their distance d from the root of the tree. Variables $N_M^{(U)}(d)$, $N_B^{(U)}(d)$, $N_L^{(U)}(d)$ for the universal network and $N_M(d)$, $N_B(d)$, $N_L(d)$ for species-specific network are illustrated using the layer $d=3$ as an example.
doi:10.1371/journal.pcbi.1001137.g001

$N_L(d)=\tau N_L^{(U)}(d)$ entering at the level d add up to the total number of metabolites at level d:

$$N_M(d)=N_M(d+1)-\left(\frac{N_M(d+1)}{N_M^{(U)}(d+1)}\right)^2 N_B^{(U)}(d)+\tau N_L^{(U)}(d) \qquad (1)$$

This equation allows one to iteratively calculate $N_M(d)$ for all d starting from $N_M(d_{\max})=\tau N_L^{(U)}(d_{\max})$. We will use this equation to derive the relationship between the number of leaves and the total number of nodes first for a critical branching tree and then for a supercritical one.

The toolbox model on a critical tree

The Galton-Watson branching process [3] is a simple stochastic process generating random trees, and we will consider its version where a node can have two, one, or zero neighbors (parents) at the previous level with probabilities p_2, p_1 and p_0 correspondingly. If the average number of parents k equals one, then the process is referred to as critical, and if k is greater than one then the process is supercritical. More generally critical and supercritical branching trees can be generated by a variety of random processes such as e.g. directed percolation [4]. While for simplicity we used the Galton-Watson branching process in our derivation below, it can be readily extended to this more general case.

The principal geometric difference between supercritical and critical trees is that in the former case the number of nodes in a layer $N_M^{(U)}(d)\sim k^d$ exponentially grows with d [3], while in a critical tree it grows at most algebraically (for the Galton-Watson critical process $N_M^{(U)}(d)\sim d$ [3]). The other difference is that while the critical branching process always stops on its own at a certain finite height $d_{\max}$, a supercritical process will go on forever so that to generate a tree one has to manually terminate it at a predefined layer $d_{\max}$. The most significant feature of a critical tree is that it has much longer branches than a supercritical one of the same size. Indeed, the diameter (the maximal height) of a random critical tree with $N_M^{(U)}$ nodes is $d_{\max}\sim\sqrt{N_M^{(U)}}$ while in a supercritical tree it is much shorter: $d_{\max}\sim\log N_M^{(U)}/\log k$. Thus supercritical trees (unlike their critical counterparts) have the small world property.

A random critical network where each node has at most has two parents in the previous layer is defined by $p_0=p_2=p\le 0.5$. Indeed, in this case $k=0\cdot p_0+1\cdot p_1+2\cdot p_2=1$. In such network $N_B^{(U)}(d)=N_L^{(U)}(d)=pN_M^{(U)}(d)$ and hence the Eq. (1) can be rewritten as

$$\frac{1}{p}\left[\mu(d)-\frac{N_M^{(U)}(d+1)}{N_M^{(U)}(d)}\mu(d+1)\right]=\tau-[\mu(d+1)]^2 \qquad (2)$$

A critical branching process that has not terminated by level d satisfies $N_M^{(U)}(d)\sim d$ or $N_M^{(U)}(d+1)/N_M^{(U)}(d)=1+1/d$. More generally if $N_M^{(U)}(d)$ algebraically increases with d, $N_M^{(U)}(d+1)/N_M^{(U)}(d)$ asymptotically approaches 1 as

$$\frac{N_M^{(U)}(d+1)}{N_M^{(U)}(d)}=1+\frac{\text{const}}{d} \qquad (3)$$

Here $\text{const}/d\to 0$ as $d\to\infty$, thus for $1\ll d\ll d_{\max}$ $\mu(d)$ remains approximately constant and according to Eq. (2) this constant ratio μ is defined by

$$\tau=\mu^2 \qquad (4)$$

This quadratic relation is exact in a critical branching tree where each node can branch out into at most two nodes at the next layer, and it is still correct to a leading order in $\mu\ll 1$ for a critical branching tree with arbitrary branching ratios (see "Quadratic

relation between μ and τ for general critical branching processes" of Text S1). Furthermore, one can show (see "Calculation of the average μ in the toolbox model on a critical tree" of Text S1) that in large critical networks the overall fraction of metabolites present in organism-specific metabolic network is very close to this stationary limit of $\mu(d)$: $N_M \Big/ N_M^{(U)} \approx \mu$.

As was explained in the previous section the ratio $N_G \Big/ N_G^{(U)}$ between the total number N_G of metabolic-related genes in the genome of an organism and its theoretical maximal value $N_G^{(U)}$ for a genome containing the entire universal network is also given by μ. Furthermore, in our model the number of leaves is equal to the number of nutrient-utilizing pathways or, alternatively, their transcriptional regulators $N_R = N_L = \tau N_L^{(U)}$. Thus like in a much simpler model of Ref. [2] the toolbox model on any critical tree-like universal network gives rise to quadratic scaling of the number of transcription factors with the total number of genes:

$$N_R \Big/ N_R^{(U)} = \left(N_G \Big/ N_G^{(U)} \right)^2 \tag{5}$$

The geometrical properties of the universal network such as its total number of nodes/edges $N_M^{(U)} \approx N_G^{(U)}$ and number of leaves/branches $N_L^{(U)} \approx N_R^{(U)}$ determine the prefactor of this scaling law. Simulation of the toolbox model on the critical tree (Figure 2) verified our mathematical predictions with the best fit to binned datapoints in Figure 2 giving the exponent $\alpha = 1.9 \pm 0.1$.

The toolbox model on a supercritical tree

For a supercritical branching process $\dfrac{N_M^{(U)}(d+1)}{N_M^{(U)}(d)} = k > 1$ and according to Eq. (1) (See SI for the derivation) the steady state value μ_* of $\mu(d)$ satisfies

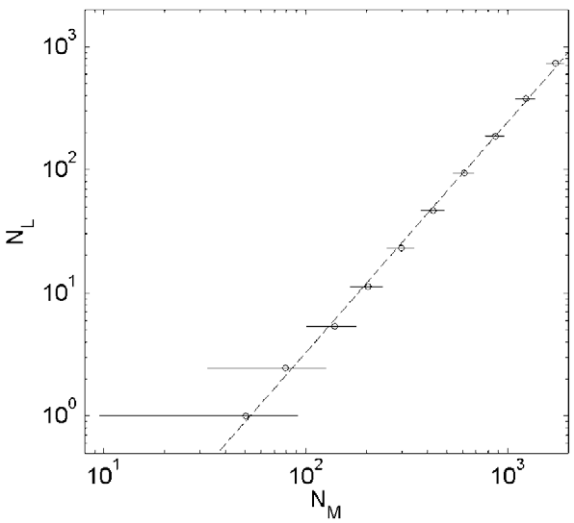


Figure 2. N_L vs. N_M. N_L is the number of leaves in an organism-specific metabolic network and equal to the number of transcriptional regulators of corresponding nutrient-utilizing pathways, while N_M is the total number of nodes/metabolites in this netowrk. The data are generated by the toolbox model on critical universal network with sizes around 2000. Solid line $N_L = N_M^{\alpha}/A$, where the exponent $\alpha = 1.9 \pm 0.1$ and the prefactor $A = 1600 \pm 400$, are the best fits to the binned data.
doi:10.1371/journal.pcbi.1001137.g002

$$\tau = -\left(\frac{k-1}{p}\right)\mu_* + \left(\frac{k-1}{p} + 1\right)\mu_*{}^2 \tag{6}$$

Here $p = p_0$ and $k = 1 - p_0 + p_2 > 1$. Notice that for $\tau = 0$ one has two solutions for μ_*: 0 and $\mu_0 = (k-1)/(k-1+p)$. This indicates transition in which for τ exactly at zero one has $\mu(d) = 0$, while for an arbitrary small yet positive τ the value of $\mu(d)$ asymptotically converges to $\mu_0 > 0$ for $d \ll d_{max}$. This transition resembles the first order phase transition, e.g., liquid-gas transition, where right at the transition point very small variation of the external parameter such as temperature (analogous to τ in this model) results in a large jump of the order parameter such as density (analogous to our $\mu(d)$). (See [5] for details), The number of layers over which this conversion is taking place is itself a function of τ and for small τ it is large. For exponentially growing supercritical networks and for small $\tau \ll 1$, the network average value of $\mu(d)$ defined as $\mu = N_M \Big/ N_M^{(U)}$ satisfies

$$\mu = \frac{\tau}{\mu_0}\frac{k-1}{k}\log_k\left(\frac{\mu_0}{\tau}\right) \tag{7}$$

Note that this equation connecting μ and τ (see SI for detailed derivation) is markedly different from Eq. (6) for steady state value μ_* in middle layers.

In conclusion, while the toolbox model on a critical universal network is characterized by a quadratic scaling between τ and μ (see Eq. (4)), the same model on a supercritical, exponentially expanding universal network gives rise to a linear scaling of τ vs. μ albeit with logarithmic corrections (see Eq. (7)). This difference in exponent equally applies to the scaling of the number of regulators N_R vs. the total number of genes N_G in the toolbox model on critical and supercritical universal network.

Simulation of the toolbox model on the KEGG network with linearized pathways

To test our mathematical results for a more realistic version of the universal tree we linearized pathways and reactions in the network formed by the union of all reactions in the KEGG database [6]. To this end we generated a random spanning tree on the KEGG network by the following algorithm: the metabolite pyruvate was selected as the root of the tree. We then randomly picked a metabolite located upstream of it and generated a linear pathway (tree branch) as a self-avoiding random walk on the KEGG network extended until it either merges with another pathway or reaches the root of the tree. This step was repeated until all upstream metabolites were covered. The resulting spanning tree was then used as the universal network on which we simulated the toolbox model by gradually increasing the number of pathways N_L and recording the total number of metabolites N_M in organism-specific metabolic networks. Our numerical simulations generated approximately quadratic scaling $\alpha = 1.8 \pm 0.1$ (see Ref. [2]).

To better understand the origins of this scaling we investigated the topology of the underlying universal tree. The criticality of a tree is defined by the asymptotic value of the ratio $N_M^{(U)}(d+1) \Big/ N_M^{(U)}(d)$ for large d: for supercritical trees it reaches a limit $k > 1$, while for critical ones it converges to 1 as described in Eq. (3). Figure 3 showing $N_M^{(U)}(d+1) \Big/ N_M^{(U)}(d)$ vs. d in the linearized KEGG network convincingly demonstrates its criticality. Thus the quadratic scaling between the number of

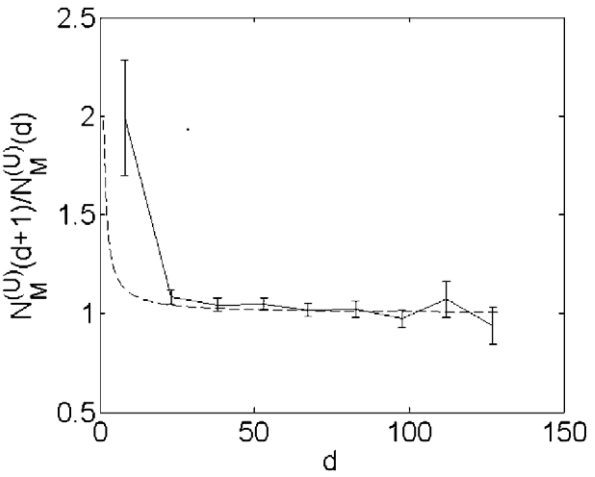


Figure 3. $N_M^{(U)}(d+1)\Big/N_M^{(U)}(d)$ vs. d for KEGG-based universal network with linearized pathways. $N_M^{(U)}(d+1)\Big/N_M^{(U)}(d)$ (the ratio of the number of metabolites at two consecutive layers) plotted as a function of d (the layer number) for KEGG-based universal network with linearized pathways. Solid line: measurement, dotted line: its expected profile, $1+1/d$, in a critical branching tree. The error bars reflect standard deviation in different spanning trees used to linearize the KEGG network.
doi:10.1371/journal.pcbi.1001137.g003

transcriptional regulators and the number of metabolites in the toolbox model simulated on the linearized KEGG network is explained by the mathematical formalism described in previous sections.

In addition to using a random spanning tree to linearize the KEGG network we also tried a version using minimal paths. In this version the universal network is generated by randomly picking a metabolite and connecting it to the root of the tree (pyruvate) by the shortest path. At a first glance such "minimal path" selection appears to be reasonable from evolutionary standpoint. Indeed, evolution would favor simpler and shorter pathways in order to minimize the expenditure of resources to achieve a given metabolic goal [7] . However, the minimal paths version of linearization of the KEGG resulted in a supercritical universal network with logarithmically short branches $d \sim \log N_M^{(U)}$. As predicted for supercritical trees (Eq. (7)) the toolbox model in this case had an approximately linear scaling of the number of transcriptional regulators (leaves of branches on the network) with the total number of metabolites: the measured best fit exponent was only 1.2 ± 0.1.

How do we reconcile the evolutionary pressure apparently selecting for minimal pathways with dramatically wrong scaling properties of this model? We believe that most of the ultra-short "small world" pathways generated by minimal paths on the KEGG network are unrealistic from biochemical standpoint. Indeed, highly connected co-factors often position metabolites with very different chemical formulas in close proximity to each other. For example, the KEGG reaction R00134: $\mathrm{Formate} + \mathrm{NADP}^+ \leftrightarrow \mathrm{CO_2} + \mathrm{NADPH}$ would appear as a miraculous "one-step" conversion of carbon dioxide into formate, while the reaction R03546: $\mathrm{CO_2} + \mathrm{Carbamate} \leftrightarrow \mathrm{Cyanate} + \mathrm{H}^+ + \mathrm{HCO_3^-}$ would artificially link carbon dioxide and cyanate. The combination of these two reactions gives rise to equally impossible two-step path: formate $\rightarrow$ CO_2 $\rightarrow$ cyanate. As a consequence of such artificial shortcuts branches of the universal network linearized by minimal paths are much shorter than they are in reality. .This problem is at least partially alleviated by 1) removing unusually high-degree nodes corresponding to common co-factors such as H_2O, ATP, NAD in the metabolic network so that some unrealistic paths are eliminated, and also 2) using random spanning tree instead of the shortest paths. In Ref. [2] we followed both of these recipes to successfully reproduce the quadratic scaling in real-life data. Still no linearization procedure could completely avoid biochemically meaningless shortcuts. In the next section we introduce and study a new considerably more realistic version of the toolbox model operating on branched and interconnected universal networks. Pathways in this version of the toolbox model satisfy the evolutionary requirements for minimal size. Proper treatment of metabolic reactions with multiple substrates prevents biochemically meaningless shortcuts and as a consequence restores the quadratic scaling.

The toolbox model on KEGG network with branched pathways and multi-substrate reactions

Real metabolic reactions routinely include multiple inputs (substrates) and multiple outputs (products) (see Table 1 and Table 2 for statistics in the KEGG database). Furthermore, metabolic networks often have two or more alternative pathways generating the same set of end-products from the same set of nutrients. Both these factors result in metabolic networks that are branched and interconnected. Here we propose and simulate a more realistic version of the toolbox model. The most prominent feature of the new model of pathways is the "AND" function acting on inputs of multi-substrate reactions. It reflects the constraint that a reaction cannot be carried out unless all its substrates are present.

The new version of the toolbox model simulates addition of anabolic pathways aimed at production of new metabolites from those the model organism can currently synthesize (its current metabolic core). The new pathways are *optimal* in the sense that they contain the smallest number of reactions necessary to synthesize the desired end-product. As for previous versions of the toolbox model, one can modify the rules of this model to apply to catabolic pathways but for simplicity we will limit the following discussion to anabolic pathways. The rules of the new model are:

1. At the beginning of the simulation, the model organism starts with a "seed" metabolic network consisting of 40 metabolites classified by the KEGG as parts of central carbohydrate metabolism, plus a number of "currency" metabolites such as water, ATP and NAD (see the section "Seed metabolites of the scope expansion" of Text S1 for additional details). It is assumed that our organism is able to generate all of these metabolites by some unspecified catabolic pathways.
2. At each step a new metabolite that cannot yet be synthesized by the organism is randomly selected from the "scope" [8] of our seed metabolites. This scope consists of all metabolites that in principle could be synthesized from the seed metabolites using all reactions listed in the KEGG database (see Ref. [8] for details).
3. To search for the minimal pathway that converts core metabolites to this target we first perform the "scope expansion" [8] of the core until it first reaches the target. In the course of this expansion reactions and metabolites are added step by step (or layer by layer). Each layer consists of all KEGG reactions that have all their substrates among the metabolites in the current metabolic core of the organism (light blue area in Figure 4) and those generated by reactions in all the previous layers. (See Figure 4 for an illustration).

Table 1. The distribution of irreversible reactions classified by their numbers of substrates and products.

	The number of products of an irreversible reaction				
The number of substrates of an irreversible reaction	**1**	**2**	**3**	**4**	**5**
1	157	141	4		
2	82	491	95	7	
3	1	123	170	31	1
4		10	73	15	
5			1		

doi:10.1371/journal.pcbi.1001137.t001

4. Next we trace back added reactions starting from the target and progressively moving to lower levels. One starts by finding the reaction responsible for fabrication of the target metabolite and adding it to the new pathway (if several such reactions exist in the last layer we randomly choose one of them). In case of multi-layer expansion process some substrates of this reaction are not among the core metabolites (otherwise this reaction would be in the first layer). One then goes down one layer and adds the reactions fabricating these missing substrates. This is repeated all the way down to the first level of the original expansion. The resulting pathway includes the minimal (or nearly minimal) set of reactions needed to generate the target metabolite from the current metabolic core of the organism. Starting from the next step of the model the target and all intermediate metabolites become part of the metabolic core. Genes for enzymes catalyzing these new reactions are assumed to be horizontally transferred to the genome of the organism. The newly added metabolic pathway is assumed to have a dedicated transcriptional regulator so that the number of transcription factors in our model is always equal to the number of pathways or their target metabolites.
5. Steps 1–5 are repeated until metabolic network of the organism reaches its maximal size. At this stage it includes the entire scope [8] of the starting set of metabolites in step 1.

Numerical simulation of this model shows that the number of transcriptional regulators scales with the number of metabolites with power $\alpha = 2.0 \pm 0.1$ (Figure 5). This is consistent with quadratic scaling we observed and mathematically derived for a simpler model with linearized pathways composed of single-substrate reactions.

The mathematical formalism derived in the previous sections is limited to tree-like universal networks and thus does not directly apply to the new model. Nevertheless, one generally expects the quadratic scaling to be limited only to critical, "large world" networks in which organisms with small genomes initially tend to acquire sufficiently long pathways. As noted before, from purely topological standpoint the KEGG network has a "small world" property making long pathways unlikely. It is important to check if the realistic treatment of multi-substrate reactions did in fact restore the "large world" property and criticality to the KEGG universal network by increasing the minimal number of steps required for connecting target metabolites to the metabolic core. To quantify the criticality of the expansion process as before we use the ratio $N_M^{(U)}(d+1)\Big/N_M^{(U)}(d)$ where $N_M^{(U)}(d)$ denotes the number of metabolites reached at step d of the scope expansion starting from the initial seed subset of metabolites. As in the case of critical branching trees this ratio asymptomatically converges to 1 thus confirming the criticality of the scope expansion process. The mere existence of $\sim$40 steps in this process (the x-axis in Figure 6) can serve as evidence in favor of "large world" character of the KEGG universal network characterized by the existence of long pathways.

Geometrical properties of branched pathways in the model

Unlike linearized pathways in the original version of the toolbox model [2], branched pathways in the more realistic model from previous section are interesting objects in their own right. We identified several geometrical properties of these pathways (see Figure 4 for illustration) quantifying their position relative to the core network to which they were added: 1) $n_{\text{border rxn}}$–the number of added reactions that are connected (as a substrate or a product) with at least one metabolite in the core, 2) n_{base}–the number of metabolites in the core that serve as substrates to reactions in the added pathway, 3) n_{feedback}–the number of core metabolites that are products of reactions in the new pathway, 4) $n_{\text{byproduct}}$–the number of final metabolic products of the added pathway that are

Table 2. The distribution of reversible reactions classified by their numbers substrates/products.

	The number of substrates/products at one end of a reversible reaction				
The number of substrates/products at the opposite end of a reversible reaction	**1**	**2**	**3**	**4**	**5**
1	143	231	6		
2		553	284	15	
3			106	69	1
4				6	3

doi:10.1371/journal.pcbi.1001137.t002

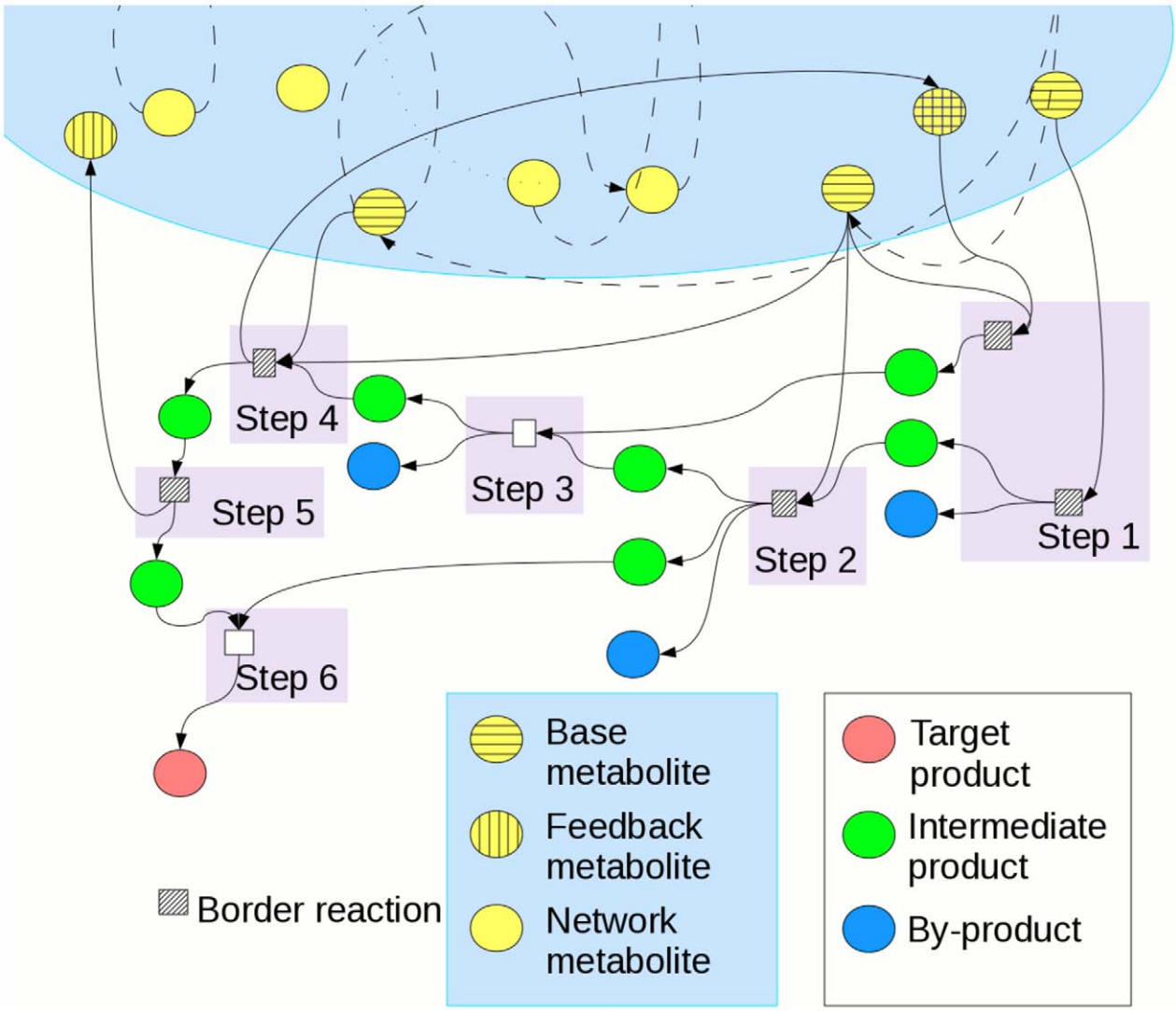


Figure 4. Diagram of a new pathway added to the metabolic network of the organism. The diagram explains different types of metabolites and reactions. Reactions (squares) in the added pathway use base substrates (yellow circles with horizontal shading) from the metabolic core of the organism (light blue area) to produce the target metabolite (the red circle). Added pathway generates intermediate products (green circles) as well as byproducts that are not further converted to the target (blue circles). Products of some reactions feed back into the metabolic core (yellow circles with vertical shading). Reactions are labeled with expansion steps at which they were added to the pathway.
doi:10.1371/journal.pcbi.1001137.g004

neither core metabolites nor the target, 5) length-the number of steps (layers of the scope expansion process) it takes to transform core metabolites into the target product. 4 illustrates the definition of these parameters while Figure 7 and Figure 8 plot these parameters as a function of n_M (the number of metabolites in the added pathway) or n_{rxn} (the number of reactions in the added pathway).

Approximately linear relationship between $n_{\text{border rxn}}$ vs. n_{rxn} (Figure 8a) suggests that added pathways tend to be located at or near the surface of the core metabolic network of the organism. Most of reactions in these pathways use metabolites from this core network either as substrates (n_{base}) or as products (n_{feedback}). Further analysis indicates that "currency metabolites" (common co-factors that serve as substrates or products of many reactions) constitute a significant fraction (~25%) of all core metabolites involved in border reactions (see the section "Analysis of the currency metabolites in the toolbox model" of Text S1 for details). On the other hand, the fact that the number of steps in a pathway (its length) constitutes a good fraction of its overall number of reactions n_{rxn} implies that, in spite of these numerous "shortcut" connections to the core, added pathways remain very thin and essentially linear. That is to say, these pathways tend to work as a single "conveyor belt" sequentially converting intermediate products into each other instead of having two or more parallel "processing lines" and assembling final products of these lines only at final stages of the pathway. This finding provides an intuitive reason why models with branched and linearized pathways have similar scaling properties. One can argue that this is because pathways in both models are essentially linear. Yet, in spite of their linearity and optimality (each has the smallest number of reactions to generate the target from the core) added pathways in the new version of the model are very different from shortest paths on the universal network. As illustrated in Figure 9 the average pathway length is several times longer than the geometrically shortest path between the target and the core.

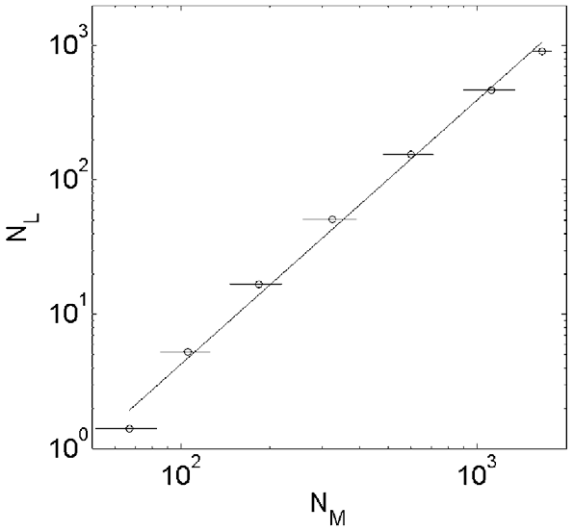


Figure 5. N_L vs. N_M of toolbox model with branched pathways and multi-substrate reactions. The scaling between the number of regulated pathways (leaves), N_L and the number of metabolites, N_M, in metabolic networks generated by the toolbox model with branched pathways and multi-substrate reactions. Solid line with slope 2.0+/−0.1 is the best fit to the data. Error bars reflect the standard deviation of N_M at a given value of N_L in 9 realizations of the model (see he section "Error analysis of the toolbox model" of Text S1 for our estimation methods and error analysis).
doi:10.1371/journal.pcbi.1001137.g005

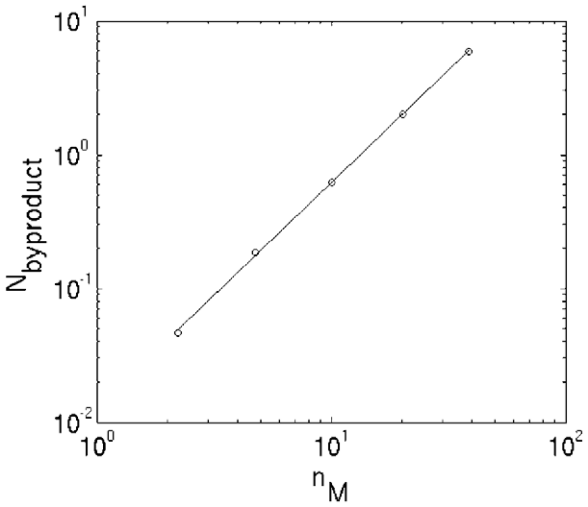


Figure 7. $n_{byproduct}$ vs. n_M. Faster-than-linear scaling of the number of byproducts, $n_{byproduct}$, and the total number of metabolites, n_M, in individual branched pathways illustrated in Figure 4. Data for individual pathways were logarithmically binned along the x-axis. Hence y-axis can be and are below 1 due to pathways with 0 byproducts. The solid line with exponent 1.7+/−0.1 is the best fit to the logarithmically-binned data shown in this plot. Readers can refer to the section "Analysis of number of by-product of the pathways of the toolbox model on the metabolic network with branched pathways and multi-substrates reactions" of Text S1 for our estimation methods and error analysis.
doi:10.1371/journal.pcbi.1001137.g007

As can be seen from Figure 7, most of added pathways (around 97%) do not generate any byproducts. They only produce the intended target and n_{feedback} metabolites in the core network of the organism to which they were added. The relative scarcity of byproducts suggests that pathways in our model satisfy the evolutionary constrains imposed on real-life organisms. Indeed, as previously proposed in Ref [9] it makes sense to assume that evolution favors pathways that achieve a given metabolic goal using the smallest number of enzymes and at the same time striving to generate the maximal possible yield. Unnecessary byproducts not only reduce the yield of the desired metabolic target, they also might become toxic in high concentrations and thus would require extra transporter proteins to pump them out.

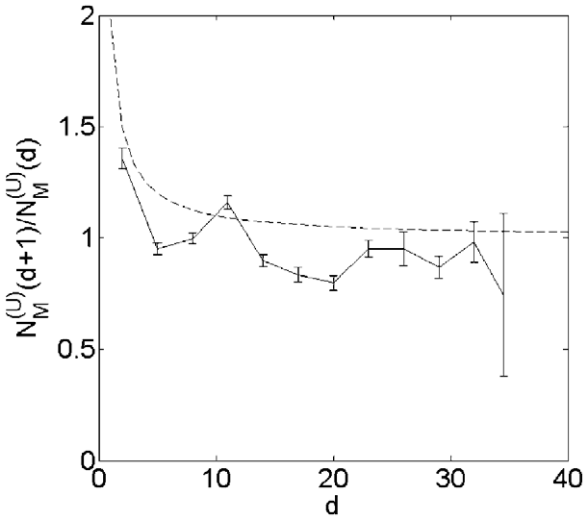


Figure 6. $N_M^{(U)}(d+1)\Big/N_M^{(U)}(d)$ vs. d for the universal network consisting of all KEGG reactions. The ratio $N_M^{(U)}(d+1)\Big/N_M^{(U)}(d)$ of the number of metabolites at two consecutive layers of the scope expansion process plotted versus the layer number d. Scope expansion was performed for the universal network consisting of all KEGG reactions. The dashed line is the mathematical expectation of the same curve in a critical branching process.
doi:10.1371/journal.pcbi.1001137.g006

Discussion

The small world property of complex biomolecular networks has been extensively discussed in the literature during the last decade (see [10–12] for earliest reports in metabolic and protein interaction networks correspondingly). It was often assumed that the small world effect positively contributes to the robustness of the network by providing multiple redundant pathways for target production in metabolic networks or for propagation of signals along regulatory and protein interaction networks. In addition to its positive aspects the small world property in biomolecular networks also has a potentially negative side by facilitating system-wide propagation of undesirable cross-talk [13]. In the course of evolution different strategies appeared allowing organism to limit and attenuate these unwelcome side effects of global connectivity.

The extent of small world topology in metabolic networks has been recently disputed in [14]. There it was argued that many shortcuts in simple graph representations of metabolic networks are meaningless from biochemical standpoint. By taking into account additional structural information about metabolites Arita [14] dramatically increased the diameter of the metabolic network in *E. coli*. In our simulations of the toolbox model we also encountered limitations of the simple graph representation giving rise to small world topology of metabolic networks. Small world by definition implies very short pathways (or equivalently supercritical

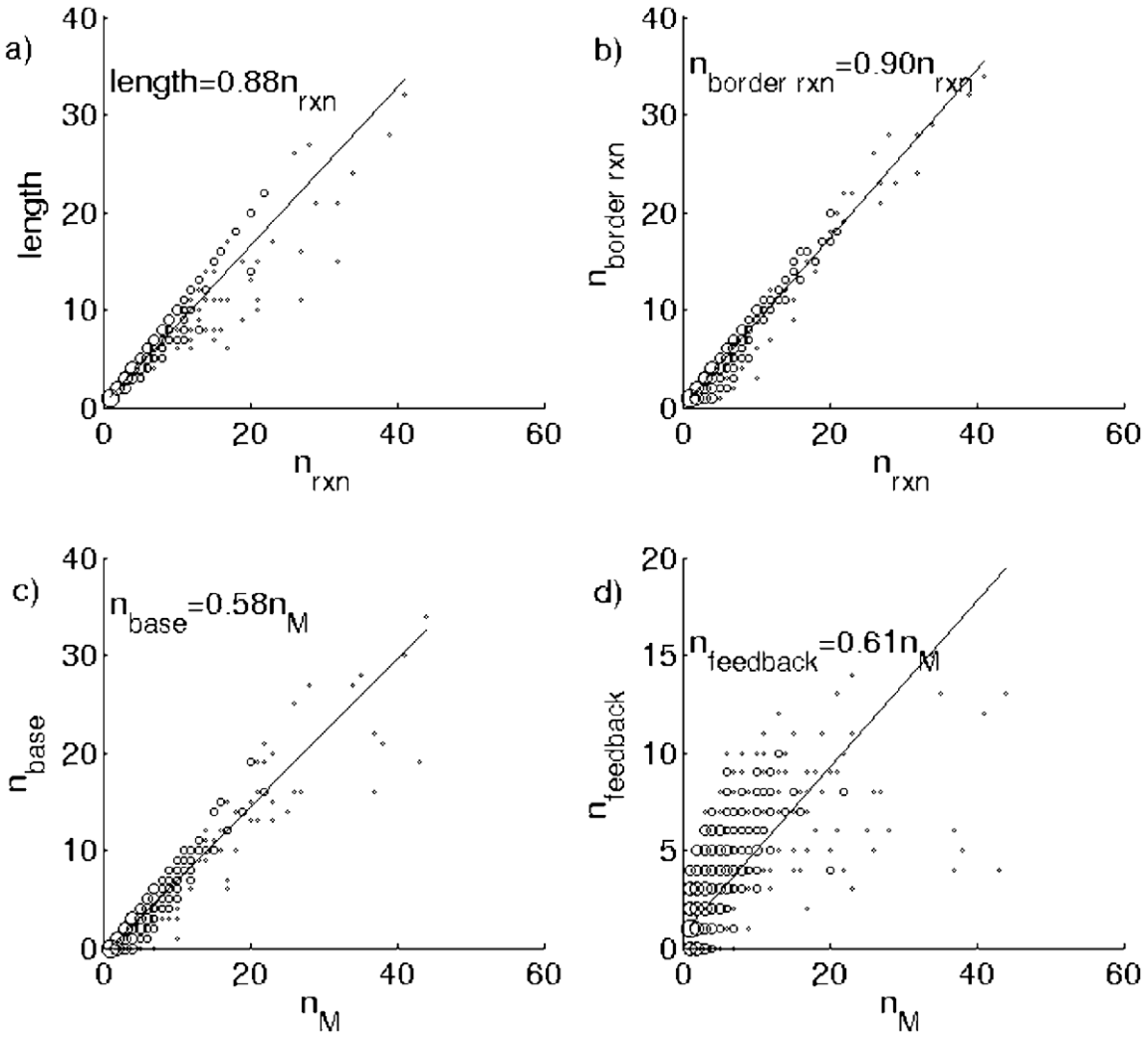


Figure 8. Various linear relationships on the individual pathways. Approximately linear relationship between a) pathway's length and its number of reactions n_{rxn},) b) the number of border reactions, $n_{border\ rxn}$, and the total number of reactions, n_{rxn}, c) the number of base metabolites, n_{base}, and the total number of metabolites, n_M, d) the number of metabolites receiving feedback, $n_{feedback}$, and the total number of metabolites, n_M. These different geometrical properties of individual pathways are illustrated in Figure 4. Sizes of circles are proportional to the logarithm of the number of discrete (x, y) pairs contributing to this point.
doi:10.1371/journal.pcbi.1001137.g008

network branching with exponentially growing lists of neighbors at distance d), which in its turn prevents the appearance of quadratic scaling in the linear toolbox model.

How to reconcile this apparent contradiction? The answer known from pioneering studies of R. Heinrich and collaborators (see e.g. [8,15,16]) is to altogether abandon the simple graph representation in favor of realistic treatment of multi-substrate reactions. A metabolic reaction with two or more substrates will not proceed at any rate until all these metabolites are present in the cell. This implicit "AND" function operating on inputs of multi-substrate metabolic reactions makes reaching a given metabolic target much harder task and ultimately leads to dramatically longer pathways (Figure 9 quantifies this effect). These longer pathways in turn reinstate the quadratic scaling in the version of the toolbox model that was introduced in the previous section. This leads to the novel conclusion of our study that, when multi-substrate reactions are properly taken into account, the small world (supercritical) topology of the metabolic universe disappears in favor of the "large world" topology characteristic of critical branching networks. The increase in the effective diameter of the network due to this effect is dramatic. One goes from 3–4 steps diameter typical of a small world network of [12,11] to ~8 steps of [14] and finally to 30–40 layers in the scope expansion process shown in Figure 6 (see also Figure 6 of [8]).

These arguments lead us to adapt the "scope expansion" algorithm by Heinrich et al [8] to pathway acquisition in the toolbox model. Not only did it restore the "large world" properties such as quadratic scaling to the model, it also made the added pathways plausible from evolutionary standpoint. Unlike linear random walk pathways on KEGG network used in [2], pathways in the new version of the toolbox model have the smallest number of KEGG reactions to achieve their metabolic task (production of the target metabolite from the set of metabolites already present in organism's network). As can be seen in Figure 7 a large fraction of these pathways also does not generate any byproducts. Accumulation of such byproducts inside a cell is potentially dangerous and would require specialized proteins to excrete them to the

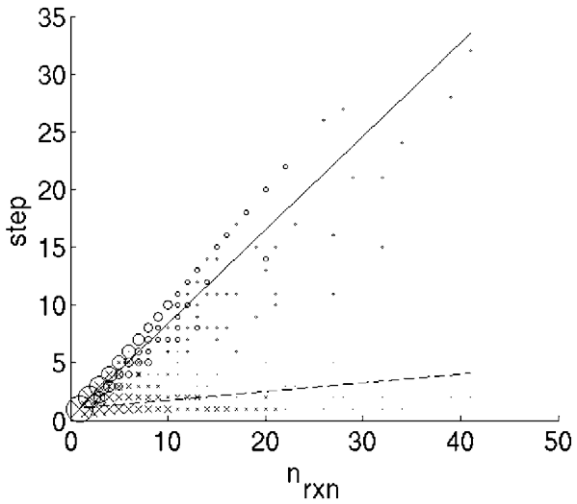


Figure 9. Comparison of lengths of the pathways and shortest distances of the targets from the core. The lengths of the pathways are represented by circles and solid line, while the shortest distances of the targets from the core are represented by crosses and dotted line.
doi:10.1371/journal.pcbi.1001137.g009

environment. The lack of byproducts also means that the useful yield of an added pathway is at or near its theoretical maximum. This is consistent with the fact that real biological pathways are optimized in the course of evolution to increase their yield while simultaneously reducing the number of reaction steps [7,17,18].

Optimality of metabolic pathways in central carbon metabolism was recently discussed in Ref. [17]. There it was shown that some (but not all) of these pathways coincide with the shortest walks in the space of possible metabolic transformations. This study also estimated a typical metabolic substrate can in principle be converted into any of the 20 different products in just one step. This quickly adds up to a very large number of biochemically feasible paths connecting metabolites to each other. However, this exponential growth does not necessarily result in a small world universal metabolic network. Indeed, evolutionary optimization leaves just a tiny fraction of these biochemically feasible reactions to be realized in any organism. The universal metabolic network formed by the union of all organism-specific metabolic networks is thus dramatically sparser than the set of all reactions allowed by the basic rules of biochemistry. Thus, as demonstrated in Ref. [8] and the present study, the number of metabolites one could generate in N steps starting from a small core network and using KEGG-listed metabolic reactions instead of expanding as 20^N grows with N much more slowly (algebraically). The overall picture consistent with both our observations and those of Ref. [17] is that exponentially large, supercritical tree of all possible biochemical transformations is first pruned to an evolutionary optimized critical universal network out of which individual organisms select a subset of reactions necessary to accomplish their metabolic goals: that is to utilize nutrients in their environment and generate metabolic targets necessary for their operation.

Simplified "toy" models based on artificial chemistry reactions have a long history of being used to reveal fundamental organizational principles of metabolic networks:

- The recent model of Riehl et al [18] uses the simplest possible metabolites distinguished from each other only by the number of atoms of one element (e.g. carbon). All reactions in this case are of ligation/cleavage type (e.g. $2+3\leftrightarrow5$) constrained only by mass conservation. In spite of utmost simplicity of this artificial chemistry, the optimal pathways in this model display a surprisingly rich set of properties and bear some similarity to real-life metabolic pathways.
- The study of Pfeiffer et el [19] emphasizes the role of different chemical groups forming metabolites. They consider another artificial chemistry where metabolites are defined by binary strings indicating presence or absence of each of N different chemical groups, and reactions transferring one such chemical group from one substrate which has it to another substrate which initially does not. Plausible evolutionary rules of their model give rise to complex scale-free metabolic networks emerging from the simple initial condition of N completely non-specific transferase enzymes.
- Finally the artificial chemistry studied by Hintze et al [20] has molecules composed of three different types of atoms with different valences. Metabolites are linear molecules in which every atom is connected to others by as many bonds as specified by its valence. This model with rather complicated rules of evolution is then used to shed light on topics such as robustness and modularity of metabolic networks.

In our study we used the real-life (even if incomplete and sometimes noisy) metabolic universe of all reactions in the KEGG database. The only simplifying approximations remaining in the new most realistic version of the toolbox model is random selection of metabolic targets in the course of evolution and easy availability of any subset of KEGG reactions for horizontal transfer. Both these approximations can be relaxed in later versions of the model. Another promising direction is to extend the toolbox model to artificial chemistry universal networks of Refs. [18], [19], [20]. While taking away from the realism of the model such extensions might help to broaden our intuition about what topological properties of the universal network determine the scaling properties of its species-specific subnetworks.

Materials and Methods

The universal network used in our study consists of the union of all reactions listed in the KEGG database. The directionality of reactions and connected pairs of metabolites were inferred from the map version of the reaction formula: ftp.genome.jp/pub/kegg/ligand/reaction/reaction?mapformula.lst. The universal network with linearized pathways used to construct Figure 2 and Figure 3 consists of 1813 metabolites upstream of pyruvate and 2745 reaction edges out of which 1782 are irreversible and 963 are reversible. The metabolic network with branched and cyclic pathways used to construct Figure 5–9 consists of 1861metabolites located downstream from the central metabolism and reachable from it by the scope expansion algorithm of Ref. [8]. It has 2819 reactions out of which 1402 are irreversible and the remaining 1417 are reversible. Table 1 and Table 2 shows the statistics for the number of substrates and products of these reactions. The list of core metabolites is obtained from KEGG Pathways Modules in the category "central carbohydrate metabolism" and extended with "currency" metabolites including water, ATP and NAD. Simulations were done in Matlab and Octave.

Supporting Information

Text S1 Supplementary information.
Found at: doi:10.1371/journal.pcbi.1001137.s001 (0.37 MB DOC)

Acknowledgments

Additional support of this work was provided by the DOE Systems Biology Knowledgebase project "Tools and Models for Integrating Multiple Cellular Networks".

Author Contributions

Conceived and designed the experiments: SM. Performed the experiments: TYP. Analyzed the data: TYP SM. Wrote the paper: TYP SM. Derived mathematical solutions: SM.

References

1. van Nimwegen E (2003) Scaling laws in the functional content of genomes. Trends Genet 19: 479–484.
2. Maslov S, Krishna S, Pang TY, Sneppen K (2009) Toolbox model of evolution of prokaryotic metabolic networks and their regulation. Proc Natl Acad Sci U S A 106: 9743–9748.
3. Athreya KB, Ney PE (2004) Branching Processes. New York: Dover Publications.
4. Broadbent SR, Hammersley JM (1957) Percolation Processes. Math. Proc. Camb. Phil. Soc 53: 629–641.
5. Huang K (1987) 17.5 THE VAN DER WAALS EQUATION OF STATE. In: Statistical mechanics. New York: Wiley. 426 p.
6. Kanehisa M, Goto S (2000) KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res 28: 27–30.
7. Meléndez-Hevia E, Torres NV (1988) Economy of design in metabolic pathways: further remarks on the game of the pentose phosphate cycle. J. Theor. Biol 132: 97–111.
8. Handorf T, Ebenhöh O, Heinrich R (2005) Expanding metabolic networks: scopes of compounds, robustness, and evolution. J. Mol. Evol 61: 498–512.
9. Meléndez-Hevia E, Torres NV (1988) Economy of design in metabolic pathways: further remarks on the game of the pentose phosphate cycle. J. Theor. Biol 132: 97–111.
10. Wagner A (2001) The Yeast Protein Interaction Network Evolves Rapidly and Contains Few Redundant Duplicate Genes. Mol Biol Evol 18: 1283–1292.
11. Wagner A, Fell DA (2001) The small world inside large metabolic networks. Proc. Biol. Sci 268: 1803–1810.
12. Jeong H, Tombor B, Albert R, Oltvai ZN, Barabasi A (2000) The large-scale organization of metabolic networks. Nature 407: 651–654.
13. Maslov S, Ispolatov I (2007) Propagation of large concentration changes in reversible protein-binding networks. Proc Natl Acad Sci U S A 104: 13655–13660.
14. Arita M (2004) The metabolic world of Escherichia coli is not small. Proc Natl Acad Sci U S A 101: 1543–1547.
15. Handorf T, Christian N, Ebenhöh O, Kahn D (2008) An environmental perspective on metabolism. J. Theor. Biol 252: 530–537.
16. Ebenhöh O, Handorf T, Heinrich R (2005) A cross species comparison of metabolic network functions. Genome Inform 16: 203–213.
17. Noor E, Eden E, Milo R, Alon U (2010) Central Carbon Metabolism as a Minimal Biochemical Walk between Precursors for Biomass and Energy. Mol. Cell 39: 809–820.
18. Riehl WJ, Krapivsky PL, Redner S, Segrè D (2010) Signatures of Arithmetic Simplicity in Metabolic Network Architecture. PLoS Comput Biol 6: e1000725.
19. Pfeiffer T, Soyer OS, Bonhoeffer S (2005) The evolution of connectivity in metabolic networks. PLoS Biol 3: e228.
20. Hintze A, Adami C (2008) Evolution of Complex Modular Biological Networks. PLoS Comput Biol 4: e23.

Calculation of the average μ in the toolbox model on a critical tree

The total fraction of metabolites from the universal network that are present in an organism specific network is given by

$$\begin{aligned}\overline{\mu} &= \sum_{d=1}^{d=d_{\max}} N_M(d) \Big/ \sum_{d=1}^{d=d_{\max}} N_M^{(U)}(d) \\ &= \sum_{d=1}^{d=d_{\max}} \mu(d) N_M^{(U)}(d) \Big/ \sum_{d=1}^{d=d_{\max}} N_M^{(U)}(d)\end{aligned}.$$

The boundary condition at the last layer of the tree does not satisfy the Eq. (4) but instead is given by $\mu(d_{\max}) = \tau$. One can easily show that for $d < d_{\max}$ $\mu(d)$ rapidly (exponentially) converges to its steady state value $\mu = \sqrt{\tau}$ and stays at this level for as long as $d \gg 1$ when it starts rising again and ultimately approaches 1 at $d = 0$. In a large critical network the number of nodes in the last and the first several layers is small compared to the total number of nodes in the network. Hence in case of a critical network one has $\overline{\mu} \approx \mu$

Quadratic relation between μ and τ for general critical branching processes

The following equation relates $\mu(d)$ between consecutive layers on arbitrary critical branching process:

$$\mu(d) = \sum_{k=0}^{\infty} p_k \left\{1 - \left[1 - \mu(d+1)\right]^k\right\} + \tau p_0$$

Here p_k is the probability for a node to branch out into *k* nodes, and the term $\left\{1 - \left[1 - \mu(d+1)\right]^k\right\}$ is the probability for a node that branched out into *k* nodes to have at least one of them picked. In the stationary state where μ is independent of *d*, we have

$$
\begin{aligned}
\mu &= \sum_{k=0}^{\infty} p_k \left\{1-\left[1-\mu\right]^k\right\} + \tau p_0 \\
&= 1-\sum_{k=0}^{\infty} p_k \left[1-\mu\right]^k + \tau p_0 \\
&= \tau p_0 + 1 - \sum_{k=0}^{\infty}\sum_{j=k}^{\infty} (-1)^{k+1}\, p_j C_k^j \mu^k \\
&= \tau p_0 + 1 - 1 + \mu\sum_{k=0}^{\infty} k p_k - \mu^2 \sum_{k=0}^{\infty} \frac{k(k-1)}{2} p_k + O\left(\mu^3\right)
\end{aligned}
$$

A critical branching process has $\sum_{k=0}^{\infty} k p_k = 1$, and this allows the first order term in the right hand side cancel with the μ on the left hand side, thus we have

$$
\tau p_0 = \mu^2 \sum_{k=0}^{\infty} \frac{k(k-1)}{2} p_k + O\left(\mu^3\right)
$$

So finally we get $\tau \sim \mu^2 + O\left(\mu^3\right)$ for $\mu, \tau < \lhd 1$, where the quadratic term dominates for small μ.

Alternative derivation:

Another derivation, which is independent of layer-to-layer uniformity of branching probabilities p_i, extends the proof from trees generated by Galton-Watson process to more general situations. This derivation starts with the conservation law describing changes of $N_M^{(U)}(d)$ in the universal network between two consecutive layers (this equation follows from the Eq. (1) in the manuscript when $\mu(d) = \tau = 1$):

$$
N_M^{(U)}(d) = N_M^{(U)}(d+1) - N_B^{(U)}(d) + N_L^{(U)}(d)
$$

For many trees the number of nodes in a layer changes slowly compared to the total number of nodes in a layer. For such trees $N_M^{(U)}(d+1) - N_M^{(U)}(d)$ is small compared with both $N_B^{(U)}(d)$ and $N_L^{(U)}(d)$ and

thus one can approximately write $N_B^{(U)}(d) \cong N_L^{(U)}(d)$. Here as in the main text we assume that the universal network does not have more than 2 branches merging at a given node. In the stationary state where $\mu(d) \cong \mu(d+1) = \mu$ the Eq. (1) from the main text becomes

$\mu N_M^{(U)}(d) = \mu N_M^{(U)}(d+1) - \mu^2 N_B^{(U)}(d) + \tau N_L^{(U)}(d)$ and since $N_B^{(U)}(d) \cong N_L^{(U)}(d)$ and

$N_M^{(U)}(d+1) \cong N_M^{(U)}(d)$ we once again get the quadratic scaling

$\tau = \mu^2$. This argument extends our proof of quadratic scaling to any tree-like universal network in which the number of nodes slowly changes from layer to layer.

Solution to the toolbox model on a supercritical tree

$\mu(d)$ is the fraction of nodes in organism-specific network at distance d from the origin of the tree satisfies the following difference equation:

$$\mu(d) = p\tau + k\mu(d+1) - (k-1+p)[\mu(d+1)]^2 \qquad \text{(S1)}$$

We are interested in small τ and so $\mu(d)$ is small, and by keeping only the leading linear term one gets $\mu(d) = p\tau + k\mu(d+1)$. The last layer is special since it contains only leaves and hence $\mu(d_{\max}) = \tau$.

Iteratively solving Eq. (S1) one gets

$$\begin{aligned}\mu\left(d_{\max}-l\right) &= p\tau + p\tau k + p\tau k^2 + \ldots + p\tau k^{l-1} + \tau k^l \\ &= p\tau\left(k^l-1\right)\big/\left(k-1\right)+\tau k^l \\ &= \tau k^l\left(k-1+p\right)\big/\left(k-1\right)-p\tau\big/\left(k-1\right) \\ &\approx \tau k^l\big/\mu_0\end{aligned} \tag{6}$$

where $\mu_0 = \left(k-1\right)\big/\left(k-1+p\right)$. To arrive at this expression we have made an approximation by dropping the quadratic term in Eq. (S1). This made our estimation for $\mu\left(d\right)$ to increase without saturating at the steady state. To rescue this we assume that $\mu\left(d\right)$ follows the linearized difference equation until it reaches the steady state at the height $d_{\max}-m$, and then $\mu\left(d\right)$ stays as a constant over the region $d < d_{\max}-m$. Setting the equation $\tau k^m/\mu_0 = \mu_0$ we will get

$$m = \log_k\left(\mu_0^2/\tau\right) \tag{7}$$

Now we use the results of eq. (6) and (7) to calculate n_M, picking only the leading order:

$$\begin{aligned}N_M &= \sum_{l=0}^{m} N_M^{(U)}\left(d_{\max}-l\right)\mu\left(d_{\max}-l\right)+\sum_{l=m+1}^{d_{\max}} N_M^{(U)}\left(d_{\max}-l\right)\mu\left(d_{\max}-l\right) \\ &= k^{d_{\max}}\tau+\sum_{l=1}^{m}\left(k^{d_{\max}}/k^l\right)\left(\tau k^l/\mu_0\right)+\sum_{l=m+1}^{d_{\max}}\mu_0 k^{d_{\max}}/k^l \\ &= k^{d_{\max}}\tau+\left(\tau k^{d_{\max}}/\mu_0\right)m+\mu_0 k^{d_{\max}}\sum_{l=m+1}^{d_{\max}} 1/k^l \\ &\approx \left(\tau k^{d_{\max}}/\mu_0\right)\log_k\left(\mu_0^2/\tau\right)\approx N_M^{(U)}\frac{\tau}{\mu_0}\frac{k-1}{k}\log_k\left(\frac{\mu_0^2}{\tau}\right)\end{aligned}$$

Finally we get $\mu = N_M/N_M^{(U)} = \frac{\tau}{\mu_0}\frac{k-1}{k}\log_k\left(\frac{\mu_0^2}{\tau}\right)$, where μ scales with $\tau\log_k\left(\tau\right)$. The

$\tau \log_k(\tau)$ term comes from the summation of $\sum_{l=1}^{m} N_M^{(U)}(d_{\max}-l)\mu(d_{\max}-l)$, which represents the contribution from the last few layers before the saturation of $\mu(d)$.

Error analysis of the toolbox model

The regression of the data of Figure 2 (N_L vs. N_M plot of the toolbox model on critical trees) and Figure 5 (N_L vs. N_M plot of the toolbox model on the metabolic network with branched pathways and multi-substrate reactions) in the manuscript was done first by logarithmically binning the data points along their *y*-coordinate (N_L), and the exponents was then calculated with ordinary least square using the binned curve and taking the *y*-coordinate as the predictor and minimizing the mean square of the difference between the x-coordinate (log N_M in our case) of the binned data and the fitted curve. The best fit coefficients of the linear regression and their 95% confidence intervals were calculated by the "regress" function of the Statistical Toolbox in Matlab 7. We used the *y*-coordinate as the predictor (as opposed to a more traditional use of the x-coordinate as the predictor) because in our simulations N_L was increased in constant (unit) steps (one leaf was added per each step of the model), while the corresponding steps in N_M (added pathways' lengths) varied from simulation to simulation. Thus it was natural to view N_M as a fluctuating function of N_L and not vice versa.

The error bars on slopes (exponents) and prefactors in best linear fits to the binned data in Figure 2 and Figure 5 are based on the 95% confidence intervals estimated by the regress function in Matlab. Figure S1a shows that different fitting methods give consistent results, which indicates that the spread of the raw data is relatively small and will not change our main conclusions obtained from Figure 2 and Figure 5.

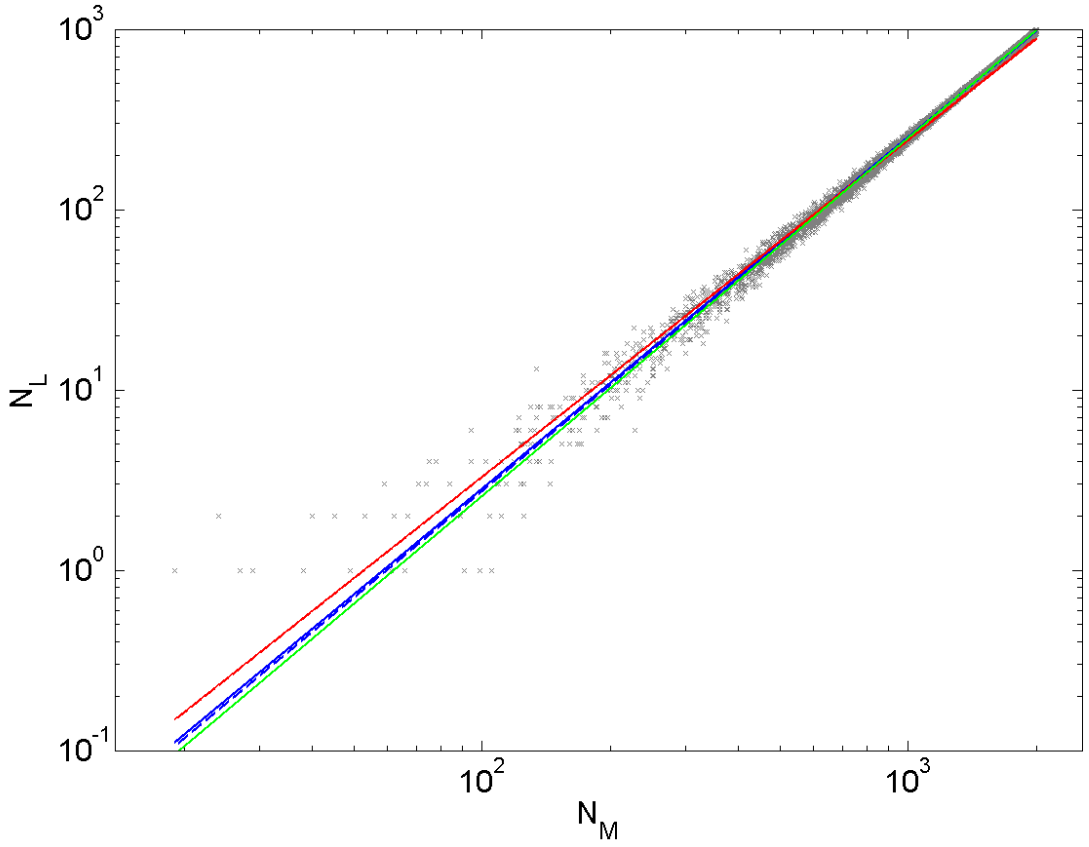


Figure S1a Comparison of different ways of performing the regression analysis.

N_L *vs.* N_M *data of the toolbox model on 10 different critical trees generated with probability* $p_0 = 0.5$ *and* $p_2 = 0.5$. *The grey crosses represents the raw data; the blue solid line is the best fit curve derived from linear regression, i.e., by taking the x-coordinate as the predictor and minimizing the mean square difference in y-coordinate between the best fit line and the data points in logarithmic scale; the blue dash line is the best fit curve derived from linear regression taking the y-coordinate as the predictor; the red solid line is the best fit curve obtained by first logarithmically binning the data along the y-coordinate followed by linear regression on the binned curve using the x-coordinate as the predictor in logarithmic scale; the red dash line is the best fit curve obtained by first logarithmically binning the data with their y-coordinate followed by linear regression on the binned curve using the y-coordinate as the predictor in logarithmic scale; the green solid line is the best fit curve obtained from powerlaw fitting of the raw*

data; the green solid curve is obtained by binning the along the y-coordinate and followed by fitting with powerlaw along the binned curve. All these best fit curves have exponent 1.9.

Furthermore, as shown in Figure S1b, the exponent varies when we consider different range of N_L in the regression analysis, and this indicates the existence of systematic error. The presence of the systematic error suggests that we cannot use the conventional regression analysis, because the error of the regression coefficients inversely depends on the square root of the number of data points. In our case we want the error to account for the systematic differences of exponent, and therefore we binned the data followed by regression. The binning of the data reduces the number of points and so the regression analysis that follows can reflect the change of the exponent along different regions, i.e., the systematic error, and get rid of the size effect of the size raw data.

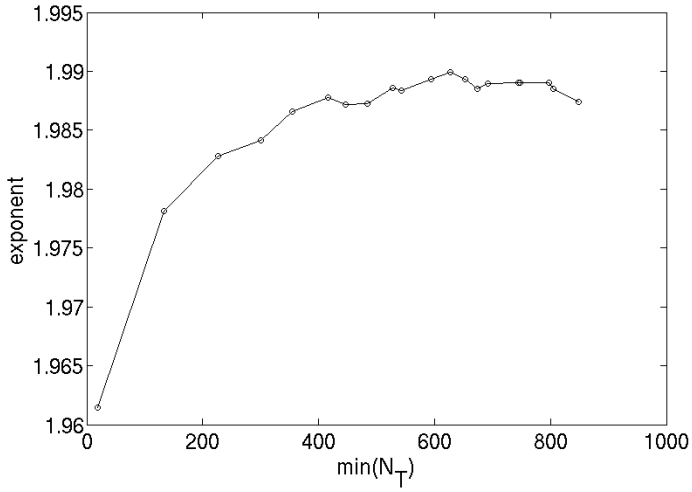


Figure S1b Exponent obtained from regression on different ranges of N_L.

The exponent is obtained by binning of data along the y-coordinates (N_L) followed by regression using the x-coordinate as the predictor (N_M) in logarithmic scale. The regression analyses are performed on different ranges of N_L, where all share the same upper limit but their lower limits have different cutoffs.

Analysis of number of by-product of the pathways of the toolbox model on the metabolic network with branched pathways and multi-substrates reactions

Figure 7 of the manuscript shows the binned curve of $n_{byproduct}$ vs. n_M but does not show any error or raw data. The raw data of the plot is shown in Figure S3, and the binning was done by first logarithmically grouping the data points according to their *x*-coordinate (n_M) followed by arithmetic averaging of the *y*-coordinates of the data points ($n_{byproduct}$). Geometric averaging on the *y*-coordinate was not used despite that the *x*-coordinate was binned logarithmically, because the majority of their values is 0 or 1.

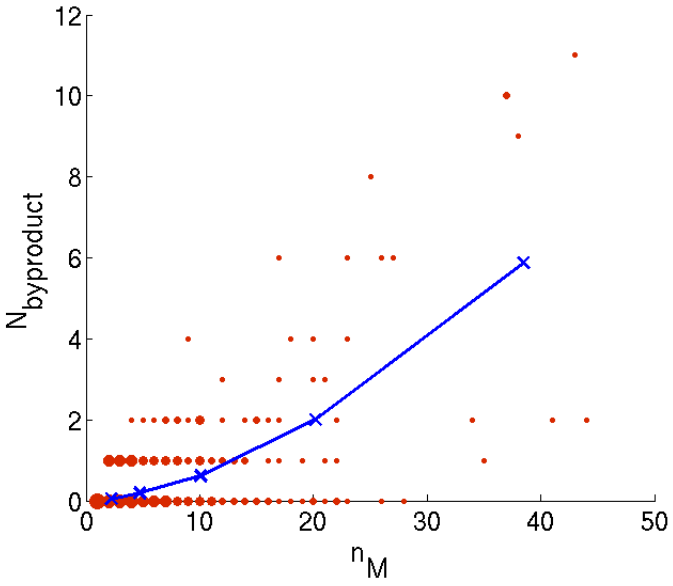


Figure S2 $n_{bzproduct}$ vs. n_M *of the toolbox model in the metabolic network with branched pathways and multi-substrate reactions. Red circles: raw data; the scale of the circles is proportional to the logarithmic of its occurrence. Blue circles and dash line: binning of data by partitioning the x-axis logarithmically and average over y coordinates arithmetically.*

Seed metabolites of the scope expansion model

The following 40 metabolites, except H2O, ATP and NAD, belong to the KEGG modules of central carbohydrate metabolism. They all serve as the starting core metabolites in the simulations of scope

expansion.

Table S1. Seed metabolites of the scope expansion model

KEGG Entry number	Name
C00001	H2O
C00002	ATP
C00003	NAD+
C00022	pyruvate
C00024	Acetyl-CoA
C00026	2-Oxoglutarate
C00036	Oxaloacetate
C00042	Succinate
C00074	Phosphoenolpyruvate
C00085	D-Fructose 6-phosphate
C00091	Succinyl-CoA
C00111	Glycerone phosphate
C00117	D-Ribose 5-phosphate
C00118	D-Glyceraldehyde 3-phosphate
C00119	5-Phospho-alpha-D-ribose 1-diphosphate
C00122	Fumarate
C00149	L-Malate
C00158	Citrate
C00197	3-Phospho-D-glycerate
C00199	D-Ribulose 5-phosphate
C00204	2-Dehydro-3-deoxy-D-gluconate
C00231	D-Xylulose 5-phosphate
C00236	3-Phospho-D-glyceroyl phosphate
C00257	D-Gluconate
C00267	alpha-D-Glucose
C00279	D-Erythrose 4-phosphate
C00311	Isocitrate
C00345	6-Phospho-D-gluconate
C00577	D-Glyceraldehyde
C00631	2-Phospho-D-glycerate
C00668	alpha-D-Glucose 6-phosphate
C01172	beta-D-Glucose 6-phosphate
C01236	D-Glucono-1,5-lactone 6-phosphate
C04442	2-Dehydro-3-deoxy-6-phospho-D-gluconate
C05345	beta-D-Fructose 6-phosphate
C05378	beta-D-Fructose 1,6-bisphosphate
C05382	Sedoheptulose 7-phosphate

C15972	Lipoamide-E
C15973	Dihydrolipoamide-E
C16254	S-Succinyldihydrolipoamide-E

Alternative model: each organism evolves its own network in the absence of horizontal gene transfers

In the toolbox model where the organisms evolve mainly by HGT, the union of all metabolic enzymes forms the universal network which can be represented by a critical tree. One might wonder if the N_L vs. N_M plot will remain quadratic if we turn off the HGT and assume that de novo formation of new protein is the only way to evolve, and the metabolic network of each organism is an independent tree. To study this alternative model, we repeated the Galton-Walton process and generate a set of critical trees, and consider the N_L vs. N_M plot. Theoretically, the N_L vs. N_M plot of this alternative model will no longer demonstrate any quadratic scaling but only a linear one, with slope being p_0, the probability for a node to terminate in a branching process. The simulation of the alternative model was performed by repeating the branching process, and that $N_L = p_0 N_M$ as verified (see Figure S3).

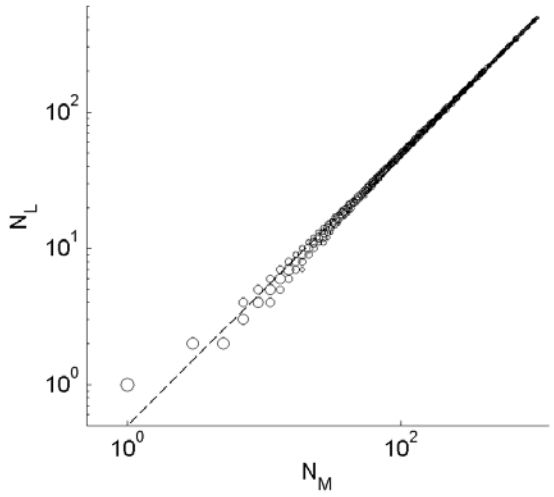


Figure S3 N_L vs. N_M *of trees. The trees are generated with Galton-Watson process with probability 0.5*

to terminate and 0.5 to branch out into 2 children. Scale of circles is proportional to the logarithm of the density of data and the dash line is the theoretical prediction, which is $N_L = p_0 N_M$

Analysis of currency metabolites in the toolbox model

The currency metabolites, i.e., metabolites that serve as inputs for a large number of pathways, have special roles in the metabolic network. In the analysis of the border reactions of on the metabolic pathways of the toolbox model in the manuscript, we concluded, without taking into account of the currency metabolites, that the pathways of the toolbox model are all most linear and lie on the surface of the core, receiving metabolites from the core and feeding them back. One might wonder how the geometry of the pathways will be affected if we take into account the currency metabolites. To study their effects, first of all, let us define l_i to be the number of times node *i* act as a substrate or feedback to a pathway in one simulation. The toolbox model simulation was repeated several times and we can get the average l_i. Analysis of the data showed that the distribution of l_i is divided into two groups (the two steep kinks in Figure S4a), and we picked the top ten metabolites with the largest average l_i to be currency metabolites and plotted the $n_{border\ rxn}$ vs. n_{rxn} (Figure S4b). The average number of currency metabolites that a typical pathway connects is around 1 (data no shown).

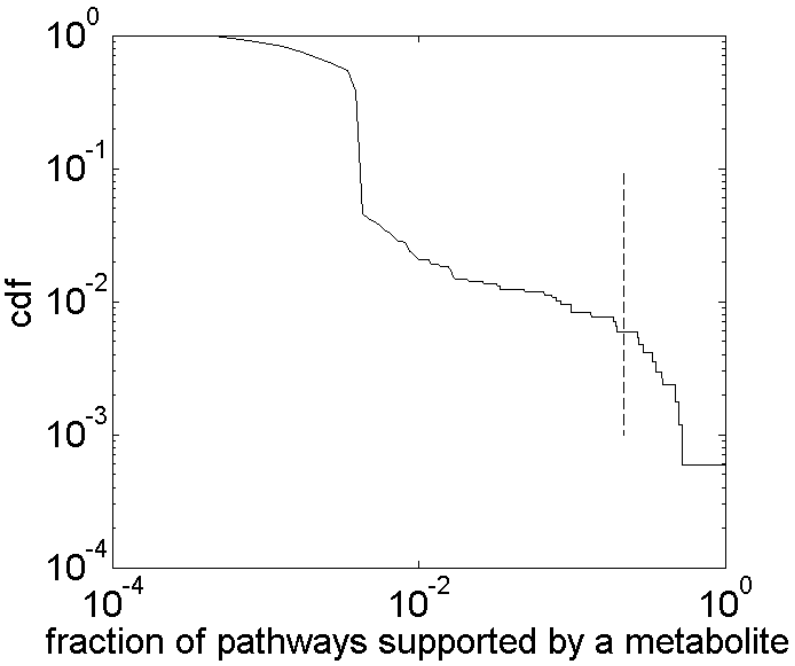


Figure S4a Distribution of the fraction of pathways a metabolite supports. The dash line indicates the cutoff, where to the right of the dash line defines the currency metabolites.

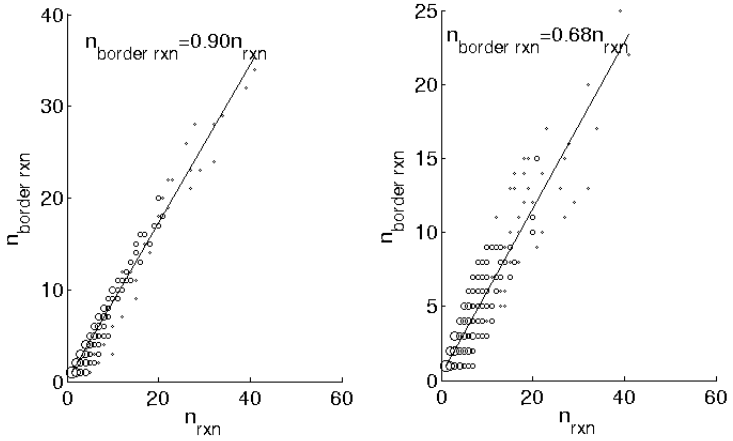


Figure S4b $n_{border\ rxn}$ vs. n_{rxn} *before (left) and after (right) the removal of currency metabolites.*

The figures indicate that the removal of currency metabolites does affect the number of border reactions and reduce their number by a quarter. Nevertheless this does not change our conclusion that pathways are rather linear and lie on the surface of metabolic core.