BIOSYNTHETIC POTENTIALS FROM SPECIES-SPECIFIC METABOLIC NETWORKS

GEORG BASLER[1],[2]
basler@mpimp-golm.mpg.de

ZORAN NIKOLOSKI[1],[2]
nikoloski@mpimp-golm.mpg.de

OLIVER EBENHÖH[1],[2]
ebenhoeh@mpimp-golm.mpg.de

THOMAS HANDORF[3]
handorf@mpimp-golm.mpg.de

[1] *Institute for Biochemistry and Biology, University of Potsdam, 14476 Potsdam, Germany*
[2] *Max Planck Institute of Molecular Plant Physiology, 14476 Potsdam, Germany*
[3] *Theoretical Biophysics, Humboldt-University Berlin, 10115 Berlin, Germany*

Studies of genome-scale metabolic networks allow for qualitative and quantitative descriptions of an organism's capability to convert nutrients into products. The set of synthesizable products strongly depends on the provided nutrients as well as on the structure of the metabolic network. Here, we apply the method of network expansion and the concept of scopes, describing the synthesizing capacities of an organism when certain nutrients are provided. We analyze the biosynthetic properties of four species: *Arabidopsis thaliana*, *Saccharomyces cerevisiae*, *Buchnera aphidicola*, and *Escherichia coli.* Matthäus *et al.* [12] have recently developed a method to identify clusters of scopes, reflecting specific biological functions and exhibiting a hierarchical arrangement, using the network comprising all reactions in KEGG. We extend this method by considering random sets of nutrients on well-curated networks of the investigated species from BioCyc. We identify structural properties of the networks that allow to differentiate their biosynthetic capabilities. Furthermore, we evaluate the quality of the clustering of scopes applied to the species-specific networks. Our study provides a novel assessment of the biosynthetic properties of different species.

Keywords: biosynthetic capabilities; clustering; scope; species-specific

1. Introduction

Recently, there has been tremendous interest in the comparison of metabolic network structures in order to quantitatively and qualitatively explain the organizational structure and identify possible intrinsic network design principles. While the research in this field historically concentrated on kinetic modelling of small parts of metabolism, *e.g.*, the glycolytic pathway [15], the emergence of biochemical databases, such as: KEGG [10], Brenda [11], and BioCyc [16], has prompted the interest for analyses of large-scale metabolic networks.

As kinetic data corresponding to genome-wide, species-specific metabolic networks are often difficult to obtain or precisely determine, novel, topology-based methods have been introduced in the last decade to allow a functional analysis of such networks. In particular, such networks have been investigated by graph-theoretic approaches [1, 18, 20], steady-state analysis, *e.g.*, elementary flux modes [17] or the related concept of extreme pathways [14], flux balance methods [5, 19], or, recently, by characterizing their synthesizing capacities using the concept of scopes [7].

The concept of a scope provides an effective method for determining which products a network can synthesize when it is provided with a given set of nutrient metabolites. In [8], it was shown that the synthesizing capacities of the nutrient metabolites, *i.e.*, their scopes, form a complex hierarchy in the species-independent network defined by the KEGG database. This hierarchy is mainly determined by the chemical composition of the metabolites—those with a larger number of chemical elements or chemical groups (and, therefore, with a larger scope) are placed on top of metabolites with a simpler composition.

In a recent paper [12], this complex hierarchy was condensed into a terse hierarchy of descriptive *consensus scopes* resulting from a clustering of scopes originating from all nutrient metabolites, taken individually. These consensus scopes represent sets of highly similar scopes, and could be assigned to characteristic combinations of chemical elements and a few chemical groups. As it is computationally impossible to calculate the synthesizing capacities of all nutrient combinations, the consensus scopes are useful to efficiently describe the biosynthetic potential of a given metabolic network.

Here, we investigate at which meaningful threshold values the formerly observed hierarchies and corresponding consensus scopes can also be found in species-specific networks. Our analysis comprises the metabolic networks of four model species: *Arabidopsis thaliana*, *Saccharomyces cerevisiae*, *Buchnera aphidicola*, and *Escherichia coli*, as defined in the BioCyc database. These species have been chosen as representatives of different domains of life and contrasting living environments. In particular, *Arabidopsis thaliana* (*abbr.* Arabidopsis, taxon 3702) is a eukaryotic multicellular CO_2 fixating plant, while *Buchnera aphidicola* (*abbr.* Buchnera, taxon 107806) is a highly specialized, intracellular parasite in aphids. *Escherichia coli* (*abbr.* E. coli, taxon 83333) is a well-studied bacteria that can grow in a variety of environments, and *Saccharomyces cerevisiae* (*abbr.* Yeast, taxon 4932) is a unicellular eukaryote and fungus that has been extensively used as a model organism.

Furthermore, we perform extensive analyses focused on the effect of different parameters on the outcome of the clustering approach. Finally, as the concept of scope strongly depends on the network structure, we discuss the influence of properties, characteristic for the investigated species-specific networks, on the scopes.

Organization and contributions: The methods employed in this study are presented in Section 2: The employed network representations and the scope algorithm are outlined in Subsections 2.1 and 2.2. In Subsections 2.3 - 2.5, the three main methods used in evaluating the influence of different parameters on the scope hierarchies, namely: the scope size distribution, (dis)similarity indices, and weighted modularity of a given clustering, are presented. The results from our analysis appear in Section 3, while discussion about the effect of the network properties on the investigated approach for determining a representative scope hierarchy is given in Section 4.

2. Methods

In this section, we describe the methods for testing the sensitivity of the approach proposed by Matthäus *et al.* [12] in order to investigate the biosynthetic potential of specific species. In Subsection 2.1, we detail the retrieval and representation of networks used in this study. The main method—calculation of the scope—is formally presented in Subsection 2.2. The size distributions of scopes on the investigated networks are discussed in Subsection 2.3, and the approach for determining the relationship between the parameters and methods for clustering is discussed in Subsections 2.4 and 2.5.

2.1. *Species-specific networks*

A *metabolic network* is typically represented by a directed bipartite graph $G = (V, E)$. The node set V of G can be partitioned into two subsets: V_r, containing *reaction nodes*, and V_m, comprised of *metabolite nodes*, such that $V_r \cup V_m = V$. The edges in E are directed either from a node $u \in V_m$ to a node $v \in V_r$, in which case the metabolite u is called a *substrate* of the reaction v, or from a node $v \in V_r$ to a node $u \in V_m$, when u is called a *product* of the reaction v. In the following, we refer to substrates as *predecessors* (*abbr.* pred), and products as *successors* (*abbr.* succ).

Such representation of a metabolic network can be retrieved from a publically available database of biochemical reactions. Here, the metabolic networks of the four investigated species were obtained from the BioCyc database [16]. Similarly to the network retrieval procedure specified in Matthäus *et al.* [12], the reactions were checked for consistency, and, consequently, those showing erroneous stoichiometry were removed. In addition, generic reactions and metabolites integrating sets of related metabolites were removed from the network, as proposed in [6]. The curation process was applied to the BioCyc database release from December 5, 2007, and resulted in networks of the following sizes: 1329 compounds and 1404 reactions (*Arabidopsis*), 1158 compounds and 1256 reactions (*E. coli*), 620 compounds and 594 reactions (*Yeast*), 356 compounds and 336 reactions (*Buchnera*).

The BioCyc database also provides information on the reversibility of biochemical reactions. Every enzymatic reaction (with a given direction), in principle, may also proceed in the reverse direction. However, the direction in which a reaction actually proceeds strongly depends on the metabolite concentrations, and may therefore vary for different physiological conditions. Thus, for analyzing the structure of a metabolic network from a given species, all reactions may be considered as being operable in both directions. Here, as a result, all reactions are assumed to be reversible. Hence, the network is represented by a bipartite graph $G = (V, E)$, where the successors and predecessors of a reaction are exchangeably considered as reactants or products.

2.2. *Biosynthetic potential of metabolites via scope*

Given a metabolic network G of an investigated species, the biosynthetic potential for a given set of metabolites, acting as substrates, can be described in terms of their scope, *i.e.*, the metabolites that can be synthesized in the network by the substrates.

The scope concept is related to reachability in the metabolic network G: A reaction node $v \in V_r$ is reachable if all of its substrates are reachable. Given a subset S of metabolite nodes, called a *seed*, a node $u \in V_m$ is reachable either if $u \in S$ or if u is a product of a reachable reaction. With these clarifications, we can present a precise mathematical formulation for the scope of a given seed [3]:

Definition 2.1. Given a metabolic network $G = (V, E)$ and a set $S \subseteq V_m$, the scope of the seed S, denoted by $R(S)$, is the set of all metabolite nodes reachable from S.

For a given metabolic network $G = (V, E)$ and a set $S \subseteq V_m$, the scope $R(S)$ can be determined in polynomial time of the order $O(|E| \cdot |V|)$, as can be established by analyzing the following algorithm:

Algorithm 1: Scope for a set of seed metabolites S in a metabolic network G

```
Input: Metabolic network $G = (V_m \cup V_r, E)$,
set of seed metabolites $S \subseteq V_m$
Output: Scope $R(S)$
mark all nodes in $V_r$ as unreachable and unvisited
$R(S) = S$
repeat
    if there is a reachable unvisited node $r \in V_r$ then
        mark $r$ as visited
        $R(S) = R(S) \cup pred(r) \cup succ(r)$
    end
    foreach node $r \in V_r$ do
        if $pred(r) \subseteq R(S)$ or $succ(r) \subseteq R(S)$ then
            mark $r$ as reachable
        end
    end
until no reachable unvisited nodes in $V_r$
```

In our analysis, the seed, S, is chosen uniformly at random from the set of metabolite nodes in a given network G. Algorithm 1 is then applied to each of $f = 3000$ sets S of a specified cardinality c. In the following, we describe how one can determine the distribution and clustering of scopes for a given cardinality, c, of the seed.

2.3. *Distribution of scope sizes*

Given a species X with a metabolic network represented by G_X, let Σ^c_X be the set of all scopes for f randomly chosen sets S, such that $c = |S|$. The scope size distribution for Σ^c_X gives the probability, $P^c_X(s)$, that a scope, randomly chosen from Σ^c_X, is of size s. The effect of the parameter c on the distribution $P(s)$ can be investigated by plotting the curves $P^c_X(s)$ for different values of c.

To investigate the (possible) difference in the scope size distribution for several species, the sizes of the scopes are normalized by the number of metabolites in the corresponding network for each species. The scope size distributions of the investigated species are analyzed in Subsection 3.1.

2.4. *Clustering of scopes*

Existing studies of biosynthetic potential [8, 12] have identified that a large number of metabolites do have scopes similar in size and metabolite composition. Here, we investigate this idea by hierarchical clustering for a set of scopes Σ^c_X generated from a seed with cardinality c and a given metabolic network of a species X.

Hierarchical clustering is based on a given distance (dissimilarity) matrix for the elements of Σ^c_X. Similar to [12], we employ the reversed Jaccard index as a distance measure for a pair of scopes, $R(S_i)$ and $R(S_j)$, $|S_i| = |S_j| = c$, $1 \leq i, j \leq f$. The computation is in the order of $O(|f|^2)$ for f scopes. For completeness, we give the definition of Jaccard distance, $J_{R(S_i)R(S_j)}$:

$$J_{R(S_i)R(S_j)} = 1 - \frac{|R(S_i) \cap R(S_j)|}{|R(S_i) \cup R(S_j)|}$$

We investigate the effect of a nearest neighbor group-average clustering algorithm [9]. Nearest neighbor clustering is a bottom up clustering method where iteratively clusters with increasing distance are joined, starting with clusters composed of single elements (scopes). Group-averaging refers to the method of defining the distance between two clusters as the average over all distances between pairs of the corresponding cluster elements.

The output of a hierarchical clustering algorithm is a tree, which can be cut at a given distance between the clusters, to retrieve the clusters of scopes. The clusters obtained from a cut at distance τ contain all scopes whose mutual distance is not greater than τ. The results of the clustering of scopes are presented in Subsection 3.2.

2.5. *Evaluation of parameter values*

To evaluate the influence of the size of the seed, c, and the distance, τ, at which the clustering tree is cut on the quality of the obtained clusters, we use weighted modularity [2]—a generalization of the graph cluster quality measure proposed by Newman and Girvan [13].

To apply graph cluster quality measures, one first has to build a graph from a given matrix of dissimilarity indices. Here, we construct a graph from the dissimilarity matrix by creating a node for each scope, with the distances between the scopes as weighted edges: let I be the dissimilarity matrix used in the hierarchical clustering. The weighted adjacency matrix A of the graph H is given by $1 - I_{R(S_i)R(S_j)}$, over all pairs $R(S_i)$ and $R(S_j)$ in Σ_X^{c}. The edges of graph H are then weighted by the similarity of the scopes S_i and S_j.

Let $\mathcal{C} = \{C_1, \ldots, C_p\}$ be the set of scope clusters obtained by cutting the clustering tree at distance τ. Given a graph H, with node set given by the f scopes and weighted edges as defined above, the modularity of $\mathcal{C}$ measures the quality of the clustering, or how separated nodes (scopes) from different clusters are from each other. It is defined as:

$$Q_{c,\tau} = \frac{1}{2m} \cdot \sum_{i,j=1}^{f} \left(A_{ij} - \frac{d(i)d(j)}{2m} \right) \delta_{ij},$$

where $m = \sum_{ij} A_{ij}$ is the weighted number of edges in H, A_{ij} is the element of the adjacency matrix in row i and column j, $d(i)$ is the weighted degree of scope i in H, $d(j)$ is the weighted degree of scope j in H, and $\delta_{ij} = 1$, if i and j are in the same cluster of $\mathcal{C}$, and 0, otherwise.

With regard to this definition, the modularity measure assesses the closeness of the scopes placed in the same cluster (according to the employed clustering algorithm) and their "distance" from the scopes placed in the other clusters with respect to the weighted adjacency matrix (*i.e.*, the similarity matrix).

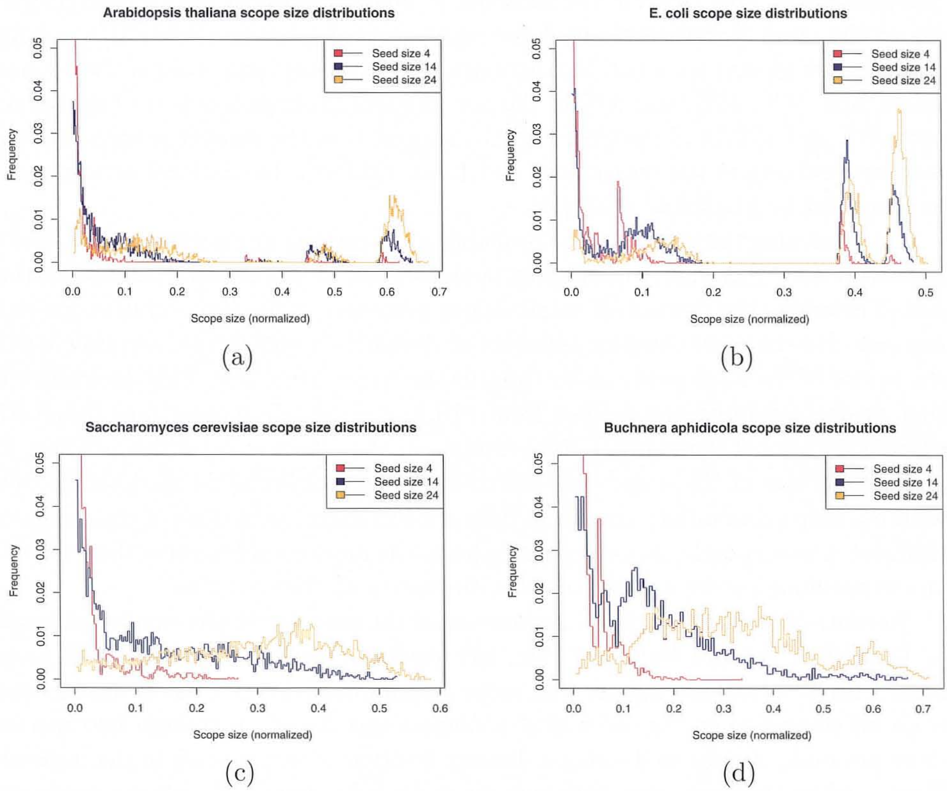
We investigate the behavior of the cluster quality for different sizes of the seed and different values for the parameter τ at which the clustering tree is cut to obtain the set of clusters $\mathcal{C}$ (see Subsection 3.3).

3. Results

Here, we analyze and compare the scope size distributions, cluster agglomeration, and weighted modularities of scope clusters, obtained from the networks of the four investigated species. The scope size distributions and cluster agglomeration reveal characteristic features of the networks, while the weighted modularities determined for different values of cut-off and seed size allow to systematically and quantitatively assess the relative influence of these parameters on the clustering.

3.1. *Scope size distributions*

Analyses of the scope concept have already identified that metabolites exhibit different biosynthetic potentials, *i.e.* the number of reachable metabolites strongly depends on the composition of the seed [3]. Therefore, we use the size of the scope to quantitatively characterize the biosynthetic potential of the seed metabolites in a given metabolic network. To this end, we empirically determine the size distributions of scopes resulting from the four investigated species (see Fig. 1). In order to enable comparability, the scope sizes were normalized by the size of the network, and the counts of scopes were turned into a probability distribution (see Subsection 2.3 for details).



(a)

(b)

(c)

(d)

Fig. 1. Scope size distributions of (a) *Arabidopsis*, (b) *E. coli*, (c) *Yeast* and (d) *Buchnera*, normalized by the number of metabolites in the corresponding network. The distributions are shown for seed sizes 4 (red), 14 (blue), and 24 (yellow). The highest frequencies for seed size 4 are excluded for clarity: $P^4_{Arabidopsis}(4) = 0.38$, $P^4_{E.\ coli}(4) = 0.35$, $P^4_{Yeast}(4) = 0.39$, and $P^4_{Buchnera}(4) = 0.38$.

We observe that with small seeds of four metabolites, the scope size distributions of all investigated networks share a high peak for very small scope sizes, indicating that a large number of seeds exhibit a very low biosynthetic potential. The remaining large isolated peaks in the networks of *Arabidopsis* (Fig. 1a) and *E. coli* (Fig. 1b) correspond to characteristic scopes reachable from a relatively large number of different seeds. These characteristic scopes correspond to large subnetworks with a high degree of mutually reachable metabolites, which we refer to as *scope communities*: If the seed contains metabolites from within such a scope community, then there is a high probability of reaching all the metabolites within the community. In addition, a scope community is self-contained in the sense that metabolites outside of the community can only be reached if the seed contains certain metabolites also outside of the community.

Note that although one characteristic peak may correspond to several such scope communities with a similar scope size, this is not observed in the networks of *Arabidopsis* and *E. coli.* Instead, the subsequent clustering reveals that scopes pertaining to the same characteristic peak are agglomerated into one cluster at a merging distance not greater than 0.2. Furthermore, the relatively large sizes of the communities (apx. 35%, 46%, and 60% of the network size in *Arabidopsis*, see Fig. 1a, and apx. 38% and 45% in *E. coli*, see Fig. 1b) suggest that the smaller scope communities form subsets of the larger ones and, thus, exhibit a hierarchical arrangement, as identified by Matthäus *et al.* [12].

By increasing the seed size, the probability of reaching any particular metabolite increases, and, therefore, one obtains larger scopes. In particular, we observe that for all networks the fraction of small scopes decreases, while the overall scope sizes increase. For the more complex networks of *Arabidopsis* and *E. coli*, we observe that the center of the large peaks shifts towards the larger scope size. This demonstrates that seeds containing metabolites from within a scope community now frequently contain additional metabolites from outside of the community, which account for a small increase of the scope size. Moreover, seeds containing no metabolites from within a scope community remain to have a small scope, regardless of the increased seed size. Consequently, scope communities in the more complex networks represent an outstanding feature that is robust with respect to the seed size.

In contrast to these findings, an increase of the seed size in the smaller networks of *Yeast* (Fig. 1c) and *Buchnera* (Fig. 1d) results in more evenly distributed scope sizes. This observation suggests that scope communities do not exist or are less pronounced compared to the cases of *Arabidopsis* and *E. coli.* For these two species, there are many scopes containing a distinct fraction of metabolites in the network. Finally, while the scope size distributions of *Arabidopsis* and *E. coli* are easily distinguishable by the frequency, relative scope size and number of scope communities, this is not the case for *Yeast* and *Buchnera.*

3.2. *Cluster agglomeration*

The dissimilarity matrix serves as the basis for the clustering described in Subsection 2.4. During the clustering process, scopes are agglomerated into clusters, starting with the most similar. At a merging distance of 0, every scope forms an individual cluster, so that the number of clusters equals the number of scopes f, *i.e.*, 3000. The number of clusters monotonically decreases with an increasing merging distance, until, at a distance of 1, all scopes form a single cluster.

The number of clusters obtained at a certain merging distance provides information on the overall mutual similarities between scopes. In the case of many highly similar scopes, a small number of clusters will be obtained for a small merging distance, while the opposite holds for the case of many dissimilar scopes. For instance, if at a distance of 0.5 the number of clusters is half the number of scopes, then more than half of the scopes have a mutual distance of at most 0.5; therefore, more than half of the scopes share at least two thirds of their metabolites with another scope (cf. Subsection 2.4).

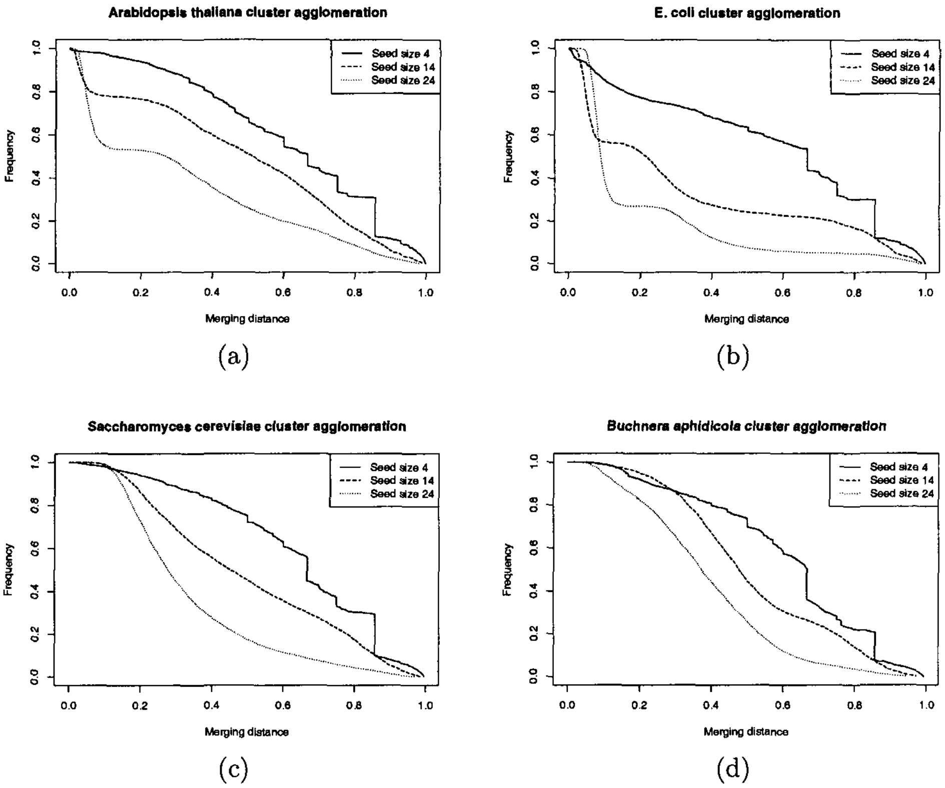


Fig. 2. Frequency of observed clusters over the merging distance for (a) *Arabidopsis*, (b) *E. coli*, (c) *Yeast* and (d) *Buchnera*. While steps appear in the frequencies for seed size of 4 (solid line) as a consequence of numerical effects of the Jaccard distance, the shapes appear continuous for seed sizes of 14 (dashed line) and 24 (dotted line). Furthermore, the overall mutual distances of scopes decrease when increasing the seed size, resulting in a smaller fraction of clusters at a particular merging distance.

As shown in Fig. 2, the mutual similarities of scopes exhibit significant differences when using varying seed sizes. As a trend, the number of clusters obtained at a certain merging distance is reduced with the increase of the seed size, demonstrating that more similar sc.opes result from a larger seed size. This conforms to the intuition, as larger seeds result in larger scopes with a higher probability of sharing common metabolites.

While the agglomeration curves from seed sizes 14 and 24 appear continuous, steps appear in the curves from seed size 4. For the latter curves, a large number of scopes is agglomerated into clusters at certain distances. For *Arabidopsis* (Fig. 2a) and *E. coli* (Fig. 2b), there are large steps of more than 160 scopes at characteristic distances of 2/3 and 3/4, and steps of more than 530 scopes at a distance of 6/7. In *Yeast* (Fig. 2c) and *Buchnera* (Fig. 2d), there are steps of more than 300 scopes at distances of 2/3 and 6/7.

These are numerical effects of the Jaccard distance which provides a discrete number of possible dissimilarity values, decreasing with smaller cardinalities of the compared entities. When using a small seed size, the fraction of small scopes is very large (*cf.* Subsection 3.1). Consequently, for a large number of scopes there is a small number of possible distances to consider. For instance, at a distance of 2/3, all scopes of size four with two metabolites in common are merged, and all scopes of size six with three metabolites in common, and so on. With many small scopes, these characteristic distances occur more frequently, leading to the observed steps.

For the clustering of *Arabidopsis* and *E. coli* with seed sizes of 14 and 24, a significant fraction of scopes is agglomerated with a merging distance of less than 0.1. This indicates that there are many scopes with a high mutual similarity. In contrast, this does not hold for *Yeast* and *Buchnera*, where the range of similarities between scopes is more uniformly distributed and, thus, results in cluster agglomerations at higher distances. Again, there are significant differences between the calculated scopes of *Arabidopsis* and *E. coli* on one hand, and *Yeast* and *Buchnera*, on the other hand.

3.3. *Influence of cut-off and seed size*

Due to the observed large impact of the employed seed size and cut-off on the calculated scopes and the resulting clustering, we aim at evaluating the influence of these parameters on the quality of clustering. Particularly, we are interested in those parameter values that allow to obtain clusters of highest weighted modularity. Moreover, thorough investigation of the parameter space may provide insights in the presented approach of scope clustering.

We determine scopes from random seeds as described in Subsection 2.2 for seed sizes $2 \leq c \leq 25$. For each set of scopes resulting from a given network and seed size, we perform the clustering of scopes as described in Subsection 2.4. Finally, we cut the obtained cluster trees at cut-off distances $0.05 \leq \tau \leq 1$ with step-size of 0.05, and determine the weighted modularities of the resulting sets of clusters, as defined in Subsection 2.5.

In Fig. 3, the resulting matrices of weighted modularities for different parameter

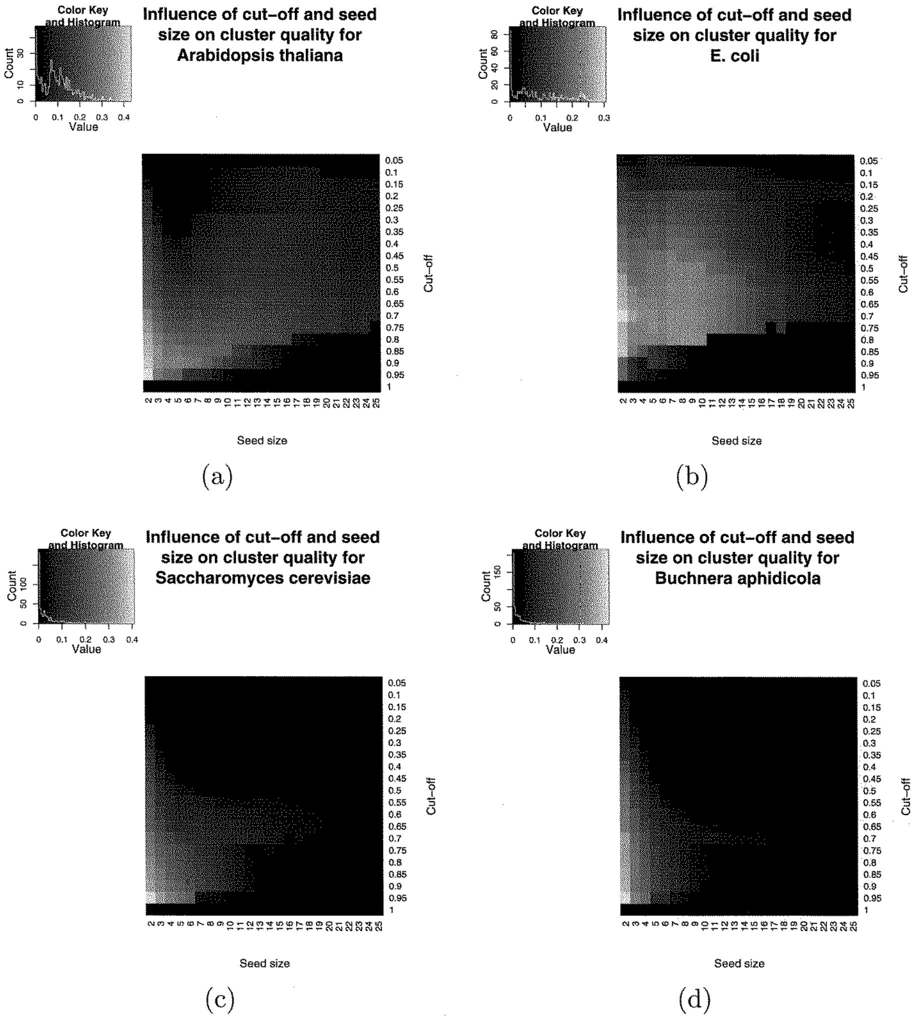


Fig. 3. Heatmaps of the weighted modularities using different seed sizes and cut-offs, for (a) *Arabidopsis*, (b) *E. coli*, (c) *Yeast* and (d) *Buchnera*. Histograms of the obtained values are shown in the top-left corners. The best cluster qualities are obtained using a seedsize of 2 and cut-off 0.7 for *E. coli*, and seed size 2 and cut-off 0.95 for the other species.

values are shown as heatmaps with corresponding value histograms. The lowest weighted modularities for all species are slightly below 0 and correspond to a cut-off distance of $\tau = 1$. This supports the intuition that a low value for the modularity should be obtained from an apparently poor clustering. The highest values differ for all species: for *Arabidopsis* (Fig. 3a) $Q_{c=2,\tau=0.95} \approx 0.43$, for *E. coli* (Fig. 3b) $Q_{c=2,\tau=0.7} \approx 0.31$, for *Yeast* (Fig. 3c) $Q_{c=2,\tau=0.95} \approx 0.41$, and for *Buchnera* (Fig. 3d) $Q_{c=2,\tau=0.95} \approx 0.43$. However, these maxima correspond to identical parameters of $c = 2$ and $\tau = 0.95$ in *Arabidopsis*, *Yeast* and *Buchnera*, while the modularity obtained from the same parameters in *E. coli* is $Q_{c=2,\tau=0.95} \approx 0.18$.

The evaluation of parameters indicate that the best clustering is achieved for a small seed size of $c = 2$ and a very high cut-off of $\tau = 0.95$ for all species except *E. coli*, for which $\tau = 0.7$ results in the highest cluster quality. The preference for small seeds demonstrates that small sets of metabolites can be well classified into distinct groups according to their biosynthetic potential using the concept of scopes. On the other hand, our analysis suggests that scopes from more complex seed compositions are harder to classify. Furthermore, the selection of a cut-off value $\tau = 0.95$ indicates that a small number of large clusters, containing scopes up to a very high distance, is preferred. Hence, the arrangement of scopes from small seeds into few very coarse groups results in the highest separation of clusters.

4. Discussion

Characterizing the biosynthetic potential by only employing the structure of metabolic networks offers a means for comparing and contrasting different species. Here, we investigated to what extent the approach proposed in [12] could be extended to determining scope clusterings and metabolite hierarchies in species-specific networks. To this end, we performed a comprehensive sensitivity analysis of the approach, which depends on the size and composition of random seeds and the cut-off distance for extracting clusters of scopes. The analysis furthermore includes the effect of the size and composition of random seeds on the scope size distributions in the four investigated species.

The findings related to the scope size distributions conform to the existing results on species-specific networks [4] as well as the network comprising all reactions from KEGG [12], *i.e.*, a large number of seeds exhibit a small biosynthetic potential. Accordingly, we observe characteristic scope sizes corresponding to scope communities for *Arabidopsis* and *E. coli*, which indicates the existence of consensus scopes and supports their hierarchical arrangement. This argument can be further strengthened by our findings regarding the scope size distributions of *Arabidopsis* and *E. coli*: With an increase of seed sizes, the overall scope sizes increase, while preserving the scope community structure.

The results from the agglomerative clustering performed on the scopes of 3000 randomly chosen seeds of different sizes suggest a plateau for the fraction of clusters at a merging distance around 0.2, *i.e.*, no significant number of scopes is agglomerated at distances close to 0.2. This is typically pronounced for seeds of larger sizes from the networks of *Arabidopsis* and *E. coli.* We point out that the phenomenon of plateau was already observed elsewhere [12] and was used as a principle for choosing a threshold in the extraction of scope clusters and the resulting metabolite hierarchies.

However, our analysis warrants caution when extending these observations to the networks of *Yeast* and *Buchnera*: While *Arabidopsis* and *E. coli* are organisms with complex metabolic networks, the opposite holds for *Buchnera.* Although *Yeast* is a generalist model organism with complex metabolic functions, its scope size distribution does not exhibit characteristic peaks and, therefore, does not contain any distinct scope communities. Likewise, there is no plateau observed in the cluster agglomeration of *Yeast.* The observed differences in scope sizes and clustering between *Arabidopsis* and *E. coli* on one hand, and *Yeast* and *Buchnera* on the other hand may be due to either differing qualities in the curation of the networks or a possible realistic difference in the biosynthetic potential of these species.

To further assess the quality of scope clusterings, we applied a generalization of the modularity measure. While for certain values of parameters (*i.e.*, cut-off distance and seed size) we obtained relatively high modularities for the respective scope clustering, the observed values have significantly different implications: The highest value for the modularity in the investigated species was obtained at cut-off distances of 0.95 and 0.7, corresponding to a small number of clusters comprising scopes with a wide range of similarities. Moreover, for most cut-off distances, the highest modularity is reached for small seed sizes ($c = 2$), suggesting that the cluster agglomeration may be highly dependent on the discretization capacity of the employed Jaccard distance. We point out that the same empirical analysis was performed and comparable results were obtained using the Manhattan distance as a (dis)similarity measure. Therefore, we can conclude that the method for extracting scope clusters and metabolite hierarchies may be most appropriate to large scope sizes, most likely resulting from large seed sizes and complex networks, for which both, the plateau principle and the observed scope communities, are clearly pronounced.

To conclude, we identified features based on the concept of scopes, which allow for a structural comparison of different species, and indicate the existence of consensus scopes and metabolite hierarchies in *Arabidopsis* and *E. coli.* In addition, our sensitivity analysis revealed a strong influence of the evaluated parameter values in the quality of clustering. Future research may aim at characterizing the scope communities via their metabolite compositions and hierarchical organization, and extending the analysis to additional organisms.

References

[1] Barabasi, A. L. and Albert, R., Emergence of scaling in random networks. *Science*, 286:509–512, 1999.

[2] Brandes, U., Delling, D., Gaertler, M., Görke, R., Hoefer, M., Nikoloski, Z. and Wagner, D., On modularity clustering. *IEEE Trans. Knowl. Data Eng.*, 20(2):172–188, 2008.

[3] Ebenhöh, O., Handorf, T., and Heinrich, R., Structural analysis of expanding metabolic networks. *Genome Informatics*, 15:35–45, 2004.

[4] Ebenhöh, O., Handorf, T., and Heinrich, R., A cross species comparison of metabolic network functions. *Genome Informatics*, 16(1):203–213, 2005.

[5] Edwards, J.S. and Palsson, B.O., The escherichia coli mg1655 in silico metabolic genotype: Its definition, characteristics, and capabilities. *Proc Natl Acad Sci USA*, 97:5528–5533, 2000.

[6] Feist, A. M., Henry, C. S., Reed, J. L., Krummenacker, M., Joyce, A. R., Karp, P. D., Broadbelt, L. J., Hatzimanikatis, V., and Palsson, B. O., A genome-scale metabolic

reconstruction for escherichia coli k-12 mg1655 that accounts for 1260 orfs and thermodynamic information. *Mol Syst Biol.*, 3(121), 2007.

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

[8] Handorf, T., Ebenhöh, O., Kahn, D., and Heinrich, R., Hierarchy of metabolic compounds based on their synthesizing capacity. *IEE Proc. Systems Biology,* 153(5):359–363, 2006.

[9] Hastie, T., Tibshirani, R., and Friedman, J., *The elements of statistical learning: Data mining, inference and prediction.* Springer, New York, 2001.

[10] Kanehisa, M., Goto, S., Hattori, M., Kinoshita, K.F., Itoh, M., Kawashima, S., Katayama, T., Araki, M., and Hirakawa, M., From genomics to chemical genomics: new developments in kegg. *Nucleic Acids Res.*, 34:D354–357, 2006.

[11] Karp, P.D., Ouzounis, C.A., Moore-Kochlacs, C., Goldovsky, L., Kaipa, P., Ahren, D., Tsoka, S., Darzentas, N., Kunin, V., and Lopez-Bigas, N., Expansion of the biocyc collection of pathway/genome databases to 160 genomes. *Nucleic Acids Research,* 19:6083–6089, 2005.

[12] Matthäus, F., Salazar, C., and Ebenhöh, O., Biosynthetic potentials of metabolites and their hierarchical organization. *PLoS Comput Biol,* 4(4):e1000049, Apr 2008.

[13] Newman, M. E. J. and Girvan, M., Finding and evaluating community structure in networks. *Physical Review E,* 69(026113), 2004.

[14] Price, N.D., Reed, J.L., Papin, J.A., Wiback, S.J., and Palsson, B.O., Network-based analysis of metabolic regulation in the human red blood cell. *Journal of Theoretical Biology,* 225:185–194, 2003.

[15] Rapoport, T. A., Heinrich, R., and Rapoport, S. M., The regulatory principles of glycolysis in erythrocytes in vivo and in vitro. a minimal comprehensive model describing steady states, quasi-steady states and time-dependent processes. *Biochem J,* 154(2):449–469, Feb 1976.

[16] Schomburg, I., Chang, A., Ebeling, C., Gremse, M., Heldt, C., Huhn, G., and Schomburg, D., Brenda, the enzyme database: updates and major new developments. *Nucleic Acids Research,* 32:D431–D433, 2004.

[17] Schuster, S. and Hilgetag, C., On elementary flux modes in biochemical reaction systems at steady state. *J. Biol. Syst.*, 2:165–182, 1994.

[18] Strogatz, S. H., Exploring complex networks. *Nature,* 410:268–276, 2001.

[19] Varma, A. and Palsson, B.O., Metabolic flux balancing:basic concepts, scientific and practical use. *Bio/Technology,* 12:994–998, 1994.

[20] Wagner, A. and Fell, D. A., The small world inside large metabolic networks. *Proc. R. Soc. Lond. B,* 268:1803–1810, 2001.