

# Large-Scale Reconstruction and Analysis of Growth Environments

Elhanan Borenstein<sup>1,2,\*</sup>, Martin Kupiec<sup>3</sup>, Marcus Feldman<sup>1</sup>, Eytan Ruppin<sup>4</sup>

1. Department of Biological Sciences, Stanford University, Stanford, USA

2. Santa Fe Institute, Santa Fe, USA

3. Department of Molecular Microbiology and Biotechnology,  
Tel Aviv University, Tel Aviv, Israel

4. School of Computer Science and School of Medicine,  
Tel Aviv University, Tel Aviv, Israel

\*E-mail: ebo@stanford.edu

In recent years, a large body of work has focused on the analysis of metabolic networks, in an attempt to uncover design principles that govern their structure and to study their function and dynamics. However, metabolic networks function within the context of biochemical environments and interact with these environments through uptake and secretion of various compounds. Their topology may therefore reflect these interactions, and its analysis can provide important insights not only into the metabolic capacity of the species under study, but also into the growth environments in which these species evolved.

This work presents the first computational large-scale reconstruction of growth environments, in which a large array of current species have evolved. To this end, we construct the metabolic networks of 478 species using data from a large-scale metabolic reaction database. The metabolic network of each species is represented as a directed graph, where nodes represent compounds (those appearing in the network are termed *occurring* compounds) and edges represent reactions. We use a graph-theory based algorithm (Figure 1A) to identify within each network a minimal subset of compounds that cannot be synthesized from other compounds in the network, and whose existence permits the production of all other compounds. These *seed* compounds represent the compounds that each species extracts from the environment, and are shown, remarkably, to serve as a good proxy for the characteristic habitat of each species (Figure 1B).

Analyzing the seed compounds characterizing each species, we find that the presence and absence pattern of many key compounds in the seed sets of the various species can be matched to major adaptation events when certain clades began or ceased to utilize novel compounds as seeds. Representative examples include, for example, phenylalanine (an essential amino acid) that appears only in the seed set of animals. Conversely, glutamate (a non-essential amino acid) is found only in the seed set of obligate intracellular parasites (*Chlamydiae* and *Buchnera*) that rely on their host for the exogenous provision of this amino acid and have lost the ability to produce it (full data not shown due to space limitations).

Turning to validate the seed sets identification on a large scale, we use the seed set composition of the various species to reconstruct a phylogenetic tree of life. The resulting tree successfully partitions the various taxonomic groups (Figure 2A). Remarkably, we find that the resulting tree, although based on seed content alone (forming, on average, only 10% of the metabolites in the network), is as accurate (measured by its distance to a reference sequence-based tree) as a tree based on the entire set of occurring compounds. This finding attests to the validity of the seed set as a fundamental characteristic of each species and its

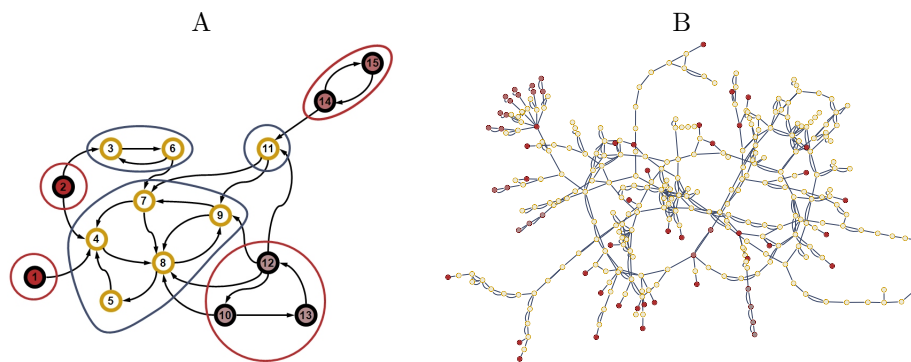


Figure 1: **(A)** The procedure for identifying seed compounds is illustrated in a simple synthetic network. The network is decomposed into its Strongly Connected Components (SCC) using Kosaraju’s algorithm. The seed set includes one compound from each *source* component (i.e., components without incoming edges and at least one outgoing edge; highlight in red). As all the compounds in a source component are equally likely to be included in the seed set, we assign to each of these compounds a confidence level,  $C = 1/(\text{component size})$ , denoted by the color saturation of the nodes. **(B)** The metabolic network of the obligate symbiont *Buchnera aphidicola* (with the seed compounds highlighted) whose habitat is well-characterized. The seed set correlates nicely with the organism’s environment; *Buchnera* has lost many biosynthetic genes and demonstrates an extremely successful symbiosis with its aphid host. It relies on the host for nutrients it cannot synthesize, such as non-essential amino acids, but also provides the host with essential amino acids that the host cannot synthesize. It has retained substrate-specific transporters only for glucose and mannitol, and is responsible for sulphate assimilation (a capability not possessed by its host). The composition of the *Buchnera* seed set obtained by our analysis is in clear agreement with the above findings: The seed set contains the most abundant non-essential amino acids, Glutamate and Glutamine, and is devoid of all the essential amino acids. It includes glucose and mannitol as the only carbon sources, as well as sulfate.

evolutionary history.

Next, we compile a detailed evolutionary history of each occurring and seed compound in our array of networks. We compute for each compound the first time in evolution in which it occurred in a metabolic network using Fitch’s parsimony algorithm. Comparing these first occurrence times with the propensity of a compound to be a seed, we find that most of the core compounds that can be extracted from the environment and utilized have been integrated into living organisms’ metabolism relatively early during life on earth (Figure 2B). Moreover, our results imply evolutionary dynamics wherein compounds are initially integrated into the metabolic network as extracted seeds, and only later lose their seed role due to the integration of new reactions into the network which enable their production (Figure 2C). Such dynamics are likely to be the outcome of adaptation to changing environments. The successful ‘reverse-engineering’ reconstruction of environments from currently existing metabolic networks and the phylogenetic reconstruction of ancient environments demonstrated here can serve to further study such adaptation dynamics.

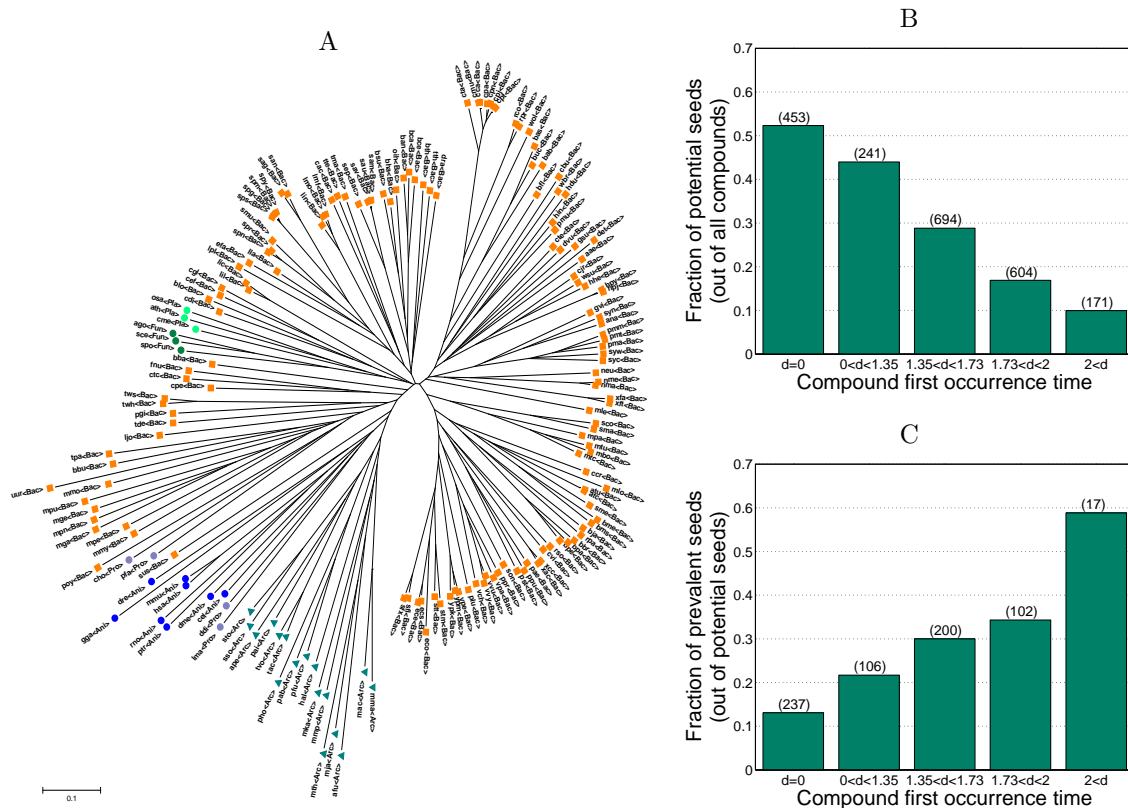


Figure 2: **(A)** Phylogenetic tree based on seed compounds content. <Bac>: Bacteria (orange squares), <Arc>: Archaea (cyan triangles), <Pla>: Plants (light green circles), <Ani>: Animals (blue circles), <Fun>: Fungi (dark green circles), <Pro> Protists (purple circles). **(B)** The fraction of potential seeds out of all occurring compounds as a function of the evolutionary time at which the compound first occurred,  $d$ . Potential seeds are defined as compounds that appear in the seed set of at least one species. The first occurrence time,  $d$ , is measured from the time of divergence between the Bacteria and the Archaea-Eukarya line and is presented in substitutions/site units. In this tree, 1.35 corresponds to a time point shortly after the divergence between Archaea and Eukarya, 1.73 corresponds to a time point shortly after the divergence of the Fungi group from the Metazoa, and 2 corresponds to a time point shortly after the divergence of mammals. The numbers in parentheses above each bar denote the number of compounds in this bin. Evidently, compounds that occurred in metabolic networks early in evolution are more likely to be potential seeds. **(C)** The fraction of prevalent seeds out of all potential seeds as a function of the evolutionary time at which the compound first occurred,  $d$ . Prevalent seeds are defined as compounds that appear in the seed sets of most of the species in which they occur. Apparently, the later in evolution a potential seed appears, the more likely it is to be prevalent. Accordingly, many of the observed prevalent seeds are those that have been integrated into the metabolic network relatively late in evolution and have not had enough time to lose their seed role.