Large-scale reconstruction and phylogenetic analysis of metabolic environments

Elhanan Borenstein*†‡, Martin Kupiec§, Marcus W. Feldman*, and Eytan Ruppin¶

*Department of Biological Sciences, Stanford University, Stanford, CA 94305; †Santa Fe Institute, Santa Fe, NM 87501; and §Department of Molecular Microbiology and Biotechnology and ¶School of Computer Science and School of Medicine, Tel Aviv University, Tel Aviv 69978, Israel

Edited by Nancy A. Moran, University of Arizona, Tucson, AZ, and approved July 22, 2008 (received for review June 26, 2008)

The topology of metabolic networks may provide important insights not only into the metabolic capacity of species, but also into the habitats in which they evolved. Here we introduce the concept of a metabolic network’s “seed set”—the set of compounds that, based on the network topology, are exogenously acquired—and provide a methodological framework to computationally infer the seed set of a given network. Such seed sets form ecological “interfaces” between metabolic networks and their surroundings, approximating the effective biochemical environment of each species. Analyzing the metabolic networks of 478 species and identifying the seed set of each species, we present a comprehensive large-scale reconstruction of such predicted metabolic environments. The seed sets’ composition significantly correlates with several basic properties characterizing the species’ environments and agrees with biological observations concerning major adaptations. Species whose environments are highly predictable (e.g., obligate parasites) tend to have smaller seed sets than species living in variable environments. Phylogenetic analysis of the seed sets reveals the complex dynamics governing gain and loss of seeds across the phylogenetic tree and the process of transition between seed and non-seed compounds. Our findings suggest that the seed state is transient and that seeds tend either to be dropped completely from the network or to become non-seed compounds relatively fast. The seed sets also permit a successful reconstruction of a phylogenetic tree of life. The “reverse ecology” approach presented lays the foundations for studying the evolutionary interplay between organisms and their habitats on a large scale.

Num erous biological systems can be represented as networks, encapsulating many of their essential properties (1). The structure and topology of these biological networks are not merely abstract descriptions of the complex interactions in a given system, but are also major determinants of the system’s function and dynamics. In particular, a wide range of analytical approaches has been used to study topological characteristics of metabolic networks and their bearings on various metabolic functional properties, including scaling (2), regulation (3), universality (4), and robustness to metabolic gene knockouts (5, 6). Furthermore, as metabolic networks function within the context of biochemical environments and interact with these environments by taking up or secreting various organic and inorganic compounds, previous studies have also addressed the effect that these environmental interactions have on the metabolic process, as manifested in, for example, the distribution of metabolic fluxes within the network (7) or the organism’s growth rate (8).

However, as the interactions with the environment must themselves be reflected in the structure of the evolved metabolic networks, these networks can be used not only to infer metabolic function but also to obtain insights into the growth environments in which the species evolved. Specifically, by analyzing the topology of a given metabolic network, we show that the set of compounds that are acquired exogenously (termed “seed set”; see also refs. 9 and 10) can be identified. Assuming that the environment of a species determines the metabolites it extracts from its surroundings to a considerable extent, the seed set can serve as a good proxy for its environment. This “reverse ecology” approach thus goes beyond previous research on the evolution of metabolic networks (11, 12) and metabolic scope analysis (9, 10, 13, 14) in enabling the evolutionary history of both metabolic networks and metabolic growth environments to be traced.

In view of this approach, in this article we first introduce the concept of metabolic networks’ seed sets and provide a formal methodology to computationally infer the seed set of a given network. We next integrate this methodology with large-scale metabolic data to compile a comprehensive large-scale dataset describing the seed sets of hundreds of species. The predicted seed sets are shown to accord with biological observations across compounds and across species, validating the potential and relevance of our computational framework and compiled dataset. This dataset is then analyzed to obtain novel insights into the evolutionary dynamics of metabolic networks and the determinants that affect their interfaces with the environment.

Results

We represent the metabolic network of a given species as a directed graph whose nodes represent compounds and whose edges represent reactions linking substrates to products [Materials and Methods and supporting information (SI) Materials and Methods]. This graph-based representation of metabolic reactions is a common tool in analyzing and studying metabolic networks (1, 2) and can be obtained from large-scale, cross-species databases [e.g., KEGG (15)]. It should be noted, however, that such directed graphs are simplifications of the actual underlying metabolic networks, ignoring, for example, reaction stoichiometry (see Discussion). Compounds that appear in the network are referred to as occurring compounds. Formally, we define the seed set of the network (9, 10) as the minimal subset of the occurring compounds that cannot be synthesized from other compounds in the network (and hence are exogenously acquired) and whose existence permits the production of all other compounds in the network (Fig. L4).

Our definition of seed compounds differs from that of essential compounds in that we require the production of all compounds in the network (and the potential activation of all of the metabolic pathways), regardless of their actual dynamic activation in a given environmental condition. In practice, organisms can survive in a wide range of environmental conditions and in each environment may activate only a subset of the pathways in the network, using a different set of exogenously acquired compounds (7, 16). Accordingly, the seed set can be conceived as the union of the essential sets required in all of these environments. Assuming that various

Author contributions: E.B., M.K., M.W.F., and E.R. designed research; E.B. performed research; E.B. analyzed data; and E.B., M.K., M.W.F., and E.R. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

*To whom correspondence should be addressed. E-mail: ebo@stanford.edu.

This article contains supporting information online at www.pnas.org/cgi/content/full/0806162105/DCSupplemental.

© 2008 by The National Academy of Sciences of the USA
Finally, it lacks all of the TCA cycle genes except for those coding transporters only for glucose and mannitol (21) and is responsible for the 2-oxoglutarate dehydrogenase complex (21). The composition of the B. aphidicola seed set obtained by our analysis (Fig. 1C and Table S2) is in clear agreement with the above observations. It contains the most abundant nonessential amino acids for aphids, glutamate and glutamine [which Buchnera uses as a substrate for the synthesis of other, essential amino acids (23)], and is devoid of all of the host essential amino acids. The seed set also includes glucose and mannitol (as the only carbon sources), 2-oxoglutarate, and sulfate, as well as thiamine (vitamin B1) and spermine (an essential growth factor).

To further confirm that the composition of the seed sets obtained by our analysis is consistent with known large-scale biological observations concerning the compounds that various species extract from their environments, we consider several key compounds and examine their presence and absence pattern in the occurring compound sets (phylectic occurrence pattern) and seed sets (phylectic seed pattern) across all of the species in our analysis (SI Materials and Methods and Fig. S2). For example, while many species can synthesize all of the amino acids they require, animals have lost their ability to make some amino acids (referred to as essential amino acids) and acquire them through their diet. Conversely, some obligate intracellular parasites have lost the ability to produce the nonessential amino acids and rely on their host for exogenous provision of these amino acids (21, 25). Comparing the resulting phylectic pattern for phenylalanine (an essential amino acid) with that obtained for glutamate (a nonessential amino acid), we find these patterns to be in complete accordance with the above observations (Fig. S3). Another example is biotin (vitamin B7), an essential cofactor for carboxylation reactions. Of the 42 species that were reported in a recent comparative genomics study (26) to synthesize biotin (and hence do not require biotin uptake from the environment), indeed, 40 have biotin as an occurring compound but not as a seed compound. Of the 24 species that were reported as lacking this capacity, 20 do have biotin as a seed. Interestingly, the four species that seem to lack the capacity to synthesize biotin and do not have biotin in their seed sets all have as a seed the same biotin biosynthetic precursor, dethiobiotin, which has been shown to allow various biotin auxotrophic bacteria to grow in the absence of biotin (27). Additional examples and details are presented in SI Text and SI Materials and Methods. Identified seeds are also correlated with the overall, static metabolic “interface” (or metabolic “potential”; see also ref. 18) of each organism (and may serve as a characterizing proxy for its effective biochemical habitat).

We developed a graph-based algorithm to detect the seed set of a given network (see Materials and Methods and Fig. 1B for more details). This algorithm is based on a fast method for strongly connected components (SCC) decomposition (19) and can therefore easily scale up and be applied to large-scale network data. Next we constructed the metabolic networks of 478 species (Table S1) using data from a large-scale metabolic reactions database (Materials and Methods) and applied the seed set detection algorithm to identify the seed compounds in each network (Dataset S1). This compilation results in a comprehensive large-scale dataset of predicted biochemical environments of hundreds of species and facilitates a cross-species comparison of such seed sets.

Clearly, large-scale metabolic data are usually based on genome annotation, largely from automated, comparison-based methods (15), and as such are bound to be incomplete and inaccurate (20). This can potentially have a marked effect on the composition of the inferred seed sets. However, examining the effect of missing or erroneous data (SI Text and Fig. S1), we find that the identified seed sets are fairly robust to perturbations of the raw metabolic data. Still, considering the inherent noise and incompleteness associated with these data, we focus here mostly on the identification of significant large-scale statistical signals and phylogenetic patterns characterizing the seed set composition across the tree of life.

To exemplify the composition of a seed set obtained by our analysis and its alignment with known findings concerning the organism’s environment, we focus first on a single species whose habitat is simple and well characterized. The obligate endocellular symbiont Buchnera aphidicola has lost many biosynthetic genes and demonstrates extremely successful symbiosis with its aphid host; it provides the host with amino acids that are essential for aphids (i.e., that aphids cannot synthesize) and relies on the host for nutrients it cannot synthesize (such as certain amino acids that are nes-
global topological features and are enriched in certain metabolic pathways (SI Text).

Further validation, spanning both multiple species and multiple compounds, builds on data concerning the biosynthetic capacity of several agents of human ehrlichiosis (an emerging infectious disease, primarily transmitted by ticks or ticks) and other insect symbionts) to synthesize amino acids and major vitamins and cofactors (ref. 29 and table 5 therein). Comparing our identified seed sets with these data (spanning nine species and 30 compounds; Table S3), we examined whether compounds reported in ref. 29 not to be synthesized in a specific species (and therefore to be exogenously acquired) are correctly classified by our algorithm as seeds, and whether compounds that can be synthesized are correctly classified as non-seeds (SI Materials and Methods). We find an overall strong agreement between the two datasets, with a classification accuracy of 79% \([P < 10^{-4};\) and accuracy of 93% \((P < 10^{-3})\) for the 10 cofactors alone; SI Materials and Methods].

Moreover, focusing only on our ability to correctly predict exogenously acquired seed compounds, we reach 95% precision (percent of correctly identified seeds out of all predicted seeds; SI Materials and Methods) and 67% recall (percent of correctly identified seeds of all exogenously acquired compounds). Considering the inherent noise in the underlying metabolic data (see SI Text), these scores attest to the ability of our method to successfully identify exogenously acquired compounds.

Having demonstrated that the obtained seed set data agree with various biological observations both across compounds and across species, we now turn to identify large-scale relations between the size and composition of the identified seed sets and the species’ environments. Here we limit our analysis only to the prokaryotic species, for which large-scale environmental data can be obtained (Materials and Methods). Furthermore, prokaryotes facilitate a comparison between different habitats without the complications associated with tissue-specific metabolism or varying trophic levels. We find that organisms that live in extreme and narrowly defined habitats (e.g., Archaea) tend to have smaller networks and smaller seed sets (see also ref. 13 and Fig. S4). This strong correlation between the organism’s environment and the network’s structure is particularly apparent for the bacterial phyla; the phyla with the smallest metabolic networks and smallest seed sets (even when normalized by the number of compounds) are Actinobacteria, Planctomyces, and Alpha/rickettsias (Fig. 2A). The average number of reactions and seed compounds across different bacterial taxonomic phyla. The number of seed compounds is estimated by the number of source components. Evidently, phyla that include mostly obligate intracellular parasites have, on average, the smallest metabolic networks and smallest seed sets. The fraction of the occurring compounds included in the seed set as a function of the ratio between the number of transcription factors and the genome size, across prokaryotic phyla. Again, phyla of intracellular parasites (e.g., Rickettsias, Mollicutes, Spirochete, and Chlamydia) inhabiting well defined and predictable environments have small seed sets (even when normalized by the number of compounds in the network) and a small number of transcription factors. The solid line represents a linear regression. The strong correlation attests to the alignment between the size of the seed set and habitat variability.

Using a covariation correlation assay (Materials and Methods) we further confirm that the growth environment of the various prokaryotic species correlates not only with the size of the seed sets but also with their composition. Data concerning the growth environment of each species are represented as a vector of four attributes (salinity, oxygen requirements, temperature range, and habitat) using discrete categories to describe each attribute (Materials and Methods, SI Materials and Methods, and SI Text). Considering the 446 bacterial and archaeal taxa for which environmental data can be obtained, a significant correlation of 0.25 (Pearson correlation test, \(P < 10^{-3}\); Materials and Methods) is found between the environmental “signature” and the seed composition of a species. This correlation is in fact higher than that found between the environmental signature and the occurring compounds composi-
tion (0.21, Pearson correlation test, $P < 10^{-3}$) despite the fact that occurring sets are markedly bigger than seed sets and potentially carry more information. For certain environmental attributes, species that share the same attribute value also tend to have similar seed sets. Specifically, species with microaerophilic, facultative, and aerobic oxygen requirements, multiple habitat, and mesophilic temperature range have significantly more similar seed sets than expected by chance ($P < 0.05$ after multiple testing correction; Materials and Methods).

Next, we characterize the evolutionary dynamics governing the gain and loss of seed and non-seed occurring compounds across a phylogenetic tree of life and the transitions between seed and non-seed states. To this end we analyzed the seed and occurring sets using methods borrowed from molecular evolution analysis (31, 32) and gene conservation analysis (33) and applying several phylogenetic analysis approaches (including both maximum parsimony algorithms and maximum likelihood models; Materials and Methods). Specifically, we considered a reference sequence-based tree and estimated the rate at which various compounds are integrated into (and lost from) evolving metabolic networks, and the rates at which seed compounds become non-seed occurring compounds and vice versa (Materials and Methods; Table 1). These estimates control for phylogenetic relations and can separate speciation dynamics from transition events.

Our findings suggest that novel compounds are integrated into metabolic networks as either seed or non-seed compounds, where integration events of non-seeds are ~1.5–2 times more frequent than integration of seeds (Table 1). Yet seed compounds have a higher tendency than non-seed compounds to be dropped completely from the network. Moreover, the rate of transition of a seed compound into a non-seed occurring compound is higher than that of the reverse process (Table 1). These findings suggest that, in general, the seed status is a transient phase in the “life” of a compound and that seed compounds tend to either be completely dropped from the network or change into non-seed compounds relatively fast. We further calculate several conservation measures (Materials and Methods) to estimate the expected evolutionary “lifespan” of the seed and non-seed states. We find that the seed state is significantly less conserved than the non-seed state (loss rates of 0.0523 and 0.0378, respectively; $P < 10^{-12}$, Wilcoxon rank sum test), confirming again the transient nature of the seed status.

Finally, the seed content of the various species included in our analysis was used to reconstruct a phylogenetic tree de novo (SI Materials and Methods) in a manner analogous to gene content-based phylogenies. Remarkably, this tree of life not only successfully clusters most of the taxonomic groups (Fig. 3) but is just as accurate (measured by its distance from a reference sequence-based tree; SI Materials and Methods) as a tree based on the entire set of occurring compounds (Table 2), despite being based on a significantly smaller number of compounds (seed compounds account, on average, for only 10.8% of the occurring compounds). This accurate reconstruction of a phylogenetic tree is not expected by chance for such small subsets of the occurring compounds, as demonstrated by the markedly less accurate trees obtained with random compound sets (Table 2), suggesting that the identified seed sets are a significant and fundamental characteristic of each species and its evolutionary history. A principal component analysis further demonstrates that the composition of the seed sets can be used to partition the major taxonomic groups (SI Text).

**Discussion**

This study introduces a large-scale reconstruction of metabolic environments using a cross-species analysis of 478 metabolic networks (and >2,200 metabolic compounds) to infer the set of compounds that each species extracts from its environment. Our

---

**Table 1. The relative frequencies of transitions across the phylogenetic tree between the various states of a compound**

<table>
<thead>
<tr>
<th>Original state</th>
<th>Absent</th>
<th>Non-seed</th>
<th>Seed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Absent</td>
<td>—</td>
<td>10.2058</td>
<td>7.2822</td>
</tr>
<tr>
<td>Non-seed</td>
<td>20.5747</td>
<td>8.0365</td>
<td></td>
</tr>
<tr>
<td>Seed</td>
<td>31.9681</td>
<td>21.9327</td>
<td>—</td>
</tr>
</tbody>
</table>

These frequencies describe the expected number of transitions from one state to the other among 100 transition events in a random set of compounds with an equal number in each state. The values presented are based on an ancestral network reconstruction (Materials and Methods; second assay). The transition rates that were obtained based on the compounds’ phylogenetic patterns (first assay) and on maximum likelihood estimates (third assay) demonstrate qualitatively similar trends and can be found in Tables S4 and S5. Note that this measure controls for the different frequencies of the various states in the data and hence is not biased by the smaller number of seed states.
remarkable capacity for adaptation to a wide range of environments, that these processes are not mutually exclusive (see also ref. 37). The strategies for externally acquiring previously produced compounds) (35) (positing a substrate-driven process where metabolic pathways can potentially yield more accurate results (34) (see also Figure S5). A strongly connected component is a maximal set of nodes such that for every pair of nodes $u$ and $v$ there is a path from $u$ to $v$ and a path from $v$ to $u$. The strongly connected components form a directed acyclic graph whose nodes are the components and whose edges are the original edges in the graph that connect nodes in two different components. In this graph, each component without incoming edges and at least one outgoing edge is defined as a source component. Each source component in the SCC decomposition forms a collection of candidate seed compounds. The set of seed compounds must include exactly one compound from each source component and should not include any other component. In the following, we briefly provide the intuition (based on a graph-based representation of the network; see also the discussion above concerning reaction stoichiometry): First, it should be noted that every strongly connected component is an equivalence class; if one of the compounds in the component can be produced then all others can be produced as well. Second, because source components do not have any incoming edges and at least one outgoing edge, if none of the organism's reactions in a source component is present in the seed set, none of the compounds in this component can be produced. Finally, if at least one compound from each source component is included in the seed set, a path from a seed compound to any other compound in the network can be found and hence all compounds in the network can be produced. All of the compounds in a source component are equally likely to be included in the seed set, each of these compounds was assigned a confidence level, $C$, (component size), denoting the compound's probability of being a seed. We used a threshold of $C \geq 0.2$ to determine whether a compound should be regarded as a seed or not (including all compounds that are part of source components of size 5 or less). With this threshold value we discarded on average only 3.3% of the seeds. Using other threshold values (specifically, $C \geq 0.1$ or $C \geq 0.01$) did not significantly change any of our results. Dataset S1 describes the composition of the seed set in each species (with the associated C values). See also Figure S6, illustrating the metabolic network of yeast with the seed compounds highlighted.

**Covariation Correlation Assay.** To examine the correlation between seed set composition and environmental attributes across all bacterial species, we applied an assay similar to the one used in ref. 39. Given $N$ species, two $N \times N$ distance matrices, $S_S$ and $S_D$, were constructed. $S_S$ represents the pairwise Jaccard distance (40) between the seed sets of the various species. $S_D$ represents the pairwise Hamming distance between the vectors of attributes describing the environments of these species. The Pearson correlation between the $(r^2 - r) / 2N$ entries forming the lower triangle of $S_S$ and $S_D$ was calculated. Statistical significance of the resulting correlation was computed by shuffling the species' labels 1,000 times and calculating the probability to achieve an equal or higher absolute value of the resulting correlation. An additional assay examines the similarity among seed sets of various species with a certain environmental attribute value. The average pairwise distance between the seed sets of all species with that specific attribute value was calculated and compared with the average distances obtained for 100,000 random collections of species (of the same size) to determine its statistical significance. The resulting $P$ values were corrected for multiple testing via the false discovery rate procedure (41).
Phylogenetic Analysis, Transition Rates, and Conservation. We consider a well established, sequence similarity-based tree as a reference phylogeny (42). This tree is based on 31 orthologs and includes a relatively large number of species, covering most of the taxonomic groups for which metabolic data are available. Our phylogenetic-based analyses were restricted to the species that could be matched to those included in the reference tree, resulting in a total of 178 species. A given compound in each species (extant or ancestral) can take one of three distinct states: absent (completely absent from the occurring compounds set), non-seed (an occurring compound that is not part of the seed set), or seed (an occurring compound that is part of the seed set). Our seed set analysis determines the state of each compound in the extant species. To calculate the evolutionary transition rate between the different states across the phylogenetic tree, we applied three assays analogous to those used in nucleotide substitution analysis (see also SI Materials and Methods). In the first assay, the compounds’ states in each internal node of the phylogenetic tree (representing ancestral species) were predicted, using Fitch’s small parsimony algorithm (43). Fitch’s algorithm finds the most parsimonious state assignment for all of the internal nodes of a phylogenetic network in each ancestral species, following Kreimer et al. (28). The seed set detection algorithm is then applied to each ancestral network to obtain the set of occurring and seed compounds in the internal nodes. Tamura and Nei’s detection algorithm is then applied to each ancestral network to obtain the set of occurring and seed compounds in the internal nodes. Tamura and Nei’s method is used again to estimate the relative transition rates. In the third assay, a maximum likelihood approach is applied to the phylectic patterns of all of the compounds in our analysis to obtain a maximum likelihood estimate of the substitution rates. This is computed with the PAML package (32) using the UNREST model. Two additional conservation measures, propensity for gene loss (PGL; a maximum parsimony measure) (44) and gene loss rate (GLR, a maximum likelihood measure) (33), were also applied to the phylectic patterns of the various compounds to compute the tendency of a compound to lose its state as a seed compound (or its state as a non-seed compound) during the evolutionary process. With these measures we do not distinguish between cases where the compound was completely dropped from the network and cases where its state converted to another state, but rather aim to characterize the level of conservation of each state. These conservation measures can therefore be conceived as representing the average “lifespan” of the state. The results obtained for the PGL measure were qualitatively similar to those obtained with the GLR measure and are not presented.

ACKNOWLEDGMENTS. We thank the anonymous reviewers and the editor for their helpful comments. We are grateful to Tomer Shlomi for numerous valuable suggestions (including the use of SCC decomposition). We thank Ruth Hershberg, Laura Kenney, Neil K. Levis, and Jeremie Van Cleve for their comments. E.B.’s research is supported by the Yeshaya Horowitz Association through the Center for Complexity Science, the Morrison Institute for Population and Resource Studies, and by a grant to the Santa Fe Institute from the James S. McDonnell Foundation. The 21st Century Collaborative Award Studying Complex Systems. This research is supported in part by National Institutes of Health Grant GM28016 (to M.W.F.). M.K.’s research is supported by grants from the Israel Science Foundation and the Israeli Ministry of Science and Technology. E.R.’s research is supported by grants from the Israel Science Foundation, the German–Israel Fund, the Yissahahu Horowitz Center for Complexity Science, and the Tauber Foundation.