Toolbox model of evolution of prokaryotic metabolic networks and their regulation

Sergei Maslov [*], **Sandeep Krishna** [†], **and Kim Sneppen** [†]

[*]Department of Condensed Matter Physics and Materials Science, Brookhaven National Laboratory, Upton, New York 11973, USA, and [†]Niels Bohr Institute, Blegdamsvej 17, DK-2100, Copenhagen, Denmark

Submitted to Proceedings of the National Academy of Sciences of the United States of America

It has been reported that the number of transcription factors encoded in prokaryotic genomes scales approximately quadratically with their total number of genes. As a the fraction of transcriptional regulators in some of the smallest genomes is below 0.5%, while in large genomes it approaches 10%. We propose a conceptual explanation of this empirical observation and illustrate it using a simple model in which metabolic and regulatory networks of prokaryotes are shaped by horizontal gene transfer of co-regulated metabolic pathways. Metabolic enzymes or by extension all non-regulatory proteins in the genome of an organism are viewed as its collection of tools. Adapting to a new condition monitored by a dedicated new transcription factor (e.g. learning to use a new nutrient) involves acquiring new enzymes as well as reusing some of the tools/enzymes that are already encoded in the genome. As the toolbox of an organism grows larger, it can reuse its tools more often, and thus needs to get fewer new ones to master each new regulated task. From this observation it logically follows that the number of functional tasks and their regulators increases faster than linearly with the total number of genes. Genomes can also shrink e.g. due to a loss of a nutrient from the environment followed by deletion of its regulator and all enzymes that become redundant. We propose several simple models of network evolution elaborating on this toolbox argument and reproducing the empirically observed quadratic scaling. The distribution of lengths of co-regulated pathways in our model quantitatively agrees with that in real-life metabolic network of E. coli. Furthermore, our model provides a qualitative explanation to a broad distribution of regulon sizes in this and other prokaryotes.

Horizontal Gene Transfer, Transcriptional regulatory networks, Functional genome analysis

Abbreviations: HGT, Horizontal Gene Transfer; KEGG, Kyoto Encyclopedia of Genes and Genomes; TF, Transcription Factor

Introduction

Biological functioning of a living cell involves coordinated activity of its metabolic and regulatory networks. While the metabolic network specifies which biochemical reactions the cell is in principle able to carry out, its actual operation in a given environment is orchestrated by the transcription regulatory network through up- or down-regulation of enzyme levels. A large size of the interface between these two networks in prokaryotes is indicated by the fact that nearly half of transcription factors in *E.coli* have a binding site for a small molecule [1], which implicates them [2] as potential regulators of metabolic pathways. The metabolic fraction of a transcription regulatory network is further increased by two component systems whose sensors bind to small molecules and only then activate a dedicated transcription factor. Thus, at least in prokaryotes, regulation of metabolism occupies the majority of all transcription factors.

Two recent empirical observations shed additional light on evolutionary processes shaping these two networks:

- The number of transcriptional regulators is shown to grow faster than linearly [3, 4, 5, 6] (approximately quadratically [4]) with the total number of proteins encoded in a prokaryotic genome.
- The distribution of sizes of co-regulated pathways (regulons), which in network language correspond to out-degrees of transcription factors in the regulatory network, has long tails [7]. As a result the set of transcription factors of each organism includes few global ("hub") regulators controlling hundreds of genes, many local regulators controlling several targets each, and all regulon sizes in-between these two extremes.

A simple evolutionary model explains both these empirical observations in a unified framework based on modular functional design of prokaryotic metabolic networks and their regulation.

A toolbox view of metabolic networks.. Metabolic networks are composed of many semi-autonomous functional modules corresponding to traditional metabolic pathways [8] or their subunits [9]). Constituent genes of such evolutionary modules tend to co-occur (be either all present or all absent) in genomes [10, 9]. These pathways overlap with each other to form branched, interconnected metabolic networks. Many of these pathways/branches include a dedicated transcription factor turning them on under appropriate environmental conditions. In prokaryotic organisms there is a strong positive correlation between the number of protein-coding genes in their genomes, the number of metabolic pathways formed by these genes, the number of transcription factors regulating these pathways, and, finally, the number of environments or conditions that organism is adapted to live in.

We propose to view the repertoire of metabolic enzymes of an organism as its toolbox. Each metabolic pathway is then a collection of tools (enzymes), which enables the organism to utilize a particular metabolite by progressively breaking it down to simpler components, or, alternatively, to synthesize a more complex metabolite from simpler ingredients. Adapting to a new environmental condition e.g. learning to metabolize a new nutrient, involves acquiring some new tools as well as reusing some of the tools/enzymes that are already encoded in the genome. From this analogy it is clear that as the toolbox of an organism grows larger, on average, it needs to acquire fewer and fewer new tools to master each new metabolic task. This is because larger toolboxes are more likely to already contain some of the tools necessary for the new function. Therefore, the number of proteins encoded in organism's genome (i.e. the size of its toolbox) is expected to increase *slower than linearly* with the number of metabolic tasks it can accomplish. Or, conversely, the number of nutrients an organism can utilize via distinct metabolic pathways is expected to scale *faster than linearly* with its number of enzymes or reactions in its metabolic network. This last prediction is empirically confirmed the data in the KEGG database [8]: as shown in Fig. S6 in supplementary materials the best powerlaw fit to the number of metabolic pathways vs the

Reserved for Publication Footnotes

number of metabolic reactions in prokaryotic genomes has the exponent 2.2 ± 0.3. This is in agreement with quadratic scaling of the number of transcription factors [4] if one assumes that most of these pathways are regulated by a dedicated transcription factor.

Results

Evolution of networks by random removal and addition of pathways. We propose a simple model of evolution of metabolic and regulatory networks based on this toolbox viewpoint. The metabolic network of a given organism constitutes a subset of the "universal biochemistry" network, formed by the union of all metabolites and metabolic reactions taking place in any organism. An approximation to this universal biochemistry can be obtained by combining all currently known metabolic reactions in the KEGG database [8]. For prokaryotes, entire metabolic pathways from this universal network could be added all at once by the virtue of Horizontal Gene Transfer (HGT), which according to Ref. [11] is the dominant form of evolution of bacterial metabolic networks. Recent studies [12] reported a number of HGT "highways" or preferential directions of horizontal gene transfer between major divisions of prokaryotes. As a result of these and other constraints the effective size of the universal network from which an organism gets most new pathways is likely be somewhat smaller than the simple union of reactions in all organisms. Metabolic networks can also shrink due to removal of pathways. This often happens when a nutrient disappears from the environment of an organism over an evolutionary significant time interval (see "use it or loose it" principle by Savageau [13]). A massive elimination of pathways occurs e.g when an organism becomes obligate parasite fully relying on its host for "pre-processing" of most nutrients.

The state-of-the-art information on metabolic networks is not adequate for a fully realistic modeling of their evolution. Fortunately, faster-than-linear scaling of the number of pathways and their regulators with the number of genes is the robust outcome of the toolbox evolution scenario and as such it is not particularly sensitive to topological structure of the universal biochemistry network. In particular we found (see Fig. S1) essentially identical scaling in two models using two very different variants of the universal biochemistry network:

- the union of KEGG reactions [8] in all organisms. The part of this network connected to the biomass production consists of $N_{univ} \simeq 1800$ metabolites;
- a tree made by random walks on the fully connected graph of N_{univ} metabolites. While certainly not realistic, this version is mathematically tractable.

Furthermore, it turned out that many details of pathway acquisition process are irrelevant for the final result as well (see Fig. S2 in Supplementary materials). In the rest of this study we use the first universal network (KEGG union) in our numerical simulations and the second network in our mathematical analysis.

While toolbox view of evolution is equally applicable to catabolic (breakdown of nutrients) and anabolic (synthesis of complex metabolites) pathways, for simplicity we will simulate only addition of catabolic branches. Given the repertoire of enzymes of an organism each of the N_{univ} universal metabolites can be categorized as either "metabolizable" (connected to biomass production), or "non-metabolizable" (currently outside of the metabolic network). To add a new branch to the network in our model we first randomly choose a non-metabolizable molecule as a new nutrient (leaf). A pathway/branch that begins at the leaf and connects it to the set of metabolizable molecules is then added to the network. This connecting pathway consists of a linear chain of reactions randomly selected from the universal network until it first intersects with the currently existing metabolic network of the organism. The leaf plus all the intermediate metabolites of this branch thereby become metabolizable. This process is illustrated in Fig. 1A.

In our model pathway additions and removals are treated in a symmetric fashion. The steps leading to pathway deletion are illustrated in Fig. 1B: First, one of the leaves of the network corresponding to the vanished nutrient is chosen randomly. The branch of the network starting at this nutrient/leaf is followed downstream to the point where it first intersects another branch of the network. This entire path, starting from the leaf down to the merging point with another pathway is then removed from the network. The selected nutrient along with all intermediate metabolites thereby become non-metabolizable.

The network in our model evolves by a random sequence of pathway additions and removals (see Methods for more details). Since our goal is to understand how properties of metabolic and regulatory networks scale with the genome size of an organism, we take multiple snapshots of the evolving network with different values of N_{met} – the current number of nodes in the metabolic network, which in our model is equal to the number of reactions or metabolic enzymes.

Assigning transcriptional regulators to metabolic pathways. Operation of metabolic networks involves regulating production of enzymes in response to nutrient availability. In prokaryotes most of this regulation is achieved at the transcriptional level. In order to investigate the interface between metabolic and regulatory networks we extend our model to include transcription factors (TFs) regulating individual metabolic pathways. In the basic version of our model shown in Fig. 2A we chose the following simple method to assign TFs to reactions: one randomly picks a leaf/nutrient and follows its reactions downstream until this branch either reaches the metabolic core or merges with a pathway regulated by a previously assigned TF. A new TF is then assigned to regulate all reactions in this part of the nutrient utilization pathway. This process is repeated until all enzymes/reactions have been assigned a (unique) transcriptional regulator (see Fig. 2A). Each TF is activated by the presence of the corresponding nutrient in the environment. Note that this method results in exactly one TF per nutrient, and that the out-degree distribution of TFs in the regulatory network is identical to the distribution of branch lengths in the metabolic network.

In addition to this simple rule we have tried several others illustrated in Figs. 2B-D. The advantage of these more complicated regulatory topologies is that they ensure that on/off states of connected metabolic pathways are properly coordinated with each other. For example, unlike the basic scheme in Fig. 2A, those shown in Figs. 2B-D turn on the downstream (and only the downstream) part of the blue pathway in response to presence of the nutrient utilized by the red branch. We will further compare network topologies generated by these rules in the Discussion section.

Comparison of the model with empirical data. In agreement with the toolbox argument outlined in the introduction, we found (see Fig. 4A) that the number of transcriptional regulators of an organism scales steeper than linearly with the total number of metabolites in its metabolic network, which in our model is equal to its number of reactions or enzymes:

$$N_{TF} \propto (N_{met})^{\alpha}, \qquad [1]$$

where the best fit to the exponent is $\alpha = 1.8 \pm 0.3$. To directly compare the results of the model (red diamonds) to the empirical scaling of the number of transcription factors with the number of genes (green circles) in Fig. 4A we multiplied the number of metabolites/reactions N_{met} by a constant factor to approximate the total number of genes N_{genes} in the corresponding genome. The ratio $N_{met}/N_{genes} \sim 0.2$ was estimated as follows: metabolic enzymes constitute about a quarter of all genes in a procaryotic genome independent of its size (see blue line in Fig. 1a of [4]). Due to presence of isoenzymes the number of different reactions catalyzed by these enzymes (equal to the number of metabolites N_{met} in our model) is somewhat smaller and its average value over 451 fully sequenced prokaryotic genomes [14] is 20%. The model results shown in Fig.

4 were simulated on the universal network formed by the union of KEGG reactions in all organisms. However, a model simulated on a random universal network of the same size $N_{univ} \simeq 1800$ produced essentially identical results (black crosses in Fig. S1). This agreement indicates that the scaling between N_{TF} and N_{met} for the most part is determined by just the number of universal metabolites – N_{univ} and is much less affected by the topology of connections between them. On the other hand, we believe that the nearly precise agreement in the actual number of regulators in real prokaryotic genomes and the model with $N_{univ} \simeq 1800$ is coincidental. Indeed, even in prokaryotes not all transcription factors are dedicated to regulation of metabolic enzymes. For example, *E. coli* this fraction appears to be above 50% [1] which means that the actual universal network might be up to twice as large as the one we use in our current model.

In addition to explaining the quadratic scaling between numbers of leaves and all nodes, our model nicely reproduces other large-scale properties of real-life metabolic networks. A sample metabolic network generated by simulating our model is shown in Fig. 3B. Its tree-like topology reflects the fact that in our simplified model each reaction converts a single substrate to a single product. The network is hierarchical in the sense that smaller linear pathways consisting of just a few reactions tend to be attached to progressively longer and longer pathways, until they finally reach the metabolic core. This architecture is reminiscent of drainage networks in which many short tributaries merge to give rise to larger rivers. For comparison, Fig. 3A shows a tree-like backbone (to match linear pathways in our model) of the *E. coli* metabolic network [8, 14]. The details of generating this backbone are described in the Methods section. The hierarchical structure of branches in this real-life metabolic network (Fig. 3A) resembles that of the model network (Fig. 3B) of a comparable size. To quantify this visual comparison in Fig. 4B we compare cumulative branch lengths distributions $P(K_{out} \geq K)$ in our model with $N_{met} = 400$ (red diamonds for $N_{univ} = 1800$ and red squares for $N_{univ} = 900$) and in real metabolic network in *E. coli* of comparable size (green circles). Branch length distributions are characterized by a long powerlaw tail: $P(K_{out}) \sim K_{out}^{-\gamma}$. Best fit value of the tail exponent $\gamma = 2.9 \pm 0.3$ is similar in model and real-life network and agrees with our analytical result $\gamma = 3$ derived in the next chapter. Furthermore, the data in our model simulated on a truncated univversal network with $N_{univ} = 900$ (red crosses in Fig. 4B) are in excellent agreement with that in *E. coli* (green circles) throughout the whole range.

In Fig. S3 we compare the distributions of regulon sizes (branch lengths) in our model (red diamonds in Fig. 4B) and in the Regulon database [15] consisting of all presently known transcriptional regulations in *E. coli*. One can immediately see that the tail of the distribution in the Regulon database with $\gamma \simeq 2$ is considerably broader than in our model. There are several possible explanations of this discrepancy: 1) coordination of activity of different metabolic pathways with each other as shown in Figs. 2B-D gives rise to larger hubs; 2) regulation of proteins other than metabolic enzymes in the same regulon; 3) an anthropogenic effect in which better studied transcription factors included in the regulon database have larger-than-average out-degrees. In the Discussion section we return to comparison real-life and model regulatory networks in more details.

Mathematical derivation of scaling behavior in toolbox model. When a new nutrient (leaf) is added to a network of size N_{met}, the length of the metabolic pathway required for its utilization is (on average) inversely proportional to N_{met}. It is easy to show for a mean-field version of the model in which the universal network is a tree formed by random walks in the space of N_{univ} metabolites. In this case each reaction in the new pathway has the same probability $p = N_{met}/N_{univ}$ to produce one of the N_{met} currently metabolizable molecules. The minimal pathway required for utilization of the new nutrient involves only the reactions until the first intersection with the already existing metabolic network. The average length of such pathway is just the inverse of this probability: $1/p = N_{univ}/N_{met}$. When this pathway is added, the number of metabolizable molecules increases by $\Delta N_{met} = N_{univ}/N_{met}$ and the number of regulators increases by one: $\Delta N_{TF} = 1$. In the steady state of the model, removal of a branch produces the opposite result: $\Delta N_{met} = -N_{univ}/N_{met}, \Delta N_{TF} = -1$. In both cases one has:

$$\frac{dN_{met}}{dN_{TF}} = \frac{N_{univ}}{N_{met}} \qquad [2]$$

the integration of which gives

$$N_{TF} = \frac{N_{met}^2}{2N_{univ}}. \qquad [3]$$

Therefore, the quadratic scaling between N_{TF} and N_{met} naturally emerges from our toolbox model of addition/removal of metabolic pathways.

Scale-free distribution of branch lengths (regulon sizes) in our model: $N(K_{out}) \sim K_{out}^{-\gamma}$ follows from the fact that the average length of a newly added metabolic pathway (or out-degree of its regulator in Fig. 2B) is $K_{out} = N_{univ}/N_{met}. = \sqrt{N_{univ}/2N_{TF}}$. As the size of the metabolic network increases, the length of each new pathway progressively shrinks. If the network was monotonically growing, pathway distribution would have the exponent $\gamma = 3$. Indeed, in this case a typical pathway of length $K_{out} \geq K$ was added at the time when the number of metabolites was smaller than N_{univ}/K or equivalently the number of transcription factors was below $N_{univ}/(2K^2)$. Therefore, $P(K_{out} \geq K) = N_{univ}/(2K^2 N_{TF})$ or $P(K_{out} = K) \sim N_{univ}/(N_{TF}K^3)$. As evident from Fig. 4B random cycling through addition and removal of pathways does not significantly change this exponent.

Discussion

Trends of average in- and out-degrees in the regulatory network as a function of genome size. As was pointed out by van Nimwegen in Ref. [4, 16, 17] faster-than-linear scaling of the number of transcription factors results in systematic differences in topology of transcriptional regulatory networks as a function of genome size. Indeed, the total number of edges in this network can be written either as $N_{genes}\langle K_{in}\rangle$ if one counts the incoming regulatory inputs of all its genes, or as $N_{TF}\langle K_{out}\rangle$ if one counts the regulatory outputs of all transcription factors. Here the brackets denote the average over a given genome. Therefore, one always has

$$\frac{N_{TF}}{N_{genes}} = \frac{\langle K_{in}\rangle}{\langle K_{out}\rangle} \qquad [4]$$

The empirical data [3, 4] indicate that the left hand side of this equation monotonically grows with N_{genes}. Therefore, an increase in genome size of an organism must be accompanied either by an increase in the average in-degree $\langle K_{in}\rangle$ of all its genes or by a decrease in the average out-degree $\langle K_{out}\rangle$ of its transcriptional regulators. The empirical observation [16] that average operon size (a lower bound on the regulon size or the length of co-regulated pathway) is negatively correlated with N_{genes}. On the other hand, a recent study by van Nimwegen and collaborators [17] found that small and large prokaryotic genomes have approximately the same average number of regulatory binding sites per gene suggesting no systematic correlation between $\langle K_{in}\rangle$ and N_{genes}. These two observations suggest that the dominant trend is that $\langle K_{out}\rangle$ decreases with N_{genes}. This is consistent with our basic model (Fig. 2A) in which K_{out} of transcription factors regulating newly added metabolic pathways progressively decreases with N_{met} (or N_{genes}). However, the lack of coordination between upstream and downstream pathways in this

model is not realistic. The model shown in Fig. 2B introduces a complete top-to-bottom regulation of utilization of each nutrient. For this model one expects $\langle K_{out} \rangle$ to be independent of and $\langle K_{in} \rangle$ to increase with N_{genes}. Real-life regulatory networks are likely to be somewhere in-between these two extreme scenarios.

Coordination of activity of different metabolic pathways. Converting a nutrient into the biomass often involves several successive pathways each regulated by its own transcription factor. The state of activity of such pathways has to be coordinated with each other. Our basic model illustrated in Fig. 2A does not involve such coordination. In this model:

- Transcription factors do not regulate other transcription factors. This results in "shallow" transcriptional regulatory networks consisting of only two hierarchical layers: the upper level including all regulators, and the lower level including all workhorse proteins (metabolic enzymes). While this assumption in its pure form is certainly unrealistic, it approximates the hierarchical structure of real prokaryotic regulatory networks, which were shown to be relatively shallow [7, 18, 19]. That is to say, the number of hierarchical layers in these networks was shown to be smaller than expected by pure chance [19].
- In the regulatory network shown in Fig. 2A every enzyme is regulated by precisely one transcription factor. Once again this feature, while obviously unrealistic, approximates topological properties of real-life regulatory networks e.g one in *E. coli*. In [7] it was shown that in this network the in-degree distribution peaks at one regulatory input per protein beyond which it rapidly (exponentially) decays. This should be contrasted with a broad out-degree (regulon size) distribution [7] which has a long power-law tail reaching as high as hundreds of targets.

Several of many possible scenarios for coordination of activity of different pathways shown in Figs. 2B-D) all deviate from these two simplified assumptions. Models shown in Fig. 2C-D solve the coordination problem by adding regulatory interactions among transcription factors of individual branches/pathways. The positive regulation TF2 $\rightarrow$ TF1 in Fig. 2C ensures that the nutrient processed by the TF2-regulated pathway would be converted to the central metabolism (dark green area) by the downstream part of the TF1-regulated pathway. Note that in biosynthetic (anabolic) pathways the direction of metabolic flow is opposite to that in a nutrient-utilization (catabolic) pathways used in our illustrations (Fig. 2A-D). As a result, the direction of regulatory interactions between transcription factors should be reversed as well. Thus in biosynthetic pathways one expects more centrally-positioned regulator with larger out-degree to regulate its more peripheral (and less connected) counterparts as is known to be the case e.g. in the leucine biosynthetic pathway (see [20] and references therein). A problem with adding just a TF2 $\rightarrow$ TF1 regulation is that it results in some wasteful enzyme production. Indeed, presence of the red nutrient triggers the production of enzymes of the entire blue pathway including those upstream of the merging point with the red pathway. These upstream enzymes are not required and in the regulatory topology proposed in Fig. 2C they are actively suppressed by TF2. Another possibility shown in Fig. 2B and 2D is that instead of suppressing the upstream enzymes of the blue pathway to activate its downstream enzymes. In Fig. 2B transcription factors regulate the entire length of the path from a leaf (nutrient) all the way down to central metabolism. On the other hand, the model in Fig. 2D adds a dedicated transcription factor (TF3) activated by the TF2 to regulate only the downstream part of the blue pathway. In both these models (Fig. 2B and 2D) enzymes located close to the central core receive more than one regulatory input.

It has been proposed [18] that some of the ubiquitous feed-forward loops in bacterial regulatory networks serve as low-pass filters buffering against fluctuations in nutrient availability. Such loops could be easily incorporated in our models. One possibility would be to add regulatory interaction between TF2 and TF1 in Fig. 2B. For the model in Fig. 2D one might extend the range of TF2 to include at least part of the targets of TF3 and/or add a regulatory interaction between TF1 and TF2. Our simulations of models in Fig. 2B-D indicate that they all give rise to very long regulons. The distribution of regulon sizes of these models shown in Fig. S4 has a tail broader than that in *E. coli* according to the Regulon database [15]. A more detailed empirical study and modeling of coordination of activity of metabolic pathways goes beyond the scope of this study and will be addressed in our future research.

Prokaryotic genomes are shaped by horizontal gene transfer and prompt removal of redundant genes. The Horizontal Gene Transfer (HGT) of whole modules of functionally related genes from other organisms is the likely mechanism by which new pathways are added to the growing metabolic network in our model. Indeed, the rules of our model imply that an organism acquires a several enzymes necessary to utilize a new nutrient not piece by piece but all at once. Indeed, a pathway converting a nutrient to a downstream product disconnected from the rest of metabolic network does not contribute to biomass production and thus makes little sense from the standpoint of evolution. The dominant role of HGT and genome contractions in shaping contents of prokaryotic genomes and thus their metabolic networks is well documented [21]. For example, a recent empirical study [11] reports that the horizontal transfer of physiologically-coupled enzymes is the dominant mode of adaptive evolution of bacterial metabolic networks. Pal and collaborators [11] show that horizontally transferred genes

- Outnumber duplicated genes (at least during the last 100 million years in the evolution of *E. coli*).
- Frequently confer condition-specific advantages, facilitating adaptation to new environments.
- As a consequence, horizontally-transferred cassettes tend to include transporter genes responsible for nutrient uptake and to be located near the periphery of the metabolic network rather than at its core.
- Physiologically coupled enzymes are frequently transferred or lost together (see also [9] for a genome-wide analysis of this trend).

These observations make the central assumptions of our model even more plausible. Another feature of prokaryotic evolution used in our model is its tendency to promptly remove redundant genes. Indeed, in our model we implicitly assume that if a set of horizontally transferred genes contains some enzymes that are already encoded in the genome, these redundant copies are promptly removed from the genome. Stopping the added metabolic branch precisely at the intersection point with the existing metabolic network corresponds to instantaneous removal of redundant genes. Evolutionary justified exponential decay of the number of redundant genes does not change the scaling exponent. Both these features (massive horizontal gene transfers and prompt removal of redundant genes) are not characteristic of eukaryotic genomes in general, and those of multicellular organisms in particular. That is consistent with our finding of approximately linear scaling of N_{TF} with N_{genes} in genomes of animals shown in Fig. S5 (the best fit exponent 1.16 ± 0.2). The best fit exponent for all eukaryotic genomes (1.3 ± 0.2) [4] is marginally higher and much lower than its value in prokaryotes (~ 2).

Several earlier modeling efforts [4, 22, 23] explained the quadratic scaling in terms of gene duplications followed by divergence of resulting paralogs. Models of this type assume that additions and deletions of *individual genes* are determined by a stochastic process decoupled from their biological function. Conversely, our model is, to the best of our knowledge, the first attempt to explain this scaling relation in purely functional terms. Instead of single genes we add and delete larger functional units (metabolic pathways) and as-

sume that they are retained by evolution only if they connect to the entire metabolic network of an organism. Besides, when compared to earlier explanations our toolbox model uses completely different evolutionary mechanism (horizontal gene transfer vs gene duplications).

How quickly new pathways acquire transcriptional regulators?. In our model we assume that regulatory network closely follows changes in the metabolic toolbox of its organism and quickly adapts to changes in nutrient availability. For the sake of simplicity we choose to assign regulators *de novo* to each new state of the metabolic network. To verify that this simplification does not distort our final results we simulated a variant of our model in which the transcriptional regulatory network dynamically follows changes in the metabolic network. The regulon size distribution in this model was essentially unchained from it static counterpart.

The nearly immediate assignment of regulators to newly acquired pathways is supported by the empirical study of Price and collaborators [24], where it was reported that horizontally transferred peripheral metabolic pathways frequently include their own transcriptional regulators. This should come as no surprise, given many cases where metabolic enzymes and their regulators belong to the same operon (e.g. Lac operon). As proposed in the selfish operon theory [25] such genomic proximity between enzymes and their regulators is favored by evolution.

Even when the horizontally transferred pathway does not include a dedicated transcriptional regulator it could nevertheless be quickly acquired in a separate HGT event or created by gene duplication of another TF in the genome. Overall, the emergent picture [26] is that regulatory networks in prokaryotic genomes are flexible, quickly adaptable, and rapidly divergent even between closely related strains.

Materials and Methods

Numerical simulations of the model. Metabolic network in our model is shaped by randomly repeating pathway addition and pathway removal steps. The boundary conditions for this stochastic process do not allow the number of metabolites to fall below 40 or exceed about 1600. Networks with different values of N_{met} are then sampled and analyzed. The universal network used in our study consists of the union of reactions listed in the he network consists of all reactions listed in the KEGG database [8]. The directionality of reactions and connected pairs of metabolites are inferred from the map version of the reaction formula: `ftp://ftp.genome.jp/pub/kegg/ligand/reaction/reaction_mapformula.lst`. Since our goal is to model the conversion of nutrients to organism's biomass we kept the metabolites located upstream of the central metabolism (reachable by a directed path from Pyruvate). This left us with 1813 metabolites connected by 2745 edges. The exact size and topological structure of the universal network is not known. To test our model on a universal network of a different size (red squares in Fig. 4B) we pruned the KEGG network down to ~ 900 metabolites. This pruning was achieved by randomly removing nodes along with branches that got disconnected from the central metabolism. In yet another version shown in Supplementary Fig. 1S the universal network is made of random walks on the fully connected graph of $N_{univ} = 1800$ metabolites. From this figure it follows that properties of our model are mainly determined by the number of nodes in the universal network and not by details of its topology.

1) Pathway addition. One randomly chooses a new leaf (nutrient) and a self-avoiding random walk on the universal network. This directed walk is started at the leaf and extended until it first intersects the subset of N_{met} presently metabolizable molecules. The leaf plus all the intermediate metabolites of this new branch thereby become metabolizable.

2) Pathway deletion. One of the N_{TF} network leaves (nutrients) is chosen randomly. The links downstream from this leaf are followed until the first merging point of two metabolic branches. All the metabolites down to this merging point are removed from the network, thereby becoming non-metabolizable.

We typically choose to begin all simulations with 20 nodes in the "metabolic core" (the dark green central circle in Figs. 1-2) that are already metabolizable. This core could be thought of as the "universal central metabolism" present in most organisms. The number of these core metabolites, N_{core}, is the second parameter of our model. However, in practice, as long as $N_{core} \ll N_{univ}$, the network topological structure in the steady state does not depend on the value of N_{core}. In our simulations we also tried different starting sets of metabolizable molecules connected by linear branches to the core but inevitably arrived to the statistically identical steady-state networks.

Sources of empirical datasets. The distribution of branch lengths in Fig. 2A was calculated as follows: first a leaf was randomly chosen and followed to the metabolic core. Subsequent branches were followed until the merging point with another branch that was previously selected. In the metabolic network of the K-12 strain of *E. coli* leaves were defined as either 1) having zero in-degree (no production within the organism) or 2) having an undirected degree of one (endpoints of linear branches formed by reversible reactions). The backbone of the *E. coli* network was defined by following random linear paths starting at these leaves and ending at the intersection with each other or at the metabolic core. This left us with a tree shown in Fig. 3A consisting of ~ 420 metabolites (including 112 leaves) located upstream of the central metabolism [8]. .

To estimate the number of transcription factors in different genomes shown in Fig. 4A (green symbols) we used the DBD database [27] (`www.transcriptionfactor.org`) with its manually curated list of 147 Pfam families of transcription factors. The resulting N_{TF} are in good agreement with those obtained in earlier studies [3, 4, 5, 6].

Acknowledgments

Work at Brookhaven National Laboratory was carried out under Contract No. DE-AC02-98CH10886, Division of Material Science, U.S. Department of Energy. Work at Niels Bohr Institute was funded by the Danish National Research Foundation through the Center for Models of Life. SK and KS thank the Theory Institute for Strongly Correlated and Complex Systems at BNL for the hospitality and financial support during visits where some of this work was accomplished. We thank Eugene Koonin and Yuri Wolf for helpful discussions and critical comments on this manuscript.

1. Madan Babu M, Teichmann SA (2003) Evolution of transcription factors and the gene regulatory network in Escherichia coli. Nucleic Acids Res 31(4): 1234-1244.

2. Anantharaman V, Koonin E, Aravind L (2001) Regulatory potential, phyletic distribution and evolution of ancient, intracellular small-molecule-binding domains. Journal of Molecular Biology 307(5): 1271-1292.

3. Stover C, Pham X, Erwin A, Mizoguchi S, Warrener P, et al. (2000) Complete genome sequence of Pseudomonas aeruginosa PA01, an opportunistic pathogen. Nature 406(6799): 959-964.

4. van Nimwegen E (2003) Scaling laws in the functional content of genomes. Trends in Genetics 19(9): 479-484.

5. Cases I, de Lorenzo V, Ouzounis C (2003) Transcription regulation and environmental adaptation in bacteria. Trends in Microbiology 11(6): 248-253.

6. Konstantinidis K, Tiedje J (2004) Trends between gene content and genome size in prokaryotic species with larger genomes. Proceedings of the National Academy of Sciences 101(9): 3160-3165.

7. Thieffry D, Huerta A, Perez-Rueda E, Collado-Vides J (1998) From specific gene regulation to genomic networks: a global analysis of transcriptional regulation in Escherichia coli. Bioessays 20(5): 433-440.
8. Kanehisa M, Goto S (2000) KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Research 28(1): 27-30
9. Spirin V, Gelfand M, Mironov A, Mirny L (2006) A metabolic network in the evolutionary context: Multiscale structure and modularity. Proceedings of the National Academy of Sciences 103(23):8774-8779.
10. von Mering C, Huynen M, Jaeggi D, Schmidt S, Bork P, et al. (2003) STRING: a database of predicted functional associations between proteins. Nucleic Acids Research 31(1): 258-261.
11. Pal C, Papp B, Lercher M (2005) Adaptive evolution of bacterial metabolic networks by horizontal gene transfer. Nat Genet 37(12): 1372-1375.
12. Beiko R, Harlow T, Ragan M (2005) Highways of gene sharing in prokaryotes. Proceedings of the National Academy of Sciences 102(40): 14332-14337.
13. Savageau M (1977) Design of molecular control mechanisms and the demand for gene expression. Proceedings of the National Academy of Sciences of the United States of America 74(12): 5647.
14. Handorf T, Ebenhoh O (2007) MetaPath Online: a web server implementation of the network expansion algorithm. Nucleic Acids Research.
15. Salgado H, Gama-Castro S, Martnez-Antonio A, Daz-Peredo E, Snchez-Solano F, et al. (2004) RegulonDB (version 4.0): transcriptional regulation, operon organization and growth conditions in Escherichia coli K-12. Nucleic Acids Research 3: D303.
16. van Nimwegen E (2004) *in Power Laws, Scale-Free Networks and Genome Biology*, eds Koonin EV, Wolf YI, Karev GP (Landes Bioscience, Georgetown) pp. 236-261
17. Molina N, van Nimwegen E (2008) Universal patterns of purifying selection at noncoding positions in bacteria. Genome Research 18(1): 148.
18. Shen-Orr S, Milo R, Mangan S, Alon U (2002) Network motifs in the transcriptional regulation network of Escherichia coli. Nature Genetics 31(1): 64-68.
19. Cosentino L, Jona P, Bassetti B, Isambert H (2007) Hierarchy and feedback in the evolution of the Escherichia coli transcription network. Proc Natl Acad Sci US A 104(13): 5516-5520.
20. Chin C, Chubukov V, Jolly E, DeRisi J, Li H (2008) Dynamics and Design Principles of a Basic Regulatory Architecture Controlling Metabolic Pathways. PLoS Biology 6:e416.
21. Koonin EV, Wolf YI (2008) Genomics of bacteria and archaea: the emerging dynamic view of the prokaryotic world. Nucleic Acids Research, 236(21): 6688-6719
22. Foster D, Kauffman S, Socolar J (2006) Network growth models and genetic regulatory networks. Physical Review E 73(3): 31912.
23. Enemark J, Sneppen K (2007) Gene duplication models for directed networks with limits on growth. Journal of Statistical Mechanics, P11007.
24. Price M, Dehal P, Arkin A (2008) Horizontal gene transfer and the evolution of transcriptional regulation in Escherichia coli. Genome Biology 9(1): R4.
25. Lawrence J, Roth JR (1996) Selfish Operons: Horizontal Transfer May Drive the Evolution of Gene Clusters. Genetics 143(4): 1843-1860.
26. Gelfand M (2006) Evolution of transcriptional regulatory networks in microbial genomes. Current Opinion in Structural Biology 16(3):420-429.
27. Wilson D, Charoensawan V, Kummerfeld SK, Teichmann SA (2008) DBD–taxonomically broad transcription factor predictions: new content and functionality. Nucleic Acids Res 36:D88-92
28. Maslov S,Sneppen K (2004) *in Power Laws, Scale-Free Networks and Genome Biology*, eds Koonin EV, Wolf YI, Karev GP (Landes Bioscience, Georgetown) pp. 25-37.

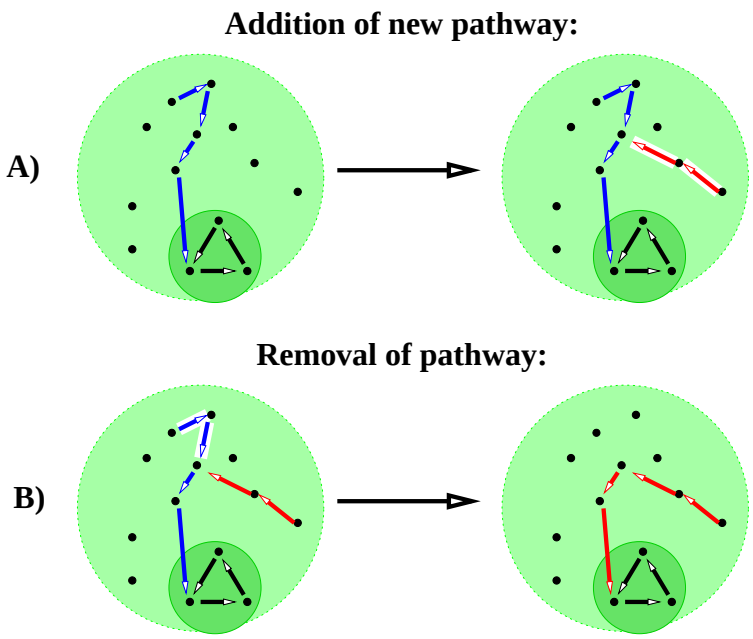

Fig. 1. "Toolbox" rules for evolving metabolic networks in our model. A) addition of a new metabolic pathway (red) connecting the red nutrient to a previously existing pathway (blue) which in turn connects it to the central metabolic core (dark green). B) removal of a pathway following the loss of the blue nutrient. The upstream part of the blue pathway that is no longer required is removed down to the point where it merges with another pathway (red).

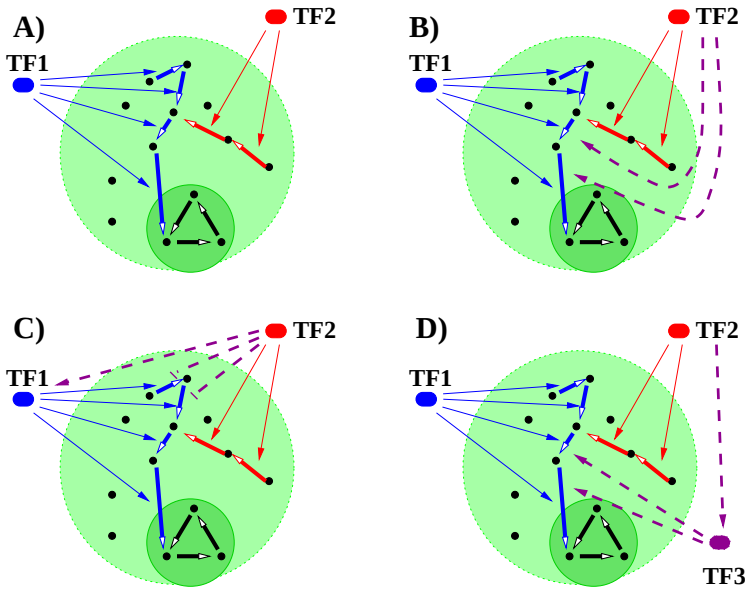

Fig. 2. Schematic diagram illustrating several scenarios of assigning transcription factors (TF) to regulate reactions in the metabolic network. Four panels correspond to four versions of our model discussed in the text. In the basic model (panel A) there is no coordination of activity between red and blue metabolic pathways. More realistic models (panels B-D) include extra regulatory interactions (purple dashed lines) and even extra transcription factors (purple TF3 in panel D) ensuring such coordination. In panels B-D when TF2 is activated by the red nutrient it turns on not only the enzymes in the red pathway but also the downstream part of the blue pathway necessary for red nutrient's utilization. Meanwhile, in the absence of the blue nutrient the part of the blue pathway upstream of the merging point of two pathways remains inactive.

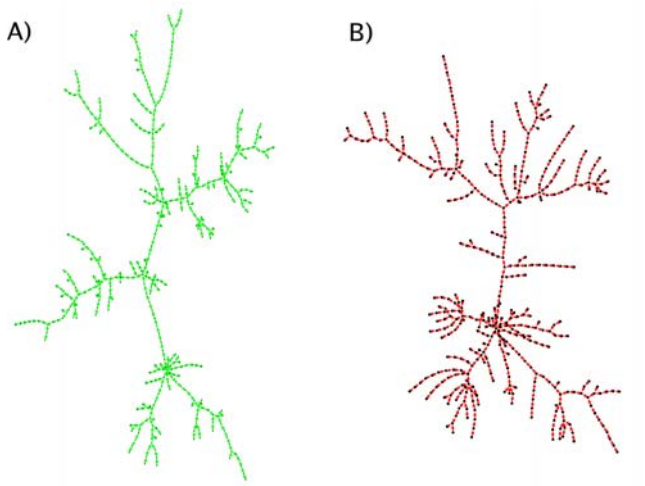

Fig. 3. A. The backbone of the metabolic network in *E. coli* [8] (green). B. An example of a similarly-sized network generated by our model (red).

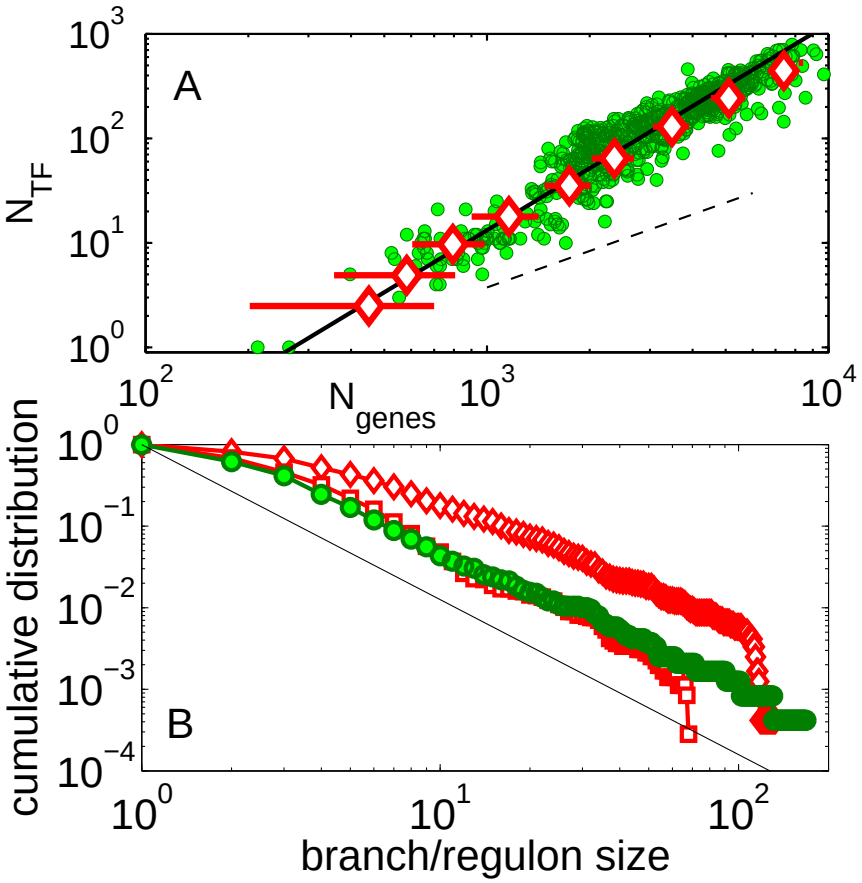

Fig. 4. A. The number of transcription factors scales quadratically with the total number of genes in our model (red) and real prokaryotic genomes (green) [8, 27]. Solid line with slope 2 ± 0.1 is the best powerlaw fit to the real data, while the dashed line with slope 1 is shown for comparison to emphasize faster-than-linear scaling. B. Cumulative distributions of pathway/branch lengths in *E. coli* metabolic network (green circles) and our model of comparable size (red symbols) have similar powerlaw tails. The best powerlaw fit $\gamma - 1 = -1.9$ (solid line) is consistent with our mean-field analytical result $\gamma = 3$ (see text for details). The toolbox model with $N_{met} = 400$ was simulated on universal networks of KEGG reactions with $N_{univ} = 1800$ (red diamonds) and $N_{univ} = 900$ (red squares) nodes.