1 Metabolic Networks : Structure

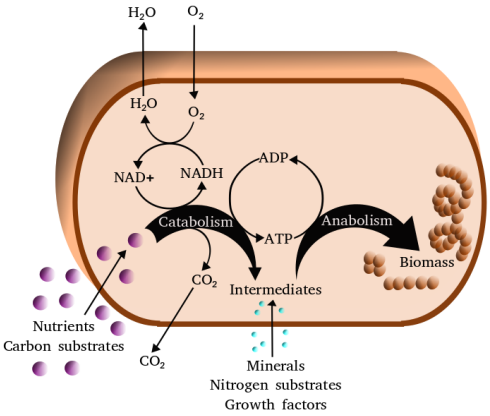
Metabolism is the set of cellular processes by which a cell

1. (**input**) acquires resources (molecular "nutrients") from its extracellular environment,
2. (**conversion**) transforms those resources intracellularly into other molecules, proteins, lipids, nucleic acids, carbohydrates, etc., and
3. (**output**) discards unneeded waste molecules back into the environment outside the cellular membrane.

Metabolic processes are often "enzymatic," meaning they are catalyzed by enzymes, a special kind of protein that lowers the activation energy for a particular molecular transformation, but is not itself "used up" in the process. Enzymatic reactions come in two flavors:

- **catabolic reactions** (catabolism), which break down compounds, e.g., breaking down glucose to pyruvate via cellular respiration, and typically release energy that can be stored in some way, e.g., as ATP, or
- **anabolic reactions** (anabolism), which build up new ("synthesis") or augment existing compounds, and typically consume energy from the cell.

The schematic below shows a simplified view of cellular metabolism.[1] Generally, metabolism defines what inputs an organism can "eat," which ones are "poisonous," and what waste it produces.



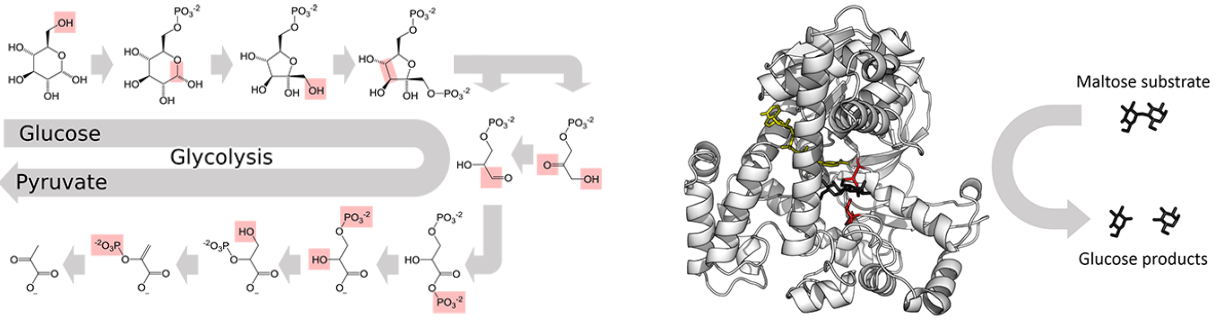
Of particular note within this picture are **ATP** and **NAD+**, which play the role of "currency metabolites," meaning they act like money inside the economy of the cell, because they are necessary components of a great many metabolic reactions. Catabolic reactions consume NAD+ and ADP (and produce NADH and ATP). Anabolic reactions consume ATP (and produce ADP).

[1] Source: `https://en.wikipedia.org/wiki/File:Metabolism.png`. This picture is simplified because it largely omits the output part of metabolism.

Enzymes and pathways

Crucially, metabolism is not a collection of passive chemical reactions. Rather, each cell actively manages ("regulates") which particular reactions occur, and at what rates, by controlling the concentration of associated **enzymes**. Enzymes are usually proteins, and are thus genetically encoded, meaning that much of metabolism is subject to biological evolution (see Lecture 10).

As an example, the *glycolysis* pathway[2] below (left) is one way a cell can convert glucose, a simple carbohydrate, into pyruvate, a common intermediate for many other metabolic reactions. In the catabolic pathway, notice the boxed portions of each molecule: each is a place where a *different enzyme* makes a structural modification to the molecule. *Glucosidase* (right) is one such enzyme, which takes as input a maltose molecule, breaks one bond, and outputs glucose products. Glucosidase is one of many structurally similar enzymes for catabolizing (breaking down) complex carbohydrates like starches and glycogens into their "monomers" (subunits).

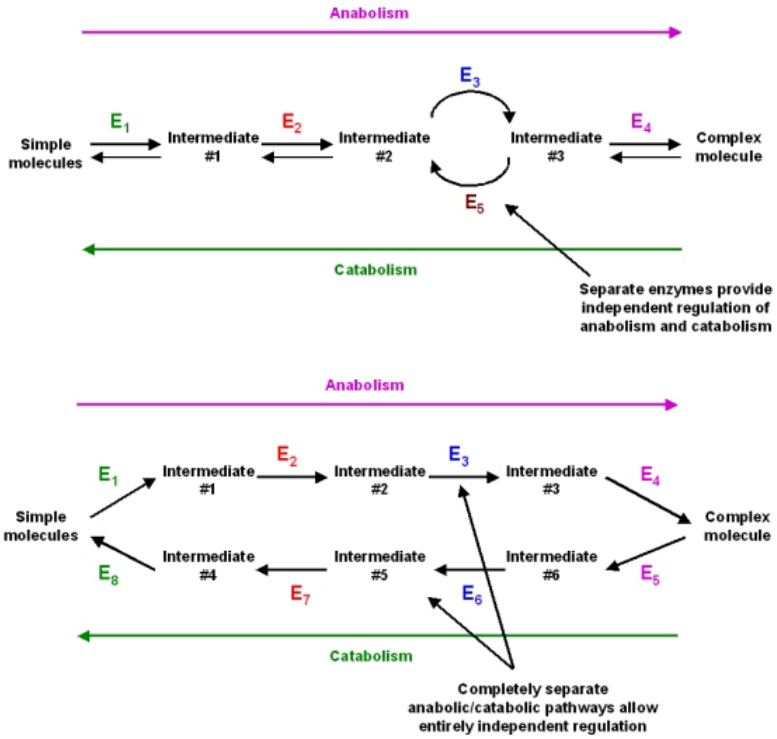


When enzymes act in a sequence on a molecule, they create a **pathway**, and much of how we think about metabolism is through the lens of pathways (much as it is for signaling, in protein interaction networks). As a rough analogy, think of a factory assembly line, in which raw materials are incrementally transformed into a finished product. At each step on the assembly line, a discrete transformation takes place, e.g., two parts are combined, an excess part is clipped off, etc. A metabolic pathway is very similar, except that every part of the process is molecules.

2 Metabolic networks

But these pathways are not independent, and instead they overlap, greatly. The union of pathways constitutes a *metabolic network*, composed of (1) its metabolic **enzymes** and (2) various **metabolites**, called substrates or products (reaction inputs and outputs). Just as there are catabolic and anabolic reactions, there are catabolic and anabolic pathways.

[2]Sources: `https://en.wikipedia.org/wiki/File:Glucosidase_enzyme.png` and `https://en.wikipedia.org/wiki/File:Glycolysis_metabolic_pathway.svg`



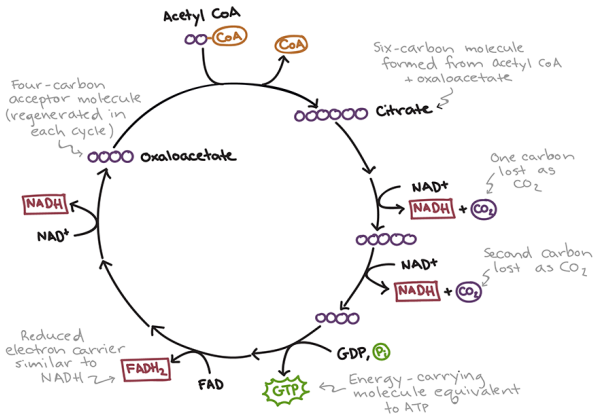
In these pathway examples,[3] each $\mathbf{E}_i$ denotes some metabolic enzyme, and the two pathways depict different ways we might organize the catabolic and anabolic directions that connect a simpler molecule (left side) to a more complex molecule (right side). These different structures provide different opportunities for regulation. For instance, in the lower version, the catabolic vs. anabolic directions can be managed independently, while in the upper version, they are less independent.

2.1 Core metabolism

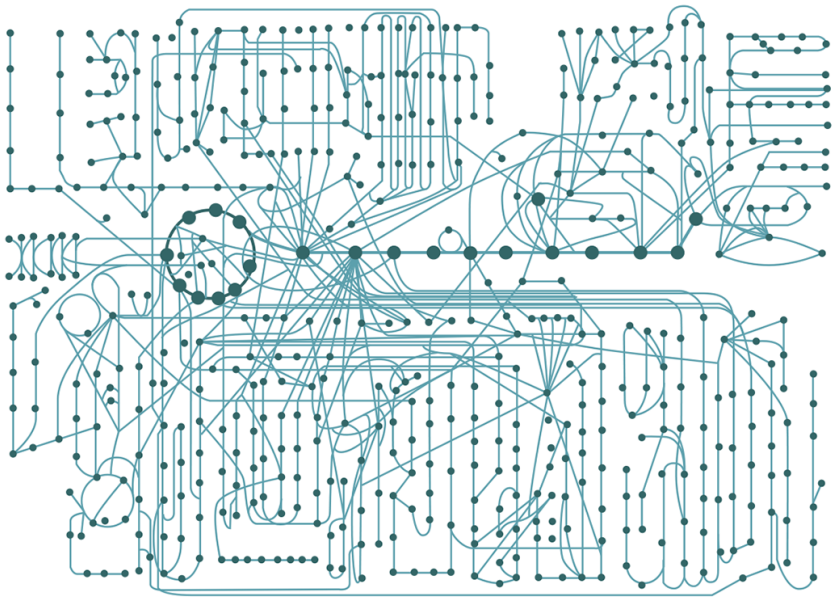
In aerobic organisms (those that use oxygen), what we call "core metabolism" is the set of reactions that surround the citric acid or Krebs cycle. (Anerobic organisms use a variety of other means to power their metabolisms.) The Krebs cycle is a core pathway that consumes NAD+ and ATP (producing NADH and ADP), and builds a core set of metabolic products that are consumed by other parts of metabolism. In a very real sense, the Krebs cycle is the "engine" of the aerobic cell. In eukaryotes, this cycle occurs within a cell's mitochondria. In prokaryotes, the cycle occurs in the cytosol (part of the liquid inside a cell), and it uses a proton gradient across its cell surface for ATP production (since it lacks the internal membrane that mitochondria provide).

[3] Source: `http://mikeblaber.org/oldwine/BCH4053/Lecture32/Lecture32.htm`

Below is a simplified view of core metabolism,[4] showing only the citric acid cycle itself:



And below is a more network-type view,[5] in which we "zoom out" at little to include some of the reactions that can feed into or out of the Krebs cycle (can you spot it?). Note that the citric acid cycle is not the only cycle in this metabolic network!



[4]Source: `https://bit.ly/3c07UTm` (`www.khanacademy.org`)
[5]Source: `https://commons.wikimedia.org/wiki/File:Metabolism_diagram.svg`

2.2 What are the nodes, and what are the edges?

What interacts with what? Metabolism is complicated, because biochemistry can be complicated. As a result, there are several different ways to represent metabolism as a network.

The most basic is a **directed bipartite** network, in which nodes are either enzymes or metabolites. A pair of directed edges $(i \to j)$ and $(j \to k)$ exist if metabolite i is a substrate (input) to a reaction catalyzed by enzyme j, while metabolite k is a product (output). But, this representation leaves something out: if an enzyme i participates in multiple metabolic reactions, each with non-identical sets of metabolites X and Y (such that $X \neq Y$), then we will not be able to tell the sets of its interacting metabolites apart. Hence, we will not be able to tell how large the input sets are to the different reactions.

From the directed bipartite representation, there are several ways to derive simpler, metabolite- or enzyme-only representations (projections) of metabolism. Two of these are particularly useful.[6] These are illustrated for two overlapping reactions, shown in row *(a)* below:

- a **substrate-product network**, row *(b)*: every substrate links to every reaction product, or
- a **substance network**, row *(d)*: edges connect all substances participating in a reaction.

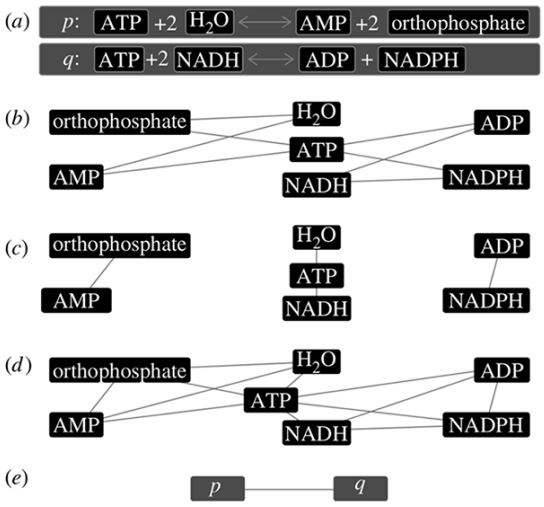


Figure 1. An illustration of different network representations derived from the two hypothetical reactions shown in (*a*). (*b*) Substrate–product network, (*c*) a substrate–substrate network, (*d*) a substance network (including both the substrate–product and substrate–substrate type edges) and (*e*) a reaction network where the vertices are reactions connected if they have a substance in common.

In the *substrate-product network*, row *(b)*, each reaction becomes a complete bipartite subgraph, in which every substrate is linked to every product (and the enzyme is omitted entirely), but there are

[6] From Holme, *J. R. Soc. Interface* **6** (2009), `dx.doi.org/10.1098/rsif.2008.0489`

no connections among substrates or among products. In this representation, interactions represent the transformation from input to output.

In contrast, in the *substance network*, row *(d)*, each reaction becomes a complete subgraph, in which every substrate or product is linked to every other substrate or product. The substance network is thus a one-mode projection of the enzyme-metabolite bipartite network, and interactions represent the participation in a common transformation (which is a subtle distinction from the substrate-product network).

Regardless of which simple graph representation of metabolism we choose, some information is lost because edges are not entirely independent and because we have represented each enzyme as a group of correlated edges. Calculations that treat edges as more or less independent may thus be skewed by the presence of these artifacts.

2.3 Where do the data come from?

The main source of metabolic network data is from DNA sequencing, combined with curated lists of known metabolic enzymes, and the corresponding reactions they catalyze. Databases like KEGG (Kyoto Encyclopedia of Genes and Genomes) and BiGG Models (`http://bigg.ucsd.edu`) contain such information, which allows us to extract, for a given organism, a set of identifiable metabolic reactions that organism contains.

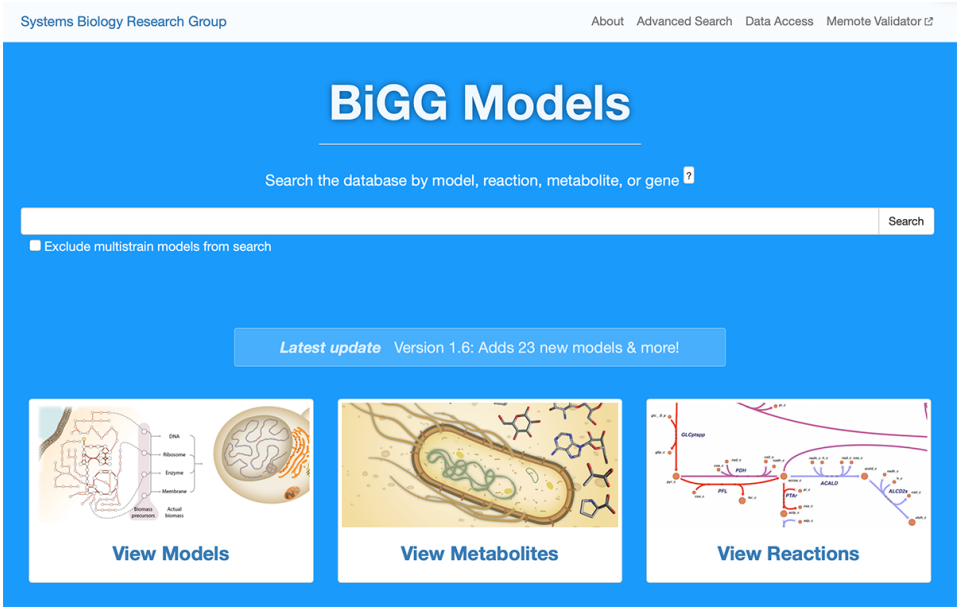
Much of the list of known reactions comes from years of study of model organisms like the bacteria *E. coli*, the nematode *C. elegans*, the fly *D. melanogaster*, and humans. This focus on model organisms means that the metabolic network we extract from a database will be incomplete, representing only the intersection of the novel genome's proteins and the set of known model organism enzymes.

The *false positive rate* of this approach is low (if something known is there, we will see it), but the *false negative rate* is unknown: we won't know what enzymes we missed, because we can only find things we already know about. In addition to the reactions themselves, we can also sometimes obtain node metadata, e.g., on

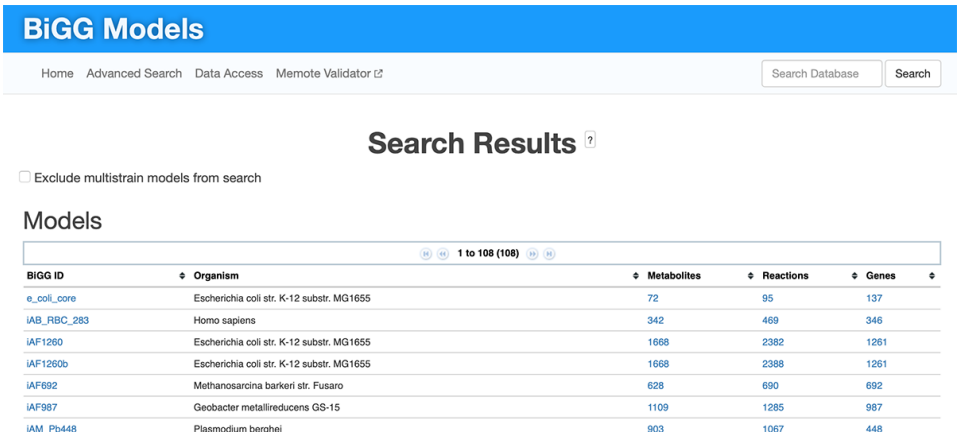
- what metabolic pathway a reaction is in,
- where in the cell the reaction takes place (e.g., mitochondria, cytoplasm, cell membrane, etc.),
- what cellular function the reaction contributes to (e.g., amino-acid metabolism, carbohydrate metabolism, lipid metabolism, nucleotide metabolism, metabolism of co-factors and vitamins, and transport), or
- what kind of energetics the reaction requires (e.g., a rate constant, or a flux).

2.3.1 BiGG Models

The BiGG Models database provides a wealth of information, drawn from many model organisms and many studies.



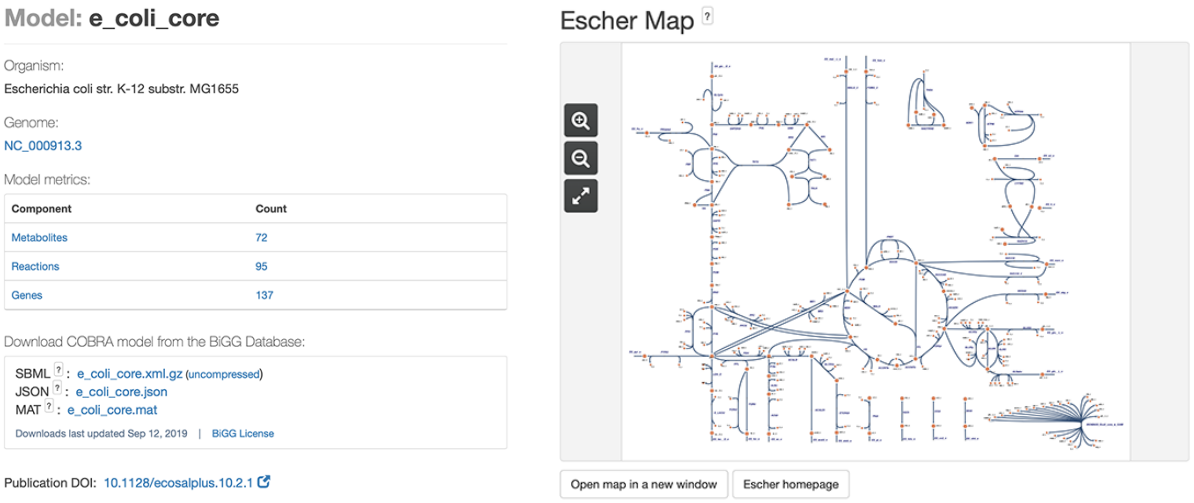
Under the "View Models" link, we can obtain a list of available networks (models):



BiGG ID	Organism	Metabolites	Reactions	Genes
e_coli_core	Escherichia coli str. K-12 substr. MG1655	72	95	137
iAB_RBC_283	Homo sapiens	342	469	346
iAF1260	Escherichia coli str. K-12 substr. MG1655	1668	2382	1261
iAF1260b	Escherichia coli str. K-12 substr. MG1655	1668	2388	1261
iAF692	Methanosarcina barkeri str. Fusaro	628	690	692
iAF987	Geobacter metallireducens GS-15	1109	1285	987
iAM_Pb448	Plasmodium berghei	903	1067	448

The first item in this list (*E. coli*'s core metabolism) reveals a bit more about this data set, including where it came from, and the underlying files. In this case (not common), BiGG also shows a nice

visualization of the metabolic network itself, in which you can clearly see the citric acid cycle, along with some of its major input and output pathways.



2.4 What do people do with metabolic networks?

1. **Systems biology**

   - What can the architecture of one or more organisms' entire metabolic networks tell us about how cells work, and how they evolved?
   - Are there general principles that structure it? Often people start with general network analysis tools, e.g., looking for modules, in order to understand a coarse-grained view of metabolic organization

2. **Flux balance analysis**

   - Every reaction in a metabolic network represents a flow of inputs to outputs, and the inputs are usually outputs of other reactions, and the outputs are inputs to still further reactions. Hence, we can model how mass (atoms) flow across the network. This is the domain of a technique called flux balance analysis (FBA), so called because the conservation of mass implies that the fluxes have to balance across the reactions, except for the food coming in, and the waste going out.
   - If we modify a metabolic network, e.g., to increase or decrease a flow here, or there, what does that impact have on the total flux across the network? can we optimize the production of certain kinds of products by modifying the network flows?
   - Can we add and optimize novel pathways, e.g., synthesizing gasoline directly from O2, CO2, and H2O, without disrupting the rest of the metabolic network?

3 How do metabolic networks get their structure?

Answering this question is still a subject of active research among scientists. One reasonably well supported explanation comes from the **toolbox model** of metabolic network evolution, which uses the evolutionary addition and deletion of metabolic pathways to maintain a "functioning" metabolic network over time.[7]

The idea of the toolbox model is that a metabolic network is composed of a repertoire (a set) of enzymes that together construct a set of overlapping pathways. On the catabolic side of metabolism, these pathways take in **nutrient molecules** and progressively decompose them into basic building blocks that are fed into core metabolism. An organism's repertoire of enzymes represent a set of "tools" by which the organism digests its nutrients. The toolbox model is only a model of catabolic pathways, although in principle, it could also be used to model the anabolic pathways that construct more complex molecules.

In the model, there are two kinds of structural changes: pathway addition (**A** in the figure below) and pathway removal (**B**). When an organism's lineage evolves to solve a novel problem, i.e., to digest a new nutrient, it adds to its repertoire only the new "tools" (enzymes) it needs, and it reuses those it already has. Hence, it only adds a pathway that connects the new nutrient to its existing network.[8] Similarly, when a lineage evolves to lose its ability to digest a nutrient, it loses from its repertoire (network) only the "tools" (enzymes, meaning edges) that it no longer needs, and it retains all others. Hence, it only deletes edges that are unique to digesting the lost nutrient.

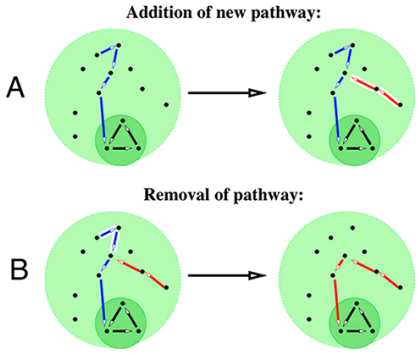


Fig. 1. "Toolbox" rules for evolving metabolic networks in our model. (*A*) Addition of a new metabolic pathway (red) that is long enough to connect the red nutrient to a previously existing pathway (blue) that further converts it to the central metabolic core (dark green). (*B*) Removal of a part of the blue pathway after loss of the blue nutrient. The upstream portion of the blue pathway that is no longer required is removed down to the point where it merges with another pathway (red). The light green circle denotes all metabolites in the universal biochemistry network from which new pathways are drawn.

[7] Introduced in Maslov et al., "Toolbox model of evolution of prokaryotic metabolic networks and their regulation." *Proc. Natl. Acad. Sci. USA* **106**, 9743-9748 (2009).

[8] We assume a separation of timescales here, meaning pathway additions and deletions occur and equilibrate rapidly in evolutionary time relative to other changes in the metabolic network's structure. Hence, the model can proceed via steps of either addition or deletion.

3.1 The universal biochemistry network

Conceptually, the toolbox model assumes the existence of a fixed network of all possible metabolic reactions. We can call this set of reactions a universal biochemistry network, and each species contains a functional subset of these reactions.[9] The citric acid cycle (aka, the Krebs cycle) is a part of this universal network, as is every other metabolic reaction found in any organism that has ever lived on Earth, *or ever could live*. An organism that contained the entire universal biochemistry network would be able to digest every nutrient known to be metabolizable by any organism.

The following pseudocode formalizes the verbal model given above for how the toolbox model works.

1. Define a **universal biochemistry network** G_u, which represents the set of all possible metabolic reactions.
2. Initialize a metabolic network G, composed only of core metabolism, which we might say represents the origin of life. Mark all edges in core metabolism so that we never delete them.
3. **A: Pathway addition**
   (a) uniformly at random, choose a new "nutrient" v, defined as leaf node of G_u that is not in G
   (b) take a "self-avoiding random walk" (no loops) on the edges of G_u that reaches from v to any node in G
   (c) call that set of edges σ, and add them all to G
4. **B: Pathway deletion**
   (a) uniformly at random, choose an existing "nutrient" v from G
   (b) starting at v, recursively delete the edges in a path toward core metabolism until we reach a node with in-degree $k_{\text{in}} > 1$, and then stop.

We could augment the model with additional assumptions, e.g., adding transcription factors (TFs) to the system to reflect the way cells need to be able to regulate the independent inputs to the network. Such a model is more realistic because cells turn on different parts of the metabolic network only when they need to use those pathways to digest nutrients, and otherwise typically turn them off to save energy. To incorporate this possibility, each "branch" of these input trees requires its own transcription factor.

Running the basic model forward in time, in which we choose to add a new or delete an existing pathway with equal probability, we can evolve a synthetic metabolic network. The toolbox model of metabolism has several nice properties:

- The model is pathway based, which reflects our understanding of metabolism being composed of sequences of enzymes operating on substrates.

[9] We say a "functional subset," because the vast majority of subsets of edges of this universal network would not correspond to a living organism.

- The model embodies the "use it or lose it" energetic selection principle that biologists believe shape the prokaryotic genome, i.e., genes that do not currently contribute to an organism's fitness will tend to be deleted from the genome.
- The model leverages known metabolic reactions (although any empirical estimate of G_u is going to be incomplete, to an unknown extent).

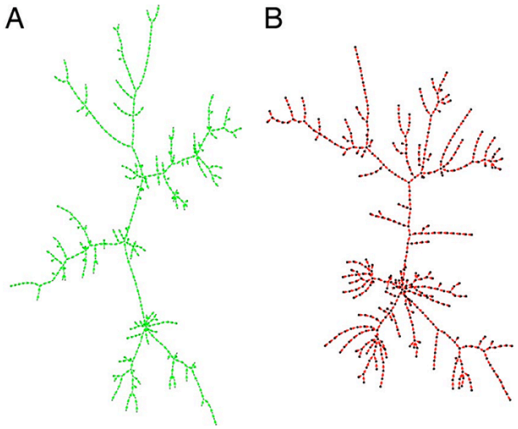


Fig. 3. Visual comparison of a real-life metabolic network with that generated by our model. (*A*) The backbone of the metabolic network in *E. coli* (8) located upstream of the central metabolism (green). (*B*) A similarly sized model network (red). Note the hierarchy of branch lengths in both images in which shorter pathways tend to be attached to progressively longer pathways. The branch length distributions in real and model networks are shown as green circles and red squares in Fig. 4*B*.

The figure above shows an example of one these simulated metabolic networks. On the left is a subset of reactions from *E. coli*, after removing certain edges (via a process called "network backboning," which basically extracts a "skeleton" (tree) of structurally important edges from a network), and a similarly-sized model network on the right. More generally, the toolbox model successfully produces synthetic metabolic network that have similar statistical properties as real-world metabolic networks, suggesting that it is, at least, on the right track.

It is also, however, very simplified, and it lacks any notion of flux, rates, or dynamics. That is, the toolbox model is a purely structural model of metabolic networks. Even within this structural setting, there are many ways we could vary the behavior of this model:

- vary the rate of pathway addition vs. deletion,
- vary the distribution of choices for novel nutrients to add (e.g., based on what's available in the environment, which may fluctuate),
- vary the "random" walk from new nutrient to existing metabolism (e.g., based on biochemistry and protein evolution), and
- vary the distribution of existing nutrients to delete (e.g., based on what's available in the environment, which may fluctuate).

Supplemental readings

1. Zhao & Holme, "Three faces of metabolites: Pathways, localizations and network positions." *Lecture Notes in Operations Research* **13**, 13-21 (2010). `http://www.aporc.org/LNOR/13/ISB2010F05.pdf`

2. Stelling et al., "Metabolic network structure determines key aspects of functionality and regulation." *Nature* **420**, 190-193 (2002) `https://www.ncbi.nlm.nih.gov/pubmed/12432396`

3. King et al., "BiGG models: A platform for integrating, standardizing and sharing genome-scale models." *Nucleic Acids Research* (2015) `https://academic.oup.com/nar/article/44/D1/D515/2502593`

4. Maslov et al., "Toolbox model of evolution of prokaryotic metabolic networks and their regulation." *Proc. Natl. Acad. Sci. USA* **106**, 9743-9748 (2009). `https://www.pnas.org/content/106/24/9743`