An environmental perspective on metabolism

published in Journal of Theoretical Biology

doi:10.1016/j.jtbi.2007.10.036

Thomas Handorf[a,*], Nils Christian[a], Oliver Ebenhöh[b] and Daniel Kahn[c]

[a] Theoretical Biophysics, Department of Biology, Humboldt University Berlin,

Invalidenstr. 42, 10115 Berlin, Germany

[b] Max Planck Institute for Molecular Plant Physiology, Wissenschaftspark Golm,

Am Mühlenberg 1, 14476 Potsdam-Golm, Germany

[c] Université de Lyon; Université Lyon 1; CNRS; INRIA;

UMR 5558, Laboratoire de Biométrie et Biologie Evolutive

43 Bd du 11 Novembre 1918, 69622 Villeurbanne cedex, France

[*] Corresponding author. Tel. +49 30 2093 8325; fax +49 30 2093 8813.

E-mail address: handorf@physik.hu-berlin.de

Abstract

In principle the knowledge of an organism's metabolic network allows to infer its biosynthetic capabilities. Handorf, Ebenhöh and Heinrich (*J. Mol. Evol.* **61**:498–512, 2005) developed a method of network expansion generating the set of all possible metabolites that can be produced from a set of compounds, given the structure of a metabolic network. Here we investigate the inverse problem: which chemical compounds or sets of compounds must be provided as external resources in order to sustain the growth or maintenance of an organism, given the structure of its metabolic network? Although this problem is highly combinatorial, we show that it is possible to calculate locally minimal nutrient sets that can be interpreted in terms of resource types. Using these types we predict broad nutritional requirements for 447 organisms, providing clues for possible environments from the knowledge of their metabolic networks.

Keywords: nutrient, scope, metabolic network

1 Introduction

Among the numerous approaches developed by Reinhart Heinrich on metabolism, one important focus was the identification of design and optimality principles (see Reinhart Heinrich's annotated bibliography in this issue). He considered such principles as essential prerequisites for understanding evolution of metabolism. It was in this line of thought that Reinhart Heinrich and some of us at the Humboldt University developed the concept of network expansion in the early 2000's (Ebenhöh *et al.*, 2004; Handorf *et al.*, 2005).

The basic principle is that a reaction can only operate if all of its substrates are available as nutrients or can be provided by other reactions of the network. This condition is applied in an iterative manner. Starting from the nutrients, which are called seed compounds, operable reactions and their products are added to an expanding network. This interaction process will end if no further reaction fulfills the above condition. The set of metabolites in the expanded network is called the scope of the seed compounds and represents all metabolites that can in principle be synthesized from the seed by the analyzed metabolic network (Handorf *et al.*, 2005).

This concept has been applied in recent papers, including a discussion on hierarchical structuring of metabolic networks (Handorf *et al.*, 2006), a comparison of metabolic capabilities of organism specific networks (Ebenhöh *et al.*, 2005), a model of metabolic evolution (Ebenhöh *et al.*, 2006) and the analysis of changes of metabolic capacities in response to environmental perturbations (Ebenhöh & Liebermeister, 2006). Further, scopes have been utilized to study the effect of oxygen in metabolic networks (Raymond & Segré, 2006) and to predict the viability of mutant strains (Wunderlich & Mirny, 2006).

In this work we consider the inverse problem of determining sets of seed compounds required for the synthesis of a specific compound or set of compounds. In particular the latter set may comprise metabolic precursors that the cell requires for maintenance or growth. Therefore solving this inverse problem may indicate minimal nutritional requirements that must be met to sustain maintenance or growth of an organism, based on the knowledge of its metabolic network. We apply this methodology to a number of organisms for which metabolic networks are defined in the KEGG database (Kanehisa *et al.*, 2006) and show that this inverse methodology can indeed provide clues on possible nutritional requirements of organisms and their environment.

2 Methods

2.1 The target set of required metabolites

A key function of metabolism is to chemically convert available nutrients into products which are required by other cellular processes. Precursors for central cellular functions like protein synthesis, DNA replication, energy or cofactor production, are ubiquitous. Since the detailed requirements may vary from cell type to cell type, we apply a systematic approach, combined with biological knowledge, to identify a universal set of necessary metabolites, referred to as the target set T.

We construct the target set by determining those metabolites which occur in at least 90% of the analyzed organisms. These include amino acids, nucleotides, many cofactors, organic acids and sugar phosphates. We manually refined this list by including plausible compounds which are missing and removing compounds whose presence in the target set seemed not reasonable. The detailed list of target metabolites as well as the removed compounds and the reasons for their removal can be found in the supplementary material.

2.2 Identifying minimal resources

To identify minimal sets of required resources that enable an organism to produce all metabolites contained in the target set, we develop an algorithm that relies on the method of expanding networks which was introduced in Handorf *et al.* (2005). Starting from a given set of initial metabolites, the seed S, the network expansion algorithm determines all those metabolites which a particular metabolic network is capable to produce when only the seed compounds are available. These metabolites are called the scope of the seed, denoted $\Sigma(S)$. The identification of minimal resources is now described as the problem to identify minimal sets of seed compounds for which the scope contains all target metabolites. A seed S is minimal if its scope contains the target T and no proper subset of S fulfills this condition:

$$S \text{ is minimal seed if } T \subseteq \Sigma(S) \text{ and } \forall S' \subset S : T \not\subset \Sigma(S') \tag{1}$$

For a given network, we determine minimal seeds with the following greedy algorithm: (1) Initially, we define an ordered list containing all metabolites occurring in the network.

Clearly, the seed composed of all metabolites from the list must produce a scope containing the target set. (2) Beginning from the top, stepwise each metabolite is removed from the list and the scope is recalculated for the corresponding reduced seed. If now the scope does not contain the full target set, the metabolite is written back to the list, otherwise it remains permanently removed. (3) Step (2) is repeated until the complete list has been traversed. The metabolites contained in the resulting list represent a minimal seed because the further removal of any metabolite would result in a scope that does not contain all target metabolites.

Since the ordering of the list in step (1) determines which metabolites are preferentially removed (those near the top) and which preferentially remain in the seed (those near the end), differently ordered lists will result in different minimal seeds. Clearly, it is impossible to test all possible orderings of the list of metabolites and thus the complete set, denoted $\mathcal{M}$, containing all minimal seeds cannot be calculated. However, significant information about the structure of $\mathcal{M}$ can be obtained by calculating seeds for a sufficient number of random orderings. Further, as not all minimal seeds are equally biologically meaningful, the size of the search space can further be reduced by incorporating biological information on which metabolites can actually be used as nutrients. Such metabolites should reside near the end of the list, which can be achieved in the following way: First, all metabolites are sorted by decreasing molecular weight. This ordering leads to a preferential removal of large metabolites from the list of possible seed compounds and has been introduced to avoid minimal seeds containing only a small number of chemically rich but large metabolites that are unlikely to be transported into a cell. Second, for several metabolites, the transport processes over the membrane are well characterized. For our calculations, we have identified from the KEGG pathways KO02010 and KO02060 all those metabolites which can be translocated by ABC transporters or the phosphotransferase system. This biological knowledge has been considered by symbolically assigning all these compounds a negative "mass" which shifts them towards the end of the list. The resulting list ensures that metabolites that are known to be transported and small molecules are preferentially chosen as seed compounds.

To identify and analyze a large number of possible minimal resource sets, we construct a large number of perturbed lists and apply the algorithm repeatedly for each network.

The random perturbation is designed in a way ensuring that large metabolites remain as a tendency near the top of the list. Specifically, we randomly chose two metabolites from the list and exchange their positions with a probability

$$p = \begin{cases} \exp(-\frac{1}{\beta}\Delta m) & \text{for } \Delta m > 0 \\ 1 & \text{for } \Delta m \leq 0 \end{cases}, \tag{2}$$

where Δm is the difference of the molecular weights of the molecules at the two positions to be exchanged. It is positive if the heavier compound is the compound situated closer to the beginning of the list. The constant β determines the degree of disorder that is allowed in the resulting list, a value $\beta = 0$ strictly forbids that a heavier compound is placed in a later position than a lighter compound and a choice of $\beta = \infty$ would completely ignore the weights. For our calculation, we chose $\beta = 20$ Da and performed 10,000 exchange operations to generate one randomized list. For each network, 1000 such lists are generated, yielding 1000 minimal, possibly identical seeds.

2.3 Identifying groups of exchangeable resource metabolites

It is typically the case that related substances can be used by a particular organism as alternative resources, for example various sugars may serve as carbon source. We call two metabolites A and B exchangeable if they appear as alternatives in minimal seeds. Formally, A and B are exchangeable if for all seeds $S_A = (A, X_0..X_i) \in \mathcal{M}$, also the seeds $S_B = (B, X_0..X_i)$ are in $\mathcal{M}$. Because the complete set $\mathcal{M}$ is not known, we test this condition on the set of minimal seeds that have been determined by the algorithm described above. First, we determine all metabolites which are found in at least one minimal seed. Then, for every pair A, B of these metabolites we (1) identify all minimal seeds S_A containing metabolite A; (2) for each S_A construct the corresponding set S_B by replacing A by B (S_B is not necessarily a minimal set); (3) test whether $\Sigma(S_B)$ contains the target set T. If the latter condition is true for all S_A, we assume A to be replaceable by B. Conversely, we test whether B is replaceable by A. The metabolites A and B are considered exchangeable if they can be replaced by each other.

Clearly, since the calculated minimal seeds represent only a fraction of the complete set $\mathcal{M}$, the result is not necessarily correct, rather, pairs of metabolites may wrongly be classified as exchangeable (false positives). The classification becomes more accurate the

more seeds are tested. Thus, metabolites occurring only a few times in the set of seeds are especially susceptible for being falsely predicted as exchangeable with other compounds. In order to keep the computing times feasible, a maximum of 50 seeds are tested even if the corresponding metabolites take part in a larger number of seeds. If the compounds A and B are found not to be exchangeable, then there exists a seed S_A where A cannot be replaced by B (or vice versa) in order to obtain the target set T. Hence, the algorithm will not predict false negatives, even though not all seeds in $\mathcal{M}$ are tested.

2.4 Graphical representation of minimal resources

The information on the exchangeability of seed compounds can be illustrated as an undirected graph with nodes representing seed compounds which are connected if they are exchangeable. To increase the reliability of the assembled information, the graph is further reduced by removing all compounds which occur in only one seed. Such a graph decomposes into clusters which may contain one or several compounds. Compounds within one cluster can be alternatively used in the seeds and can never occur together in a minimal seed. Metabolites from different clusters generally cannot be used as alternative seed compounds.

The total number of seeds in which compounds of a cluster occur determine how important it is to include one of the metabolites into a minimal resource. If this number equals the number of distinct minimal seeds, the presence of one of the metabolites from the cluster is essential for the organism. If it is lower, compounds of that cluster can actually be exchanged by metabolites from other clusters. It is also possible that one metabolite is exchangeable by two other metabolites which together provide the same required chemical structures. However, our algorithm is not able to detect such equivalences. For this reason, we call a cluster essential if compounds of this cluster occur in at least 90% of all minimal seeds.

In principle, if no exchangeable metabolites were falsely predicted, all compounds within a cluster should be pairwise exchangeable by transitivity and the cluster should form a complete subgraph (a clique). Therefore, if a cluster is not fully connected, it must contain at least two metabolites which have been falsely predicted as exchangeable.

2.5 Global classification of resource types

To consolidate the information contained in the graphs characterizing the required resources for each single organism, we construct a graph that represents classes of metabolites on a global level. For this, two metabolites are linked if they tend to be exchangeable in most organisms. Specifically, for two metabolites A and B, we determine the numbers of clusters they are found in across all considered organisms, denoted x_A and x_B, respectively, as well as the number of clusters containing both compounds, denoted x_{AB}. We introduce the coefficient

$$C_{AB} = \frac{x_{AB}}{\min(x_A, x_B)} \tag{3}$$

reflecting their co-occurrence as exchangeable metabolites. $C_{AB} = 1$ if these compounds appear exchangeable in all organisms. $C_{AB} = 0$ if they are not found exchangeable in any organism considered. In the present analysis we join two nodes if $C_{AB} \geq 0.8$. The resulting graph consists of separate connected components that can be interpreted as global resource types.

3 Results

We retrieved 447 out of 489 organism specific metabolic networks from the KEGG database (Kanehisa *et al.*, 2006) including information on the reversibility of the reactions (see supplementary material for the detailed procedure). For each of these metabolic networks, we calculated possible resource compounds, determined which of these compounds are exchangeable and represented this exchangeability as a graph. As an example, we present the resource graph for *E. coli* K12 strain MG1655 in Fig. 1. Interestingly, of the 1000 calculated minimal seeds, only 560 are distinct. The fact that many identical seeds were found indicates that our algorithm has searched a considerable part of the biologically relevant minimal seeds. In total, 77 metabolites have been found in at least one of the seeds. All of the separated clusters in the graph are complete subgraphs, indicating that the number of wrongly predicted pairs of exchangeable metabolites is low. The numbers below the metabolite names give the number of distinct seeds in which a compound was found. The sum of these numbers over a cluster characterizes its essentiality.

Thiamin appears as an isolated compound and was found in 559 minimal seeds, which

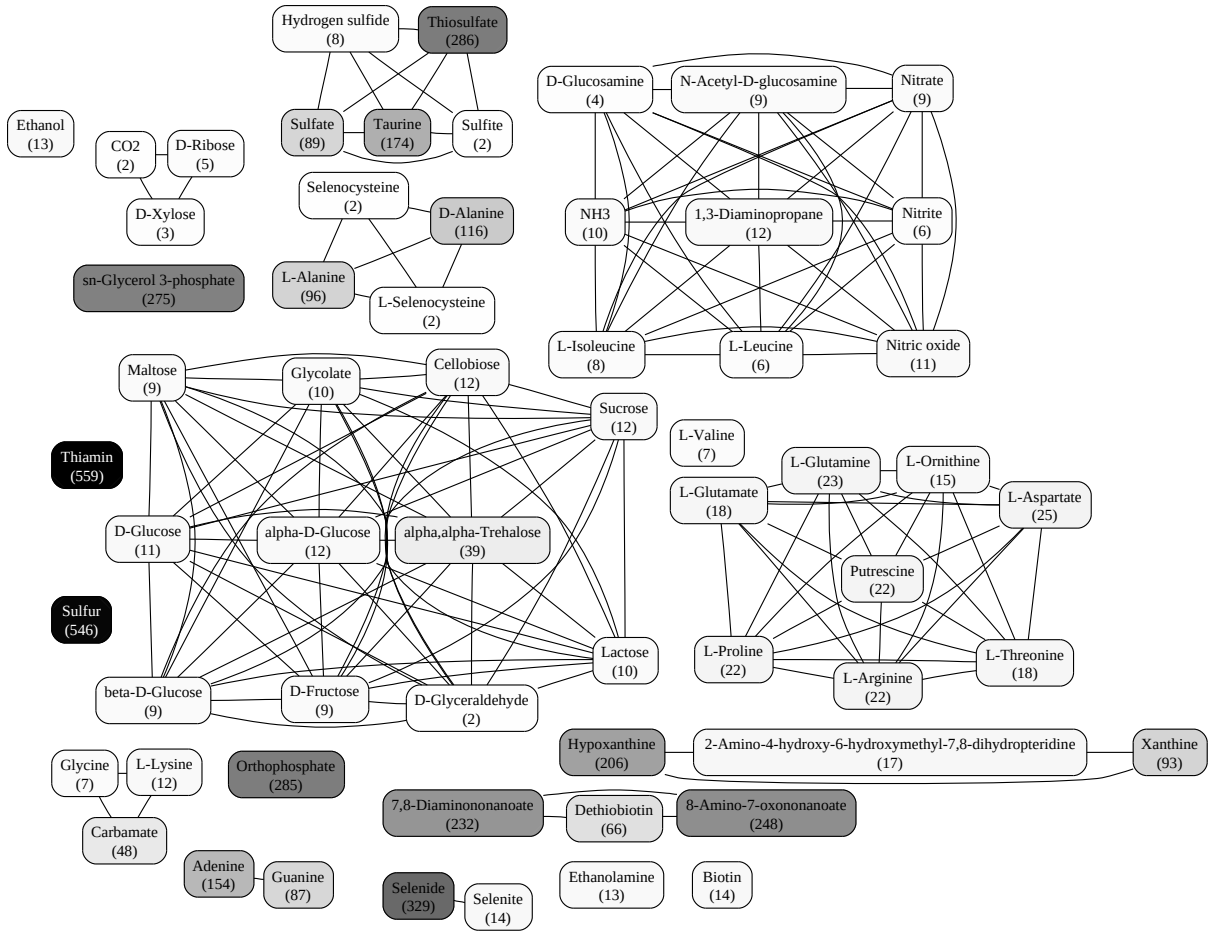


Figure 1: Graph indicating exchangeable seed compounds for *E. coli* K12 strain MG1655. The numbers in parentheses indicate in how many of the 560 distinct minimal seeds particular compounds occur. The shading supports this information by displaying frequent compounds as dark and rare compounds as light nodes.

is consistent with the known auxotrophy of *E. coli* K12. The essential cluster at the top of Fig. 1 consists of sulfur sources including sulfate, thiosulfate and taurine. Dethiobiotin and two other metabolites form a cluster occurring in 546 minimal seeds. It appears in parallel with a compound labelled in KEGG as 'sulfur' (C00087), which is in fact an uncharacterized sulfur source in biotin synthesis. The fact that the biotin cluster appears essential contradicts the known prototrophy of *E. coli* K12 and is due to the lack of annotated pathway for pimeloyl-CoA synthesis in *E. coli*. Other resource clusters are found in significantly fewer minimal seeds, indicating that they are optional resources that can be substituted or synthesized from other types of resources. For example the carbohydrate cluster contains 11 exchangeable compounds that are found altogether in only 135 out of 560 minimal seeds, consistent with the fact that *E. coli* can use alternative carbon sources.

In order to assess our predictions, we also compared the resource graphs found for *Rickettsia prowazekii* and *Tropheryma whipplei* with the information provided in the Metagrowth database (Ogata & Claverie, 2005). The Metagrowth database provides evidence and hypotheses on culture conditions of selected obligate parasitic bacteria. Our calculations are in good agreement with most deficiencies of biosynthetic pathways reported in Metagrowth. Most compounds lacking a biosynthetic pathway according to Metagrowth could be predicted by our algorithm. The others were not found because they were not necessary for the synthesis of any compound in the target set. Details can be found in the supplementary material.

Resource graphs were identified for all 447 analyzed organisms. In most cases these clusters can also be categorized, for example as carbohydrates, amino acids, or vitamins. However they usually differ in their exact compositions, which is not surprising because of the structural differences of the underlying networks. As described in the methods section, we merged the information contained in the organism specific resource graphs into a graph representing global resource type clusters (see supplementary material). Due to the variety of the organisms and the fact that the algorithm is based on a statistical approach, different clusters in this graph may actually represent the same resource types. Therefore we merged manually some of the clusters containing closely related biochemical compounds. Also some compounds not normally used as nutrients, such as sugar phos-

phates, have not been considered further in this analysis. Table 1 provides a representative for each of the 45 resource types, with the full lists of the corresponding compounds given in the supplementary material.

Acetate	Bases	beta-Alanine	Biotin	CO2
D-Fructose	D-Glucose	D-Mannose	D-Ribose	Folate
Glycerol	Glycine	Isopentenyl	L-Alanine	L-Arginine
L-Asparagine	L-Aspartate	L-Cysteine	L-Glutamate	L-Histidine
L-Isoleucine	L-Leucine	L-Lysine	L-Methionine	L-Ornithine
L-Phenylalanine	L-Proline	L-Serine	L-Threonine	L-Tryptophan
L-Tyrosine	L-Valine	Maltose	Nicotinate	Nucleoside
OrganicAcids	Orthophosphate	Pantetheine	Propanoate	Riboflavin
Shikimate	Succinate	Sucrose	Sulfate	Thiamin

Table 1: List of resource types used in the comparative analysis of the nutritional profiles of the analyzed organisms.

The global resource types have been used for a comparative analysis of the nutritional requirements of all organisms. For each organism, all resource types have been matched with seed compound clusters of that organism if they have at least one compound in common. A specific resource type has been characterized as essential for a particular organism if at least one of the corresponding clusters is present in at least 90% of all its distinct minimal seeds. Consequently, it is characterized as optional if it is present in fewer than 90% of the seeds and it is unused if no compound associated with the resource type is present in any seed. The results for the Proteobacteria phylum are presented as a matrix in Figure 2, where each row corresponds to an organism and each column to a global resource type. The corresponding entries in the matrix indicate the category of the resource type, red denoting essential, green optional and black unused resources. To visualize the relatedness of the considered organisms, they are ordered according to the topology of the phylogenetic tree as defined in KEGG (Kanehisa *et al.*, 2006). The global resource types were sorted, maximizing the similarity of neighboring columns using a seriation method as introduced in Gelfand (1971) and an euclidian distance between the nutrient profiles. The complete matrix comprising all considered species can be found in the supplementary material.

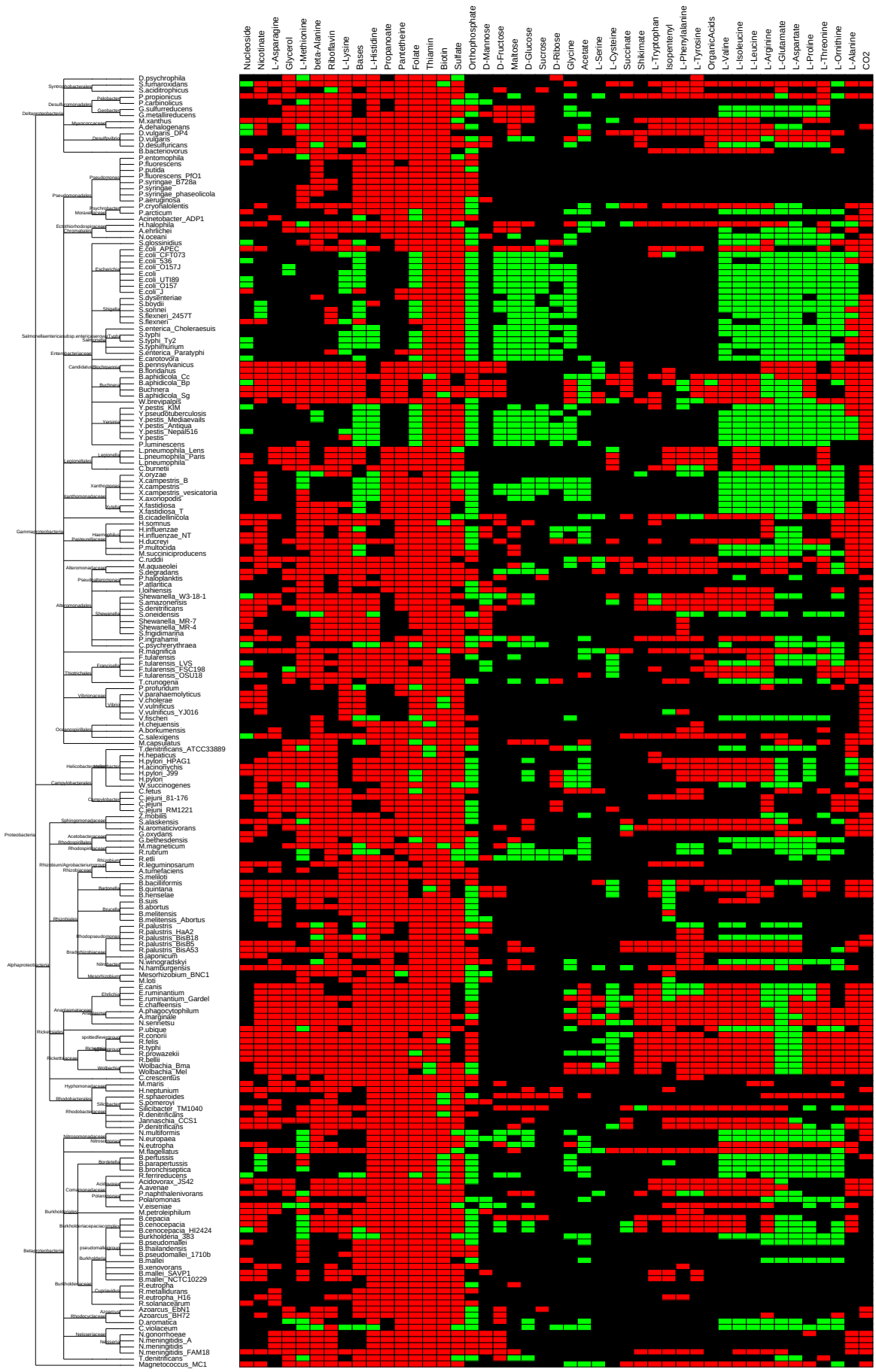


Figure 2: Predicted nutritional profiles of proteobacteria. For each organism it is indicated whether it uses the resource types of Table 1 in an essential (red) or optional way (green).

4 Discussion

In this work we presented a novel method for the analysis of nutritional requirements of organisms based on the structural analysis of their metabolic networks. The algorithm is effective in predicting precursors of essential compounds that cannot be synthesized by the organism itself. It also allows to determine sets of alternative non-essential nutrients representing metabolic options for potential different environments. Predicted nutrient sets differ among organisms, depending on their metabolic networks. In order to perform a comparative analysis, we defined general resource types containing well known resources, such as sugars, aminoacids or vitamins or their direct precursors. Organisms could thus be compared in terms of their usage of nutritional resource types in an essential or non essential way.

We focused on Proteobacteria as an example for this inverse metabolic analysis. This group of bacteria is quite diverse in terms of metabolic requirements, with organisms ranging from prototrophs to multiple auxotrophs. The analysis of the nutritional requirements of different species generally indicates similarities in the requirements of closely related species (Fig. 2), which is consistent with classical approaches of microbial identification based on spectra of growth substrates or fermentable products. For example among Enterobacteriaceae, genera like *Escherichia, Yersinia, Salmonella* or *Shigella* have similar nutritional profiles and can utilize many nutrient types in a non essential way. On the other hand and in the same family, the *Buchnera* genus possesses a completely different profile, requiring many nutrient types including amino acids in an essential way. This is consistent with the endosymbiotic nature of these organisms whose metabolism relies heavily on their hosts (aphids). Other predicted multiple auxotrophs include many obligate symbionts or parasites. For instance all Rickettsiales including the symbiotic *Wolbachia* belong to this group, as well as mollicutes and chlamydiales outside the proteobacteria. Thus our results clearly distinguish between organisms having a relatively low number of essential nutrients and those exhibiting numerous predicted auxotrophies, allowing for a broad prediction of possible life-sustaining environments.

We compared the amino acid nutrition profile obtained for the *Buchnera* strain APS with the analysis of amino-acid essentiality of Shigenobu *et al.* (2000). We found that Ala, Asn, Tyr and Pro are correctly predicted as required for *Buchnera* as they are provided

by the host. Gly can be viewed as required instead of Ser since they are interchangeable. His, Trp, Thr, Phe and Arg are correctly predicted as not required, even though Phe and Arg are found to be utilizable. Asp and Glu on the other hand should be required according to Shigenobu *et al.* (2000). However, they are only partially used in our calculations. Pathways for Ile, Leu, Val and Met are incompletely annotated according to Shigenobu *et al.* (2000), which explains our prediction that they are essential. While the correspondence with our results is quite good, this example also shows some limitations of our algorithm. In particular, the discrimination between essential and non-essential nutrients is dependent on a tunable threshold value. With the current value, Asp and Glu were not predicted as exchangeable. When looking at the resource graph of *Buchnera* (supplementary material) it can however be seen that their occurrences are significantly higher that those of Phe and Arg that are not required according to Shigenobu *et al.* (2000).

Additional information will be required in order to reach a more precise and detailed prediction of life-sustaining environments. First, the list of preferred seed compounds and its sorting are critical. Different organisms may use compounds with different priority and the set of metabolites that they can transport will vary. Therefore our analysis will miss some specific requirements of individual organisms. Second, our analysis is sensitive to errors or gaps in the described metabolic networks. Here we used metabolic networks from the KEGG database as derived from KEGG orthology groups (Kanehisa *et al.*, 2006). These networks are constructed from a systematic analysis of putative enzyme coding genes on the basis of homology, which inevitably leads to a small rate of false or missing predictions (see for instance Claudel-Renard *et al.* (2003)). Missing enzyme predictions may cause false inference of auxotrophies, as seen for instance for biotin auxotrophy with *E. coli*. Third, information on reaction directionality is not yet systematically available for all enzymes, which may cause some unrealistic predictions. In addition our approach could be extended by adding thermodynamic constraints or flux balance constraints to the structural constraints considered here, which would require a different set of methods (see for instance Imielinski *et al.* (2006) for a recent approach based on convex cone analysis).

The inverse metabolic analysis presented here could prove useful in several respects. For instance there are numerous fastidious bacteria that are not readily cultivatable. They can be found among pathogenic species, in symbiotic associations or in various environ-

mental samples. For these bacteria it has paradoxically become more accessible to obtain their complete genome sequence than to manipulate them experimentally, because of the lack of a culture medium. The method proposed here could be used to infer essential additions to culture media, which would complement manually curated metabolic knowledge bases such as Metagrowth (Ogata & Claverie, 2005). Inverse metabolic analysis could also provide clues on possible environments for poorly characterized organisms for which genome data are available. On the other hand, for well known organisms, the method can be used to uncover shortcomings in metabolic network annotation. In the longer term we envision to combine this type of inverse metabolic approach with the reconstitution of ancestral metabolic networks (Ebenhöh *et al.*, 2006), which could provide clues on environments associated with the emergence and evolution of ancestral phyla.

5 Acknowledgements

We thank the following organizations for financial support: The German Research Foundation, in particular the Collaborative Research Center "Theoretical Biology: Robustness, Modularity and Evolutionary Design of Living Systems" (Handorf, T.) and the International Research Training Group "Genomics and Systems Biology of Molecular Networks" (Christian, N.), the German Federal Ministry of Education and Research, Systems Biology Research Initiative "GoFORSYS" (Ebenhöh, O.) and the Institut National de la Recherche en Informatique et Automatique (Kahn, D.).

References

Claudel-Renard, C., Chevalet, C., Faraut, T. & Kahn, D. (2003). Enzyme-specific profiles for genome annotation: PRIAM. *Nucleic Acids Res.* **31**, 6633–6639.

Ebenhöh, O., Handorf, T. & Heinrich, R. (2004). Structural analysis of expanding metabolic networks. *Genome Informatics* **15**, 35–45.

Ebenhöh, O., Handorf, T. & Heinrich, R. (2005). A cross species comparison of metabolic network functions. *Genome Informatics* **16**, 203–213.

Ebenhöh, O., Handorf, T. & Kahn, D. (2006). Evolutionary changes of metabolic networks and their biosynthetic capacities. *IEE Proc. Systems Biol.* **153**, 354–358.

Ebenhöh, O. & Liebermeister, W. (2006). Structural analysis of expressed metabolic subnetworks. *Genome Informatics* **17**, 163–172.

Gelfand, A. E. (1971). Rapid seriation methods with archaeological applications. *Mathematics in the Archaeological and Historical Sciences* , 186–201.

Handorf, T., Ebenhöh, O. & Heinrich, R. (2005). Expanding metabolic networks: Scopes of compounds, robustness, and evolution. *J. Mol. Evol.* **61**, 498–512.

Handorf, T., Ebenhöh, O., Kahn, D. & Heinrich, R. (2006). Hierarchy of metabolic compounds based on their synthesizing capacity. *IEE Proc. Systems Biol.* **153**, 359–363.

Imielinski, M., Belta, C., Rubin, H. & Halasz, A. (2006). Systematic analysis of conservation relations in *Escherichia coli* genome-scale metabolic network reveals novel growth media. *Biophys. J.* **90**, 2659–2672.

Kanehisa, M., Goto, S., Hattori, M., Kinoshita, K., Itoh, M., Kawashima, S., Katayama, T., Araki, M. & Hirakawa, M. (2006). From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res.* **34**, D354–357.

Ogata, H. & Claverie, J. (2005). Metagrowth: a new resource for the building of metabolic hypotheses in microbiology. *Nucleic Acids Res.* **33**, D321–324.

Raymond, J. & Segré, D. (2006). The effect of oxygen on biochemical networks and the evolution of complex life. *Science* **311**, 1764–1767.

Shigenobu, S., Watanabe, H., Hattori, M., Sakaki, Y. & Ishikawa, H. (2000). Genome sequence of the endocellular bacterial symbiont of aphids *Buchnera* sp. APS. *Nature* **407**, 81–86.

Wunderlich, Z. & Mirny, L. (2006). Using the topology of metabolic networks to predict viability of mutant strains. *Biophys. J.* **91**, 2304–2311.

Supplementary material:

An environmental perspective on metabolism

March 10, 2008

1 Data import

We have extracted the metabolic networks for 447 organisms the KEGG database (as of Feb 13, 2007). The organisms have been selected in the following way: It has been verified that the number of reactions was realistic compared to similar organisms, if available. Otherwise, when the number of reactions seemed abnormally low, the original genome sequence paper was checked to verify that the low number is in line with biological knowledge, for example in case of a low number of genes, a metabolic deficiency and/or parasitism.

The corresponding metabolic networks have been extracted as follows. First, from the LIGAND subdivision (plain text file), the complete list of 6825 reactions has been imported. The reactions have been checked for consistency. We rejected 290 reactions because they showed an erroneous stoichiometry, by which we mean that some atomic species occurred in different numbers on both sides of the reaction. Further, we did not include 342 reactions involved in glycan synthesis because the focus of our investigation lies on the metabolism of small chemical species and does not include macromolecular syntheses.

Information on the reversibility of reactions has been extracted from the KGML files which specify the pathways for all organisms included in KEGG. In general, a particular reaction is listed in several KGML files and the information on its reversibility may be ambiguous. In fact, we identified 136 reactions for which this is the case. For the present

calculations, we consider a reaction to be irreversible only if it is defined as irreversible in all corresponding occurrences in the KGML files. This is the case for 2622 reactions.

The organism specific networks were determined using the 'reaction' and 'enzyme' files from the KEGG/LIGAND database. In a first step, for all reactions the EC numbers of the catalyzing enzymes were retrieved from their corresponding entries in the 'reaction' file (section ENZYME). Subsequently, from the 'enzyme' file, for each enzyme a list of organisms is obtained in which there exists a corresponding gene (section GENES). Thus, for each organism the metabolic network is defined by all those reactions for which a catalyzing enzyme is encoded in its genome. In all cases where an enzyme is not fully classified (e.g. EC1.3.1.-), the corresponding entry in the 'enzyme' file contains no GENES section. As a consequence, no such reactions are included in organism specific networks.

Further, the KO section of the database is inspected. Reactions specified in the DBLINKS/RN section of a KO entry are also assigned to the set of reactions of the organisms listed in the GENES section of this entry.

2 Target metabolites

We define a set of metabolites, called the target set T, that are required for the synthesis of compounds necessary to sustain cellular maintainance or growth. In our analysis this set contains all chemical compounds that are present in at least 90 percent of the organism specific networks. The set contains amino acids, nucleotides, cofactors, organic acids and phosphorylated sugars. From this list, the following metabolites have been removed because they usually require themselves or a related compound in the seed: IDP, ITP and dCMP due to missing phosphorylation or dephosphorylation reactions; 'Glutamate', as it is a generic version of L-Glutamate; Mercaptopyruvate, as it requires toxic compounds for it synthesis; 4-Trimethylammoniobutanal and 3-Hydroxy-N6,N6,N6-trimethyl-L-Lysine as they are part of the degradation of protein bound Lysine that cannot be produced with the present reactions; 3-Phospho-D-erythronate and Phosphoenol-4-deoxy-3-tetrulosonate, as they are separated from the remaining network; Selenomethionine and Se-Adenosylselenomethionine as they are isolated in most cases; 2-Deoxy-5-keto-D-gluconic acid 6-phosphate which is part of the poorly annotated inos-

itol degradation pathway. Further we removed from the list all compounds containing variable elements such as residues (indicated by an "R" in the sum formula) or multiple chain elements because it is not assured whether these can be synthesized from regular compounds by the defined reaction set. The complete list of compounds within the target set is given in Table 1.

3 Preferred seed compounds

For each organism, we determine minimal sets of seed compounds from which all metabolites of the target set T can be produced. The algorithm allows to provide a set of compounds that are preferably used as seed compounds. In this work we assumed that small organic and inorganic precursors can be taken up by the cell and used as nutrients. Further, we assumed that compounds transported by ABC transporters or the Phosphotransferase system, as defined by the KEGG pathways KO02010 and KO02060, respectively, can be utilized by metabolism. Table 2 shows a list of the preferred seed compounds.

4 Compound clusters and resource types

The algorithm predicts clusters of exchangeable seed compounds for each organism. These clusters are provided as supplementary files for a few example organisms listed in Table 3. Organism specific clusters are compiled into general clusters, containing seed compounds that are exchangeable in most organisms. In general these clusters contain classes of metabolites such as sugars, amino acids or vitamins, clustered with their respective precursors or derivatives. The supplementary file "global_clusters.eps" shows these automatically determined clusters. The nutrients sets are further processed manually. If compounds in a cluster are obviously direct precursors of compounds in another cluster these two are united. The resulting sets of seed compounds are called 'resource types' and labelled according to their most familiar members. Table 4 shows a list of the defined nutrient types and their member metabolites.

Acetate	Acetyl-CoA
Adenine	Adenosine
ADP	alpha-D-Glucose 6-phosphate
AMP	ATP
beta-D-Fructose 1,6-bisphosphate	beta-D-Fructose 6-phosphate
beta-D-Glucose 6-phosphate	Biotin
Carbamoyl phosphate	CDP
CMP	CO2
CoA	CTP
D-Erythrose 4-phosphate	D-Fructose 1,6-bisphosphate
D-Fructose 1-phosphate	D-Fructose 6-phosphate
D-Glucosamine 6-phosphate	D-Glucose 1-phosphate
D-Glucose 6-phosphate	D-Glyceraldehyde
D-Mannose 6-phosphate	D-Ribose 5-phosphate
D-Ribulose 5-phosphate	D-Sedoheptulose 7-phosphate
D-Xylulose 5-phosphate	dADP
dAMP	dATP
dCDP	dCTP
Deamino-NAD+	Deoxyadenosine
Deoxyguanosine	Dephospho-CoA
dGDP	dGMP
dGTP	Dihydrofolate
Dihydropteroate	Dimethylallyl diphosphate
dTDP	dTMP
dTTP	dUDP
dUMP	dUTP
FAD	FADH2
Formate	Fumarate
GDP	Glycerone phosphate
Glycine	GMP
GTP	Guanine
H+	H2O
H2O2	HCO3-
IMP	Inosine
Isopentenyl diphosphate	L-Alanine
L-Arginine	L-Asparagine
L-Aspartate	L-Cysteine
L-Glutamate	L-Glutamine
L-Histidine	L-Isoleucine
L-Leucine	L-Lysine
L-Methionine	L-Ornithine
L-Phenylalanine	L-Proline
L-Serine	L-Threonine
L-Tryptophan	L-Tyrosine
L-Valine	N6-(1,2-Dicarboxyethyl)-AMP
NAD+	NADH
NADP+	NADPH
NH3	Nicotinamide D-ribonucleotide
Nicotinate D-ribonucleotide	Orthophosphate
Oxaloacetate	Oxygen
Phenylpyruvate	Phosphoenolpyruvate
Propanoyl-CoA	Pyrophosphate
Pyruvate	S-Adenosyl-L-homocysteine
S-Adenosyl-L-methionine	Sedoheptulose 1,7-bisphosphate
Sedoheptulose 7-phosphate	sn-Glycerol 3-phosphate
Succinate	Tetrahydrofolate
Thiamin diphosphate	UDP
UDP-N-acetyl-D-glucosamine	UMP
UTP	Xanthosine 5'-phosphate
(2R)-2-Hydroxy-3-(phosphonooxy)-propanal	(S)-Malate
1-(5'-Phosphoribosyl)-5-amino-4-(N-succinocarboxamide)-imidazole	1-(5'-Phosphoribosyl)-5-amino-4-imidazolecarboxamide
10-Formyltetrahydrofolate	2,3-Bisphospho-D-glycerate
2-(alpha-Hydroxyethyl)thiamine diphosphate	2-Oxobutanoate
2-Oxoglutarate	2-Phospho-D-glycerate
3-(4-Hydroxyphenyl)pyruvate	3-Dehydroquinate
3-Dehydroshikimate	3-Methyl-2-oxobutanoic acid
3-Oxopropanoate	3-Phospho-D-glycerate
3-Phospho-D-glyceroyl phosphate	5,10-Methenyltetrahydrofolate
5,10-Methylenetetrahydrofolate	5-Phospho-alpha-D-ribose 1-diphosphate

Table 1: List of compounds in the target set, alphabetically sorted.

ABC transported	PTS transported	Small metabolites
Betaine	alpha,alpha-Trehalose	Acetamide
Butyro-betaine	alpha-D-Glucose	Acetate
Carnitine	Arbutin	Allyl alcohol
Choline	Ascorbate	Carbamate
Choline sulfate	beta-D-Glucose	Carbonic acid
Cobalt	beta-D-Glucoside	Chloride
Crotono-betaine	Cellobiose	Cl-
Cyclomaltodextrin	D-Fructose	CO2
D-Allose	D-Glucosamine	Cobalt
D-Aspartate	D-Glucose	Dimethylamine
D-Galactose	D-Sorbitol	Ethanol
D-Glucose	Galactitol	Ethanolamine
D-Methionine	Glucose	Ethylamine
D-Ribose	Lactose	Fe2+
D-Xylose	Maltose	Fe3+
Fe(III)dicitrate	Mannitol	Formate
Fe(III)hydroxamate	N-Acetyl-D-glucosamine	Glycine
Fe-enterobactin	N-Acetylgalactosamine	Glycolate
Fe2+	Nitrogen	H+
Fe3+	Salicin	H2O
Ferrichrome	Sorbose	HCO3-
Heme	Sucrose	HO-
Hemine		Imidazole
Iron chelate		Iron
L-Arabinose		Magnesium
L-Arginine		Manganese
L-Aspartate		Methane
L-Glutamate		Methanol
L-Glutamine		Methylguanidine
L-Histidine		NH3
L-Isoleucine		Nitrate
L-Leucine		Nitric oxide
L-Lysine		Nitrite
L-Methionine		Nitrogen
L-Ornithine		Nitrous oxide
L-Proline		Oxygen
L-Threonine		Propan-2-ol
L-Valine		Propane-1-ol
Maltose		Propanoate
Manganese		Sulfur
Molybdate		Trimethylamine N-oxide
Nickel		Urea
Nitrate		(R)-1-Aminopropan-2-ol
Orthophosphate		1,3-Diaminopropane
Putrescine		1-Aminopropan-2-ol
sn-Glycerol 3-phosphate		1-Butanol
Sodium		
Spermidine		
Sulfate		
Taurine		
Teichoic acid		
Tetrabenazine		
Thiamin		
Thiosulfate		
Tungsten		
Urea		
Vitamin B12		
Zinc		
2,6-Dimethoxybenzoquinone		
2-(beta-D-Glucosyl)-sn-glycerol		

Table 2: List of compounds preferably used as seed compounds, sorted alphabetically

Organism	file name
Buchnera aphidicola	seedcluster_BUC.eps
Escheria coli	seedcluster_ECO.eps
Mycoplasma mobile	seedcluster_MMO.eps
Wolbachia	seedcluster_WOL.eps
Chlamydia pneumoniae	seedcluster_CPJ.eps
Homo sapiens	seedcluster_HSA.eps
Saccharomyces cerevisiae	seedcluster_SCE.eps
Rickettsia prowazekii	seedcluster_RPR.eps
Tropheryma whipplei	seedcluster_TWH.eps

Table 3: List of organisms for which metabolic resource graphs are provided.

5 Comparison to Metagrowth

The results of our analysis have been compared to the information of the Metagrowth database for the examples *Rickettsia prowazekii* (table 5) and *Tropheryma whipplei* (table 6). For that, metabolites for which a deficiency in their synthesis pathways has been noted in Metagrowth and seeds that have been found as essential in our calculations have been listed. The essential seed compounds have been taken from the organism specific resource graphs which can be found in the supplementary files "seedcluster_RPR.eps" and "seedcluster_TWH.eps". The lists show, that in 55% (*R. prowazekii*) and 78% (*T. whipplei*) of the cases our results and the Metagrowth data predict the same necessary metabolites. Certain metabolites were not predicted by our algorithm because they were either not necessary for synthesis of any compound in the target set or absent from the organism's metabolic network as dervied from KEGG.

Resource type	Member metabolites
Acetate	Acetate
Bases	Uracil, Adenine, Xanthine, Hypoxanthine, Thymidine, Guanine, Inosine
Biotin	Biotin, 7,8-Diaminononanoate, Dethiobiotin, 8-Amino-7-oxononanoate, 6-Carboxyhexanoate
CO2	HCO3-
D-Fructose	D-Fructose
D-Glucose	D-Glucose, beta-D-Glucose, alpha-D-Glucose
D-Mannose	D-Mannose
D-Ribose	D-Ribose
Folate	Tetrahydrofolate, 10-Formyltetrahydrofolate, Folate, 5,10-Methylenetetrahydrofolate, 5,10-Methenyltetrahydrofolate, 5-Formyltetrahydrofolate, 5-Methyltetrahydrofolate, Dihydrofolate, 4-Aminobenzoate, 2-Amino-4-hydroxy-6-hydroxymethyl-7,8-dihydropteridine, 2-Amino-4-hydroxy-6-(D-erythro-1,2,3-trihydroxypropyl)-7,8-dihydropteridine
Glycerol	Glycerol, D-Glyceraldehyde, D-Glycerate
Glycine	Glycine
Isopentenyl	Isopentenyl diphosphate, Dimethylallyl diphosphate, 1-Hydroxy-2-methyl-2-butenyl 4-diphosphate, 2-C-Methyl-D-erythritol 2,4-cyclodiphosphate
L-Alanine	L-Alanine, D-Alanine, Selenocysteine, Selenide, L-Selenocysteine
L-Arginine	L-Arginine
L-Asparagine	L-Asparagine
L-Aspartate	L-Aspartate
L-Cysteine	L-Cysteine
L-Glutamate	L-Glutamate, L-Glutamine
L-Histidine	L-Histidine, Urocanate, L-Histidinal, L-Histidinol
L-Isoleucine	L-Isoleucine, (S)-3-Methyl-2-oxopentanoic acid
L-Leucine	L-Leucine, 4-Methyl-2-oxopentanoate
L-Lysine	L-Lysine
L-Methionine	L-Methionine, Methanethiol
L-Ornithine	L-Ornithine
L-Phenylalanine	L-Phenylalanine, Phenylpyruvate, L-Arogenate, D-Phenylalanine, Prephenate
L-Proline	L-Proline
L-Serine	L-Serine
L-Threonine	L-Threonine
L-Tryptophan	L-Tryptophan, Indole
L-Tyrosine	L-Tyrosine, 3-(4-Hydroxyphenyl)pyruvate
L-Valine	L-Valine
Maltose	Maltose
Nicotinate	NAD+, NADH, NADPH, NADP+, Nicotinamide, Nicotinate, Nicotinamide D-ribonucleotide, Deamino-NAD+, Nicotinate D-ribonucleotide, Pyridine-2,3-dicarboxylate, Nicotinate D-ribonucleoside
Nucleoside	Adenosine, Deoxyguanosine, Deoxyadenosine, Deoxyinosine
OrganicAcids	2-Oxobutanoate, D-erythro-3-Methylmalate, 2-Hydroxybutanoic acid, 3-Methyl-2-oxobutanoic acid, 3-Hydroxy-3-methyl-2-oxobutanoic acid, (S)-2-Acetolactate, 2,3-Dihydroxy-3-methylbutanoate, (R)-2,3-Dihydroxy-3-methylbutanoate, 2-Acetolactate
Orthophosphate	Orthophosphate, sn-Glycerol 3-phosphate
Pantetheine	Pantetheine, Pantothenate, Dephospho-CoA, Dihydropteroate, Pantetheine 4'-phosphate, D-4'-Phosphopantothenate, (R)-4'-Phosphopantothenoyl-L-cysteine
Propanoate	Propanoyl-CoA, Propanoate, L-3-Amino-isobutanoate, 3-Hydroxypropanoate, (S)-Methylmalonate semialdehyde, (S)-3-Hydroxyisobutyrate, 3-Hydroxy-2-methylpropanoate, 2-Methyl-3-oxopropanoate, Propanoyl phosphate
Riboflavin	FAD, FADH2, FMN, Riboflavin, 6,7-Dimethyl-8-(1-D-ribityl)lumazine, 4-(1-D-Ribitylamino)-5-amino-2,6-dihydroxypyrimidine
Shikimate	3,4-Dihydroxybenzoate, 3-Dehydroquinate, Shikimate, 3-Dehydroshikimate, 5-Dehydroshikimate
Succinate	Succinate, Succinate semialdehyde, Fumarate
Sucrose	Sucrose
Sulfate	Sulfate, Hydrogen sulfide, Thiosulfate, Sulfite, Taurine
Thiamin	Thiamin diphosphate, alpha,beta-Dihydroxyethyl-TPP, 2-(alpha-Hydroxyethyl)thiamine diphosphate, Thiamin, Thiamin monophosphate, 5-(2-Hydroxyethyl)-4-methylthiazole, 4-Methyl-5-(2-phosphoethyl)-thiazole, 4-Amino-5-hydroxymethyl-2-methylpyrimidine, 4-Amino-2-methyl-5-phosphomethylpyrimidine, 2-Methyl-4-amino-5-hydroxymethylpyrimidine diphosphate
beta-Alanine	beta-Alanine, beta-Aminopropion aldehyde, 3-Oxopropanoate

Table 4: List of resource types and their equivalent compounds.

Compounds	Metagrowth	seed calculation	remarks
Amino acid	+	+	our method showed several aminoacids as essential
AMP/dAMP	+	+	
Biotin	+	+	
CoA	+	+	
FAD/Riboflavin-5-phosphate	+	+	
Glutathione	+	-	not in target set
GMP/dGMP	+	+	
HCO3	-	+	
Heme/Protoporphyrin	+	-	not in target set
Isopentenyl diphosphate	-	+	
Malonyl-CoA	+	-	not in target set
Mannose 6 phosphate	-	+	
N-Acetyl-D-mannosamine	-	+	
NAD+	+	+	
Phosphatidate/sn-Glycerol 3-phosphate/Fatty acid	+	o	sn-Glycerol 3-phosphate is the initial compound for the production of Phosphatidate and Fatty acid. Both can neither be produced nor are in the target set.
Propanoate	-	+	
Pyridoxine/Pyridoxal phosphate	+	-	not in target set
Pyruvate	+	+	
Ribose 5 phosphate	-	+	
S-Adenosyl-L-methionine	+	+	
Tetrahydrofolate/Folic acid/Dihydropteroate	+	+	
Thiamin diphosphate/Thiamin (vitamin B1)	+	+	
Thymidine/CMP/UMP	+	+	
Ubiquinone (Coenzyme Q)/Chorismate	+	-	not in target set

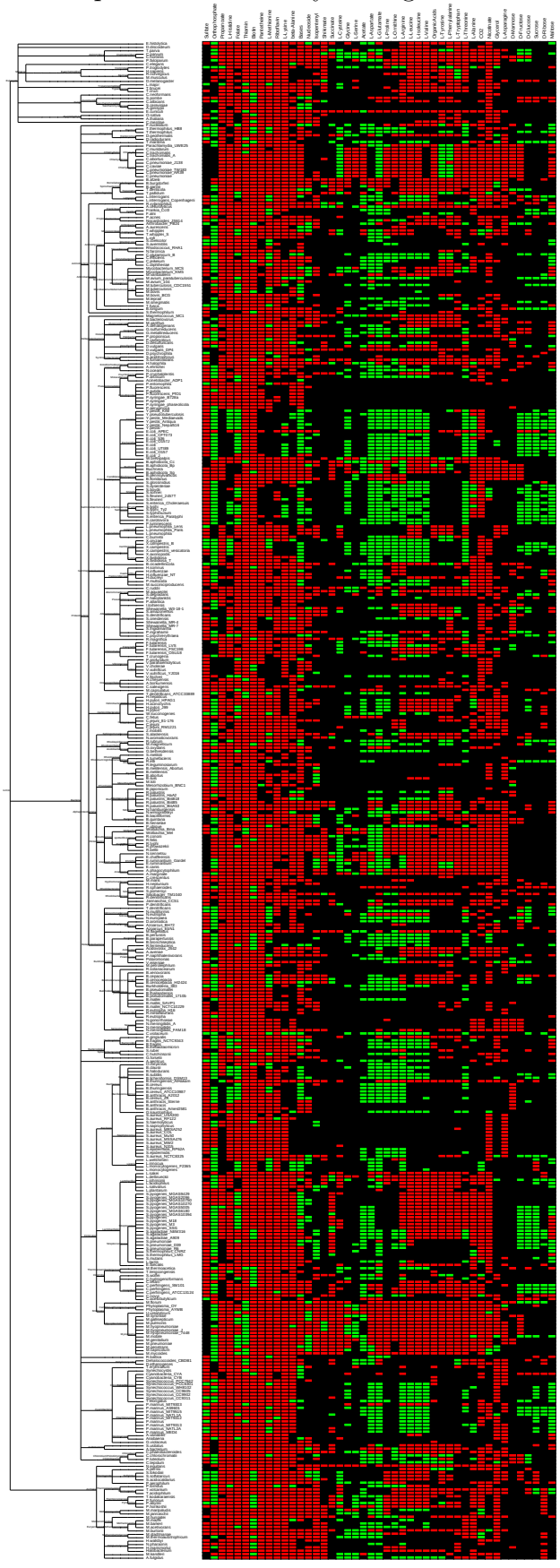
Table 5: Comparison of Metagrowth data to the calculations performed in this paper for *Rickettsia prowazekii* (RPR): "+" indicates a metabolite deficiency in Metagrowth or an essential seed compound in the calculations, "-" indicates a compound not identified as an essential nutrient and "o" indicates a potential, non essential nutrient.

Compounds	Metagrowth	seed calculation	remarks
1-Acyl-sn-glycerol 3-phosphatePhosphatidate	+	+	glycerol and sn-glycerol 3 phosphate predicted by seed calculation
beta-Alanine	+	+	
Biotin	+	+	
D-Glucose	o	+	not defined as deficiency in Metagrowth, but marked as primary energy source
D-Glutamate	+	o	generic compound "Glutamate" (i.e. D- or L-Glutamtate) found in seed calculation
Heme/Protoporphyrin	+	-	not in target set
L-Arginine	+	+	
L-Asparagine	+	+	
L-Aspartate	-	+	
L-Cysteine	+	o	
L-Glutamate	+	o	
L-Glutamine	+	o	
L-Histidine	+	+	
L-Leucine	+	+	
L-Lysine	+	+	
L-Methionine	+	+	
L-Phenylalanine	+	+	
L-Proline	+	+	
L-Tryptophan	+	+	
Malonyl-CoA	+	-	not in target set
Nicotinamide/Nicotinate	+	+	found in 2 different clusters in seed calculation
Propanoyl-Coa	-	+	
Pyridoxal phosphate/Pyridoxine	+	-	not in target set
(R)-Pantoate	+	+	Pantetheine found in seed calculation
Succinate	-	+	
Sugar phosphates	+	+	
Tetrahydrofolate	+	+	found in 2 different clusters in seed calculation
Thiamin diphosphate/Thiamin	+	+	
UDP-N-acetyl-D-glucosamine/N-Acetyl-D-glucosamine 1-phosphate	+	+	

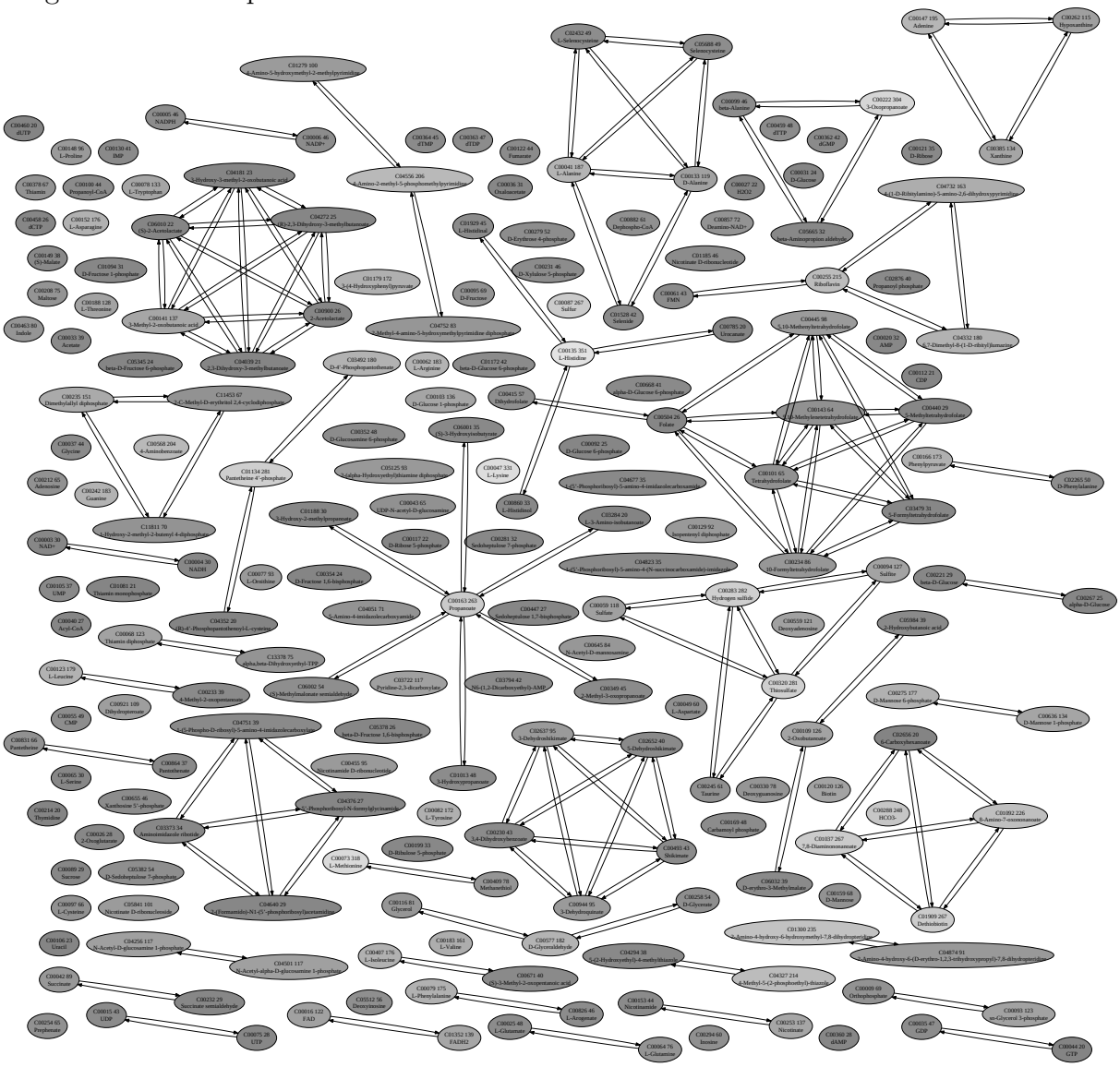
Table 6: Comparison of Metagrowth data to the calculations performed in this paper for *Tropheryma whipplei* (TWH): "+" indicates a metabolite deficiency in Metagrowth or an essential seed compound in the calculations, "-" indicates a compound not identified as an essential nutrient and "o" indicates a potential, non essential nutrient.

6 Supplementary Files

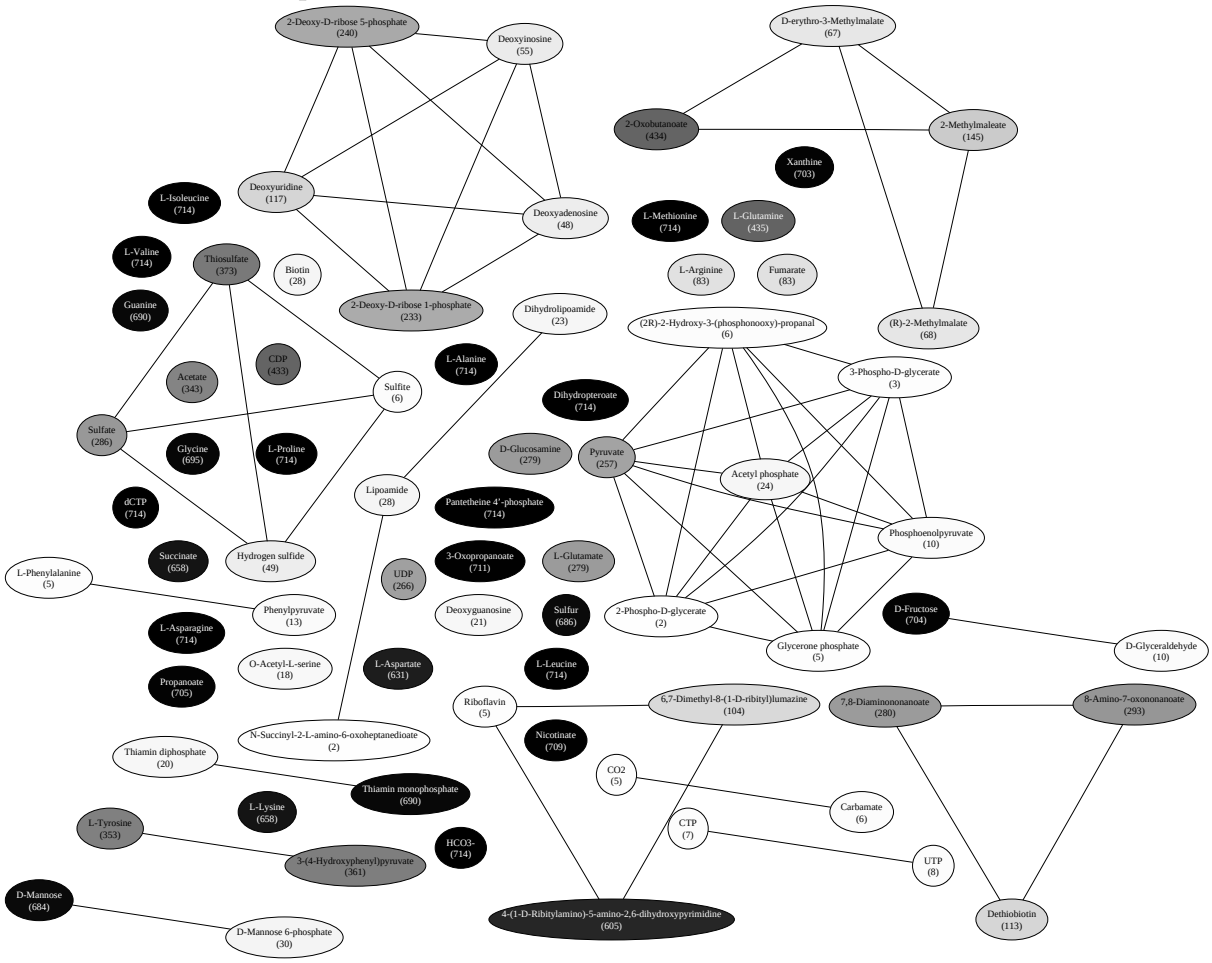
nutrient profiles of all analyzed organisms



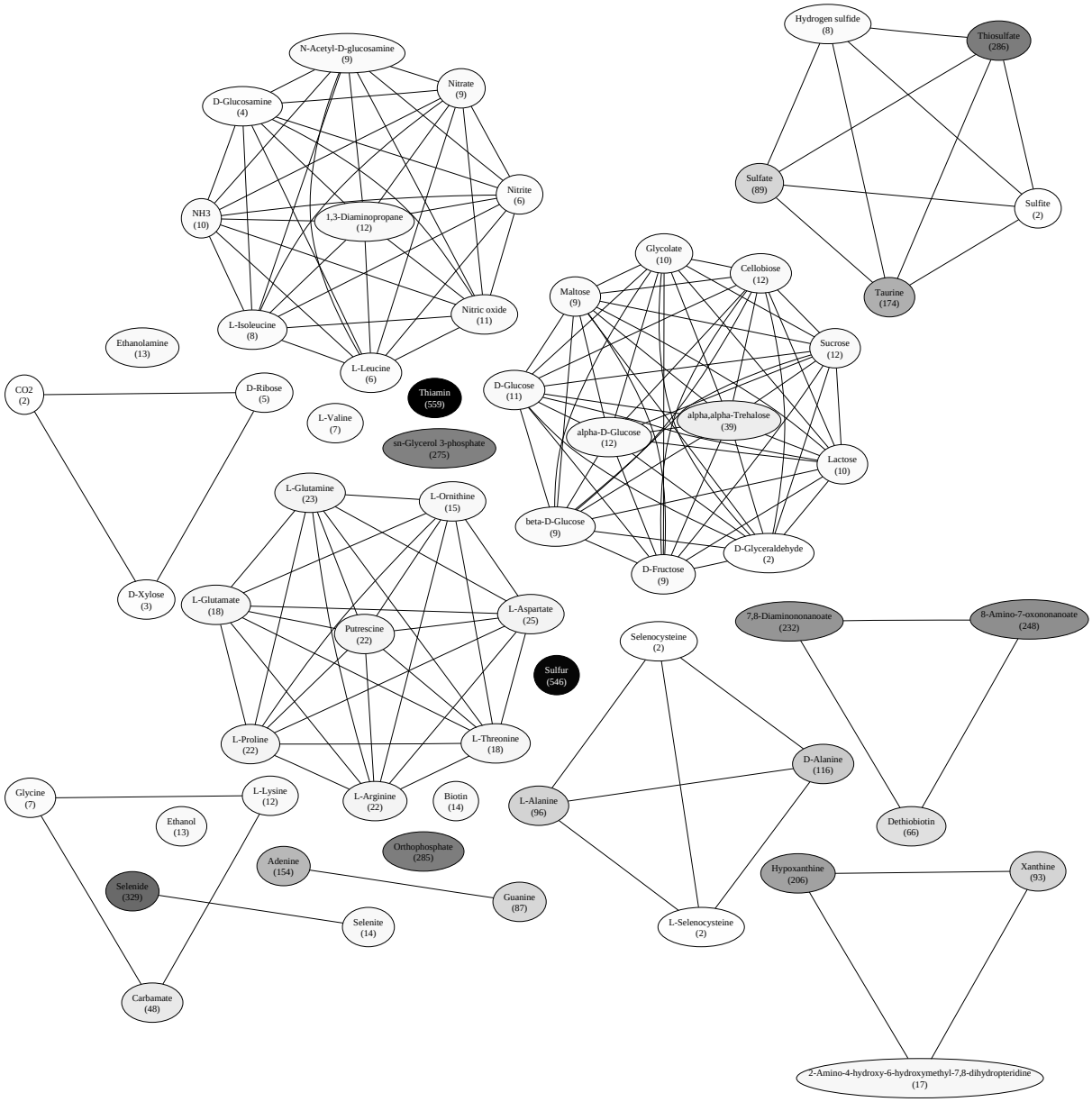
global_clusters.eps



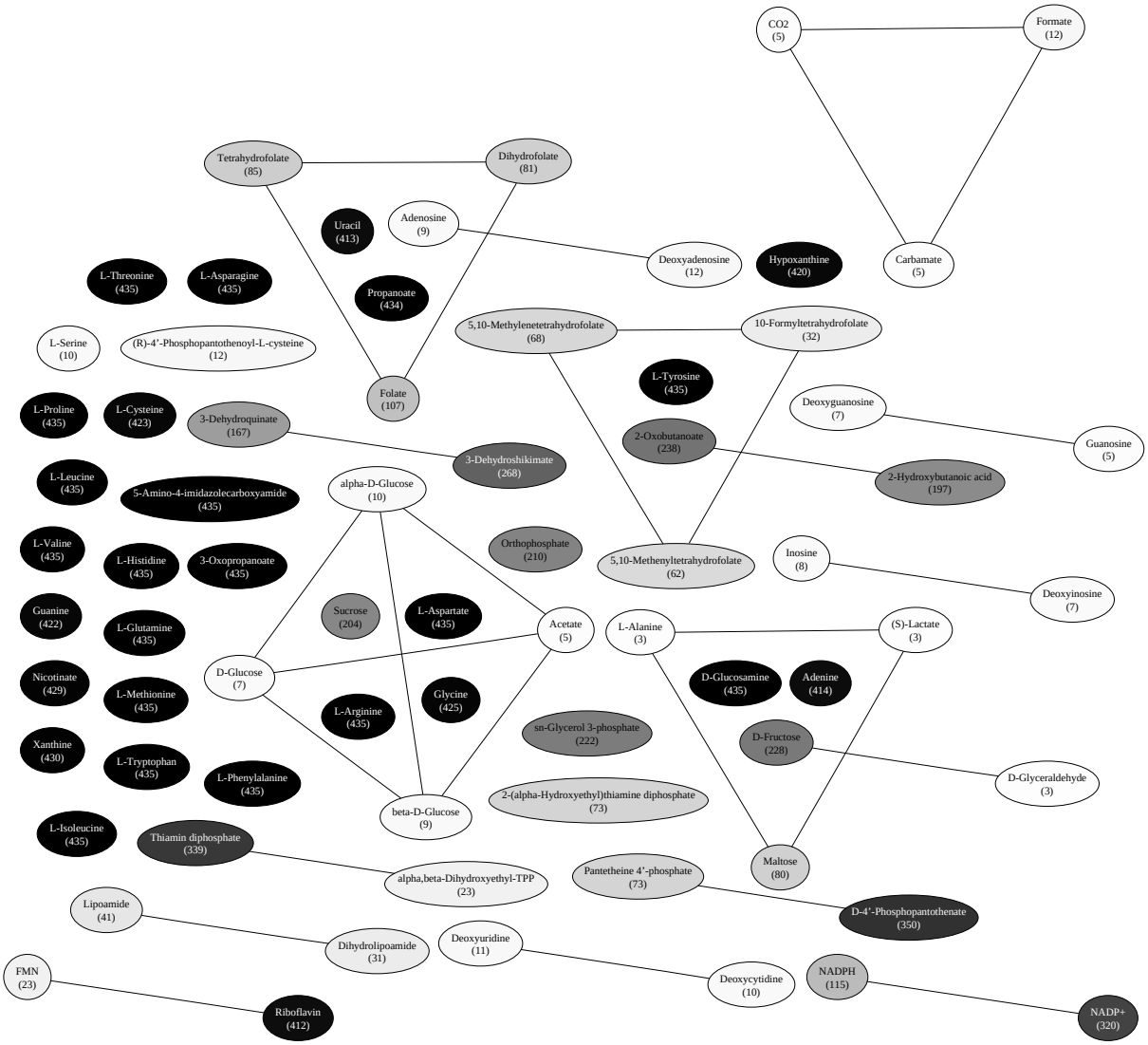
seedcluster_BUC.eps



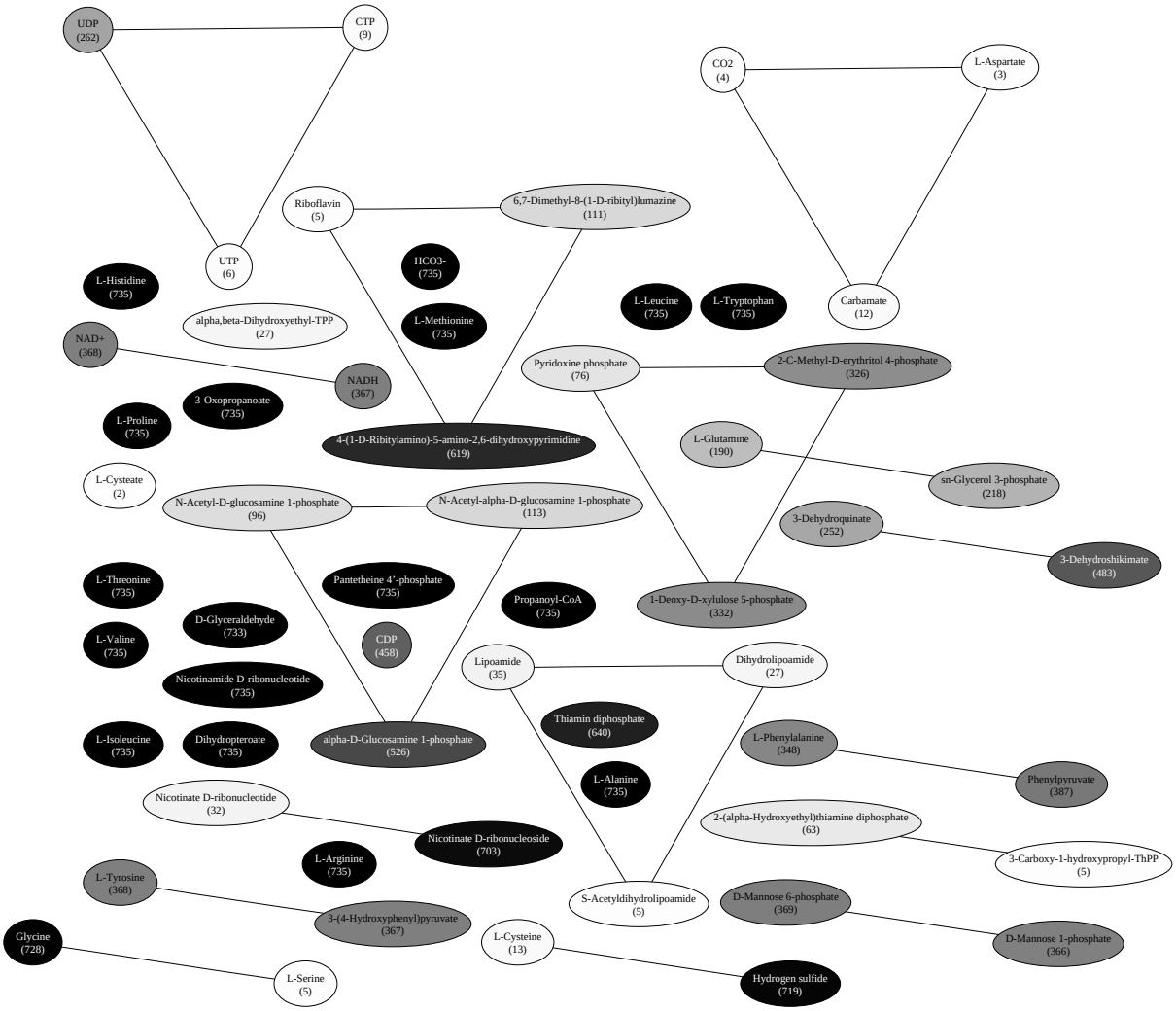
seedcluster_ECO.eps



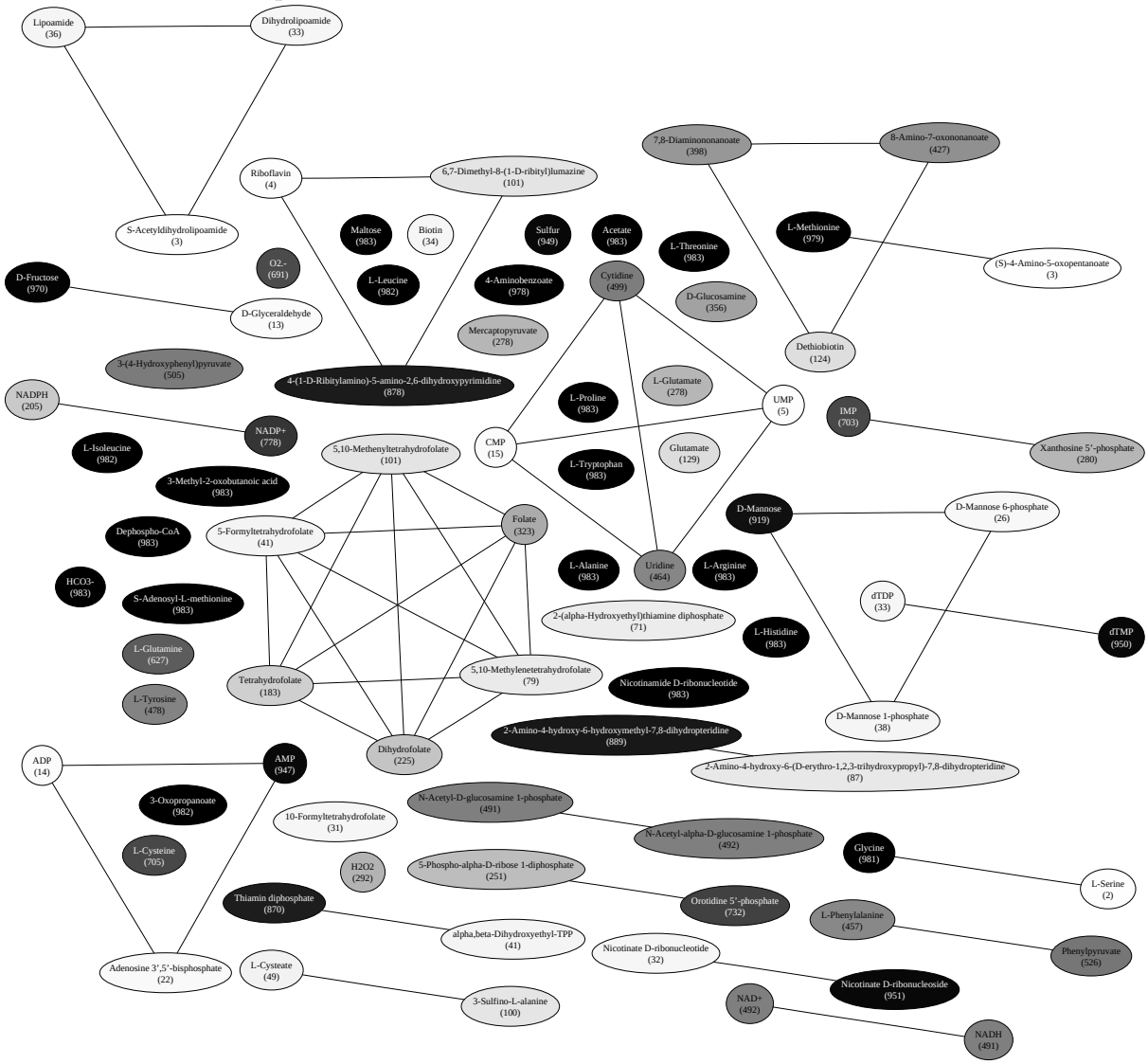
seedcluster_MMO.eps



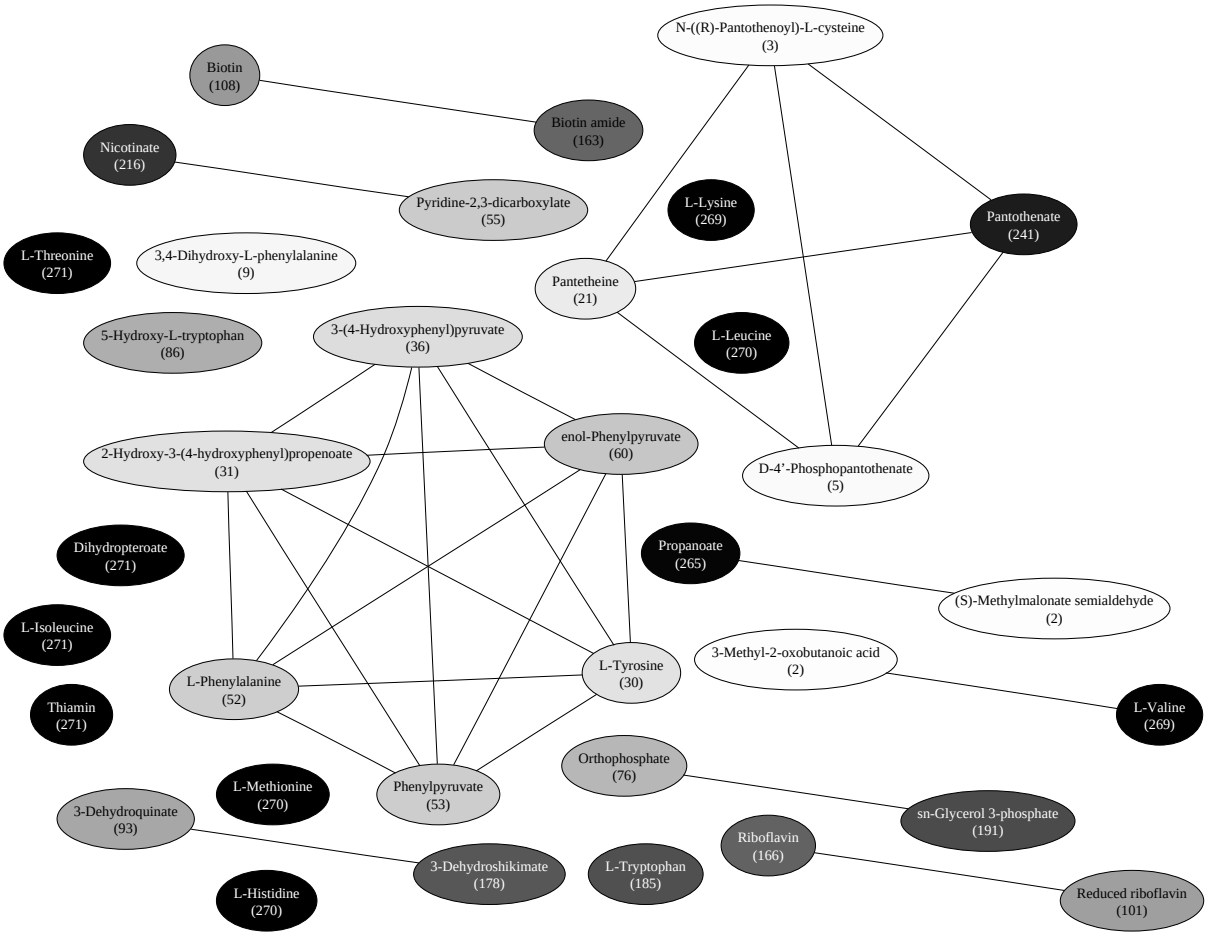
seedcluster_WOL.eps



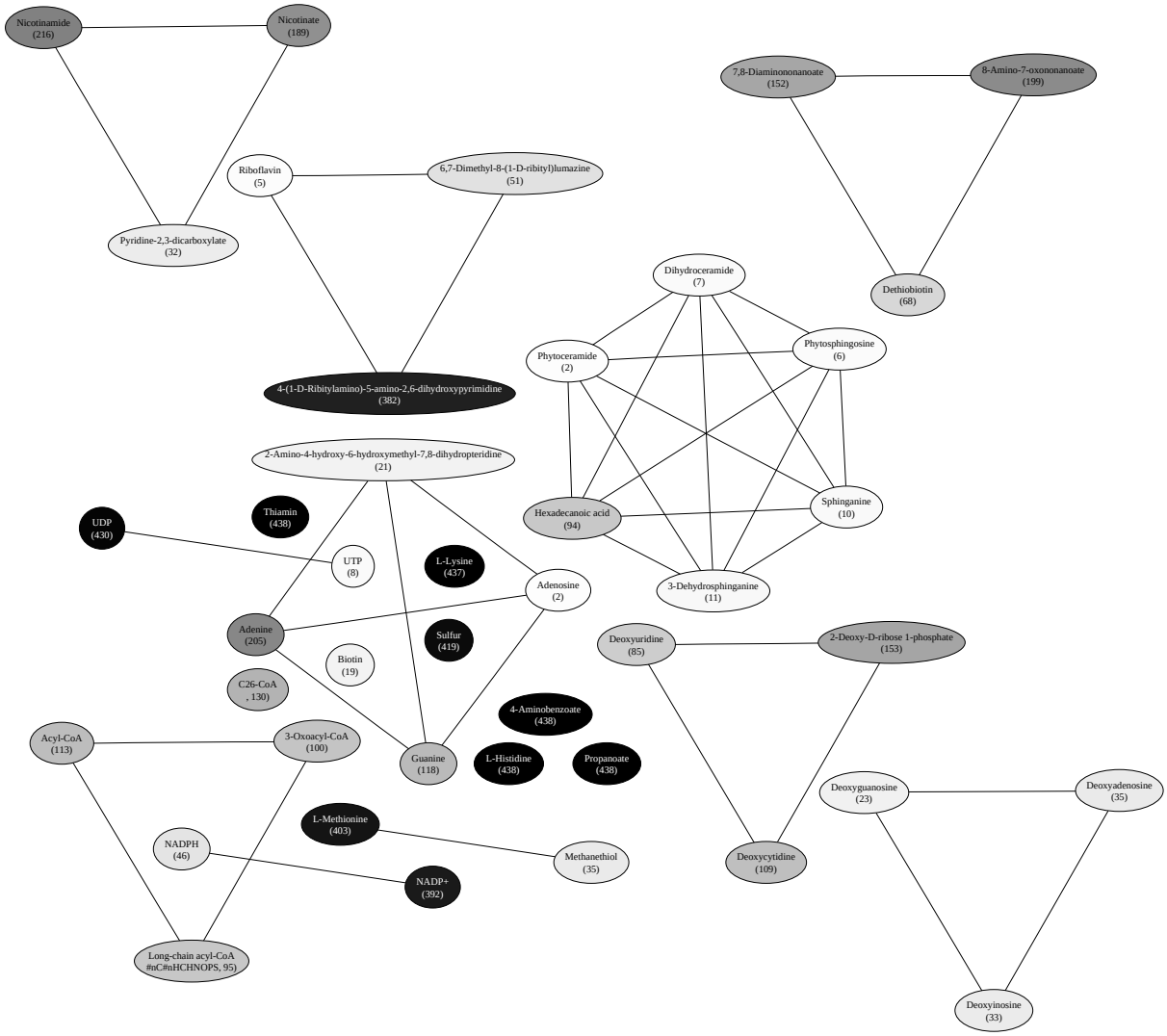
seedcluster_CPJ.eps



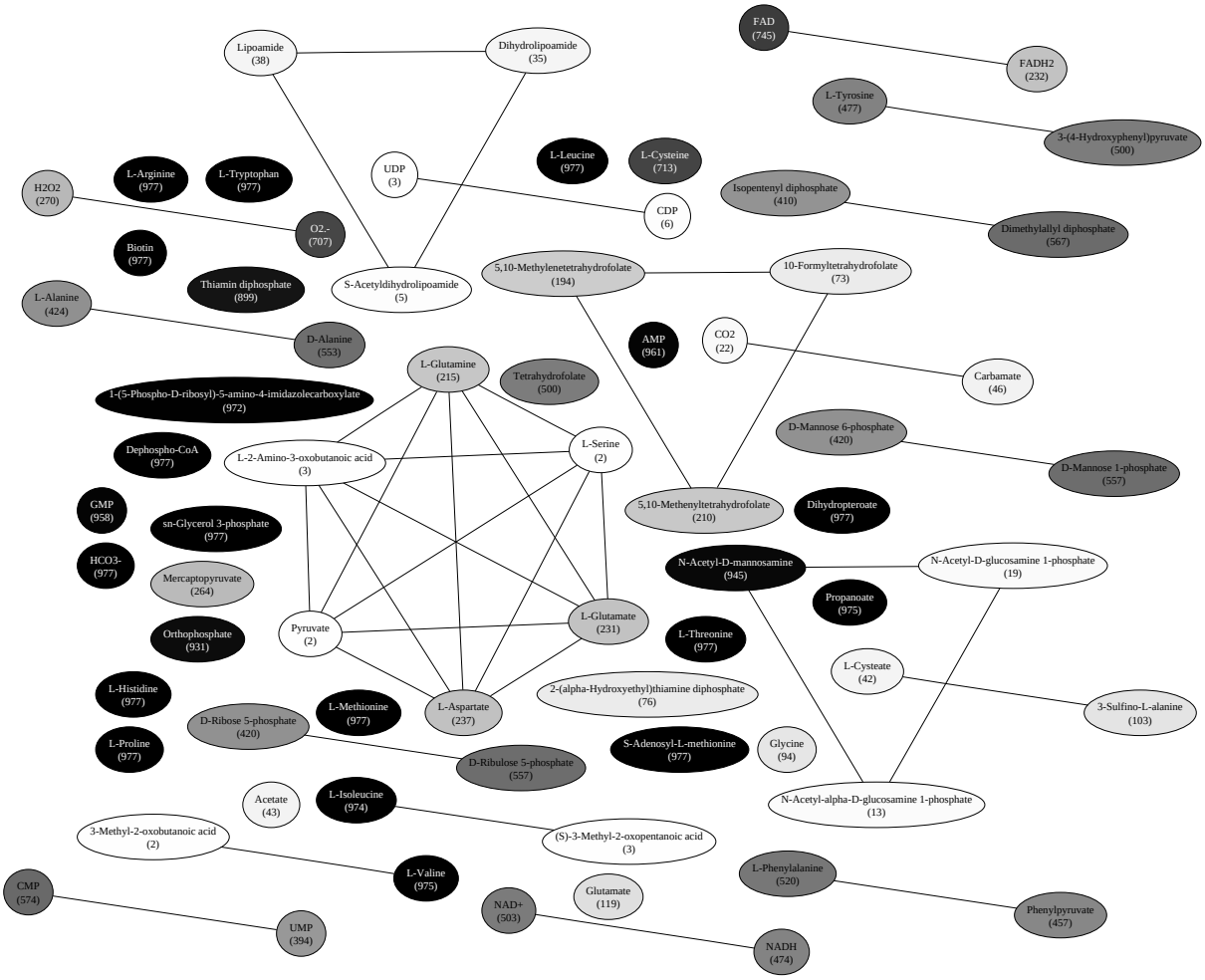
seedcluster_HSA.eps



seedcluster_SCE.eps



seedcluster_RPR.eps



seedcluster_TWH.eps