Final project report

Zhifeng Yang

11/21/2020

Abstract

Network has been developed as a common method in microbial ecology nowadays. Metabolic network, which describes how metabolites are utilized and transformed inside or around cells, is the key to interpreting the unique functions and traits of each species. But it remained unclear how two or more species with different metabolic network interacted when forming a community. Previous paper has provided methodology of how to analyze the metabolic competition and complementarity of two species. In this project, I will follow the methodology and use the public data to study the potential interaction of species with known metabolic network from KEGG. It is also a potential direction for me to apply this methodology to a real community to see how this potential metabolic interaction influence the correlation in relative abundance of species.Current analysis has show that I could build a metabolic network for known species and calculate the indexes for metabolic competition and complementarity. And I have perfomed this analysis to all species in KEGG database. The result shows the middle level of competition commonly exist while the complementary relationship is at a low level.

Introduction

Network has been developed as a common method in microbial ecology nowadays. Metabolic network, which describes how metabolites are utilized and transformed inside or around cells, is the key to interpreting the unique functions and traits of each species. Interaction network, which describes the positive or negative interaction between paired species, determines the dynamics of microbial communities. Basically, what determined the interaction of microbes is the traits of microbes. For example, the negative interaction may result from competition for shared food sources of two species, which could share some nodes pointing to the main metabolic network. On the contrary, the positive interaction may result from mutualism of two species, one of which may provide food to another. So metabolic network can somehow imply the interaction between species.

Ideally, if we know all metabolic networks of a microbial community, which may have over thousands of species, we can build an interaction network based on edges between two metabolic networks of paired species. However, the real interaction network may be much different due to a lot of issues. For example, the metabolic network doesn't involve all microbial processes such as reproduction and death, which may influence the interaction. Besides, the interaction network generated may not reflect the real interaction, which is hard to be validated in experiment. In this project, I want to study how metabolic network can imply interactions of species and valid this with interaction network.

The work is based on the paper published in 2013 (Levy and Borenstein 2013). The paper test if the metabolic-network-informed competition or complementarity could explain the co- occurrence of species. The Fig. 1 shows how metabolic network informed competition or complementarity.

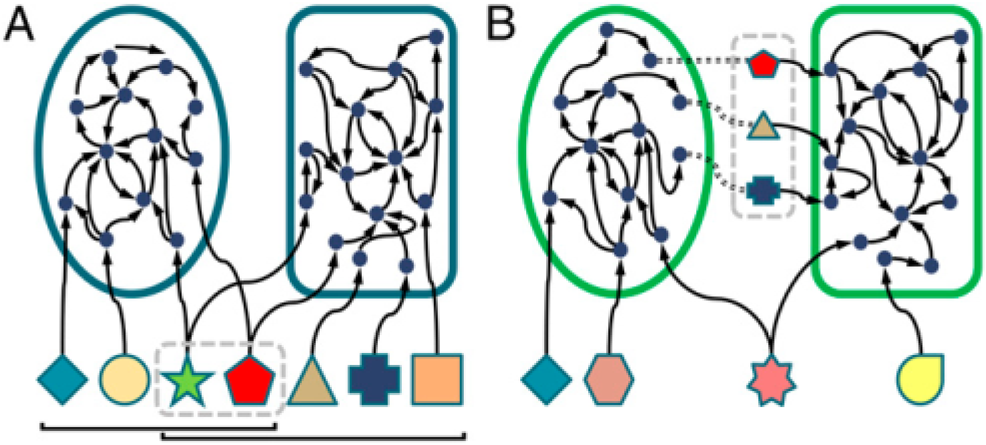


Fig. 1 The big circle and rectangle represents metabolic network while small icons represent metabolites which are also the nodes in network. The edges represent the pathway. The A graph shows the competition for food (the green and red icons), while B graph shows complementarity because the left cell provides some metabolites to right cell (Levy and Borenstein 2013).

Data preprocessing steps

The data processing is done by R as I mainly used it for my daily work. There are several steps to fulfill my research object:

1. Collect metabolic network from KEGG and input it into R.
2. Simplify the individual network based on the method in reference paper (Borenstein, Kupiec et al. 2008).

- Simplification is necessary for determine the seed set(required nutrient) and product set in the following step.

3. For the simplified network, determine the seed set, which is defined as the nodes with only outdegrees, and the product set, which is defined as the left nodes.
4. Calculated the competition index and complementarity index. The competition index, which is the $\frac{seed\ set(A \cap B)}{seed\ set(A)}$ or $\frac{seed\ set(A \cap B)}{seed\ set(B)}$, shows the competition pressure on nutrients from B->A or A->B. The complementarity index, which is the $\frac{seed\ set(A) \cap product\ set(B)}{seed\ set(A)}$ or $\frac{seed\ set(B) \cap product\ set(A)}{seed\ set(B)}$, shows how much the required nutrient could be supplied from B->A or A->B.

5. Calculate the metabolic interaction indexes for all microbes with available information in KEGG database.

6. build a microbial interaction network using abundance correlation.

7. test if metabolic network has some prediction of the structure of microbial interaction network based on correlation of abundance.

Ideas implemented up until the point

I mainly fulfilled the step 1-5.

Preliminary results

Step 1: download the metabolic network of E. Coli

Escherichia coli (E. coli) is a bacteria that normally lives in the intestines of both healthy people and animals. It is also a model species in microbial research. Here, I use it as an example to show the metabolic network of this species.

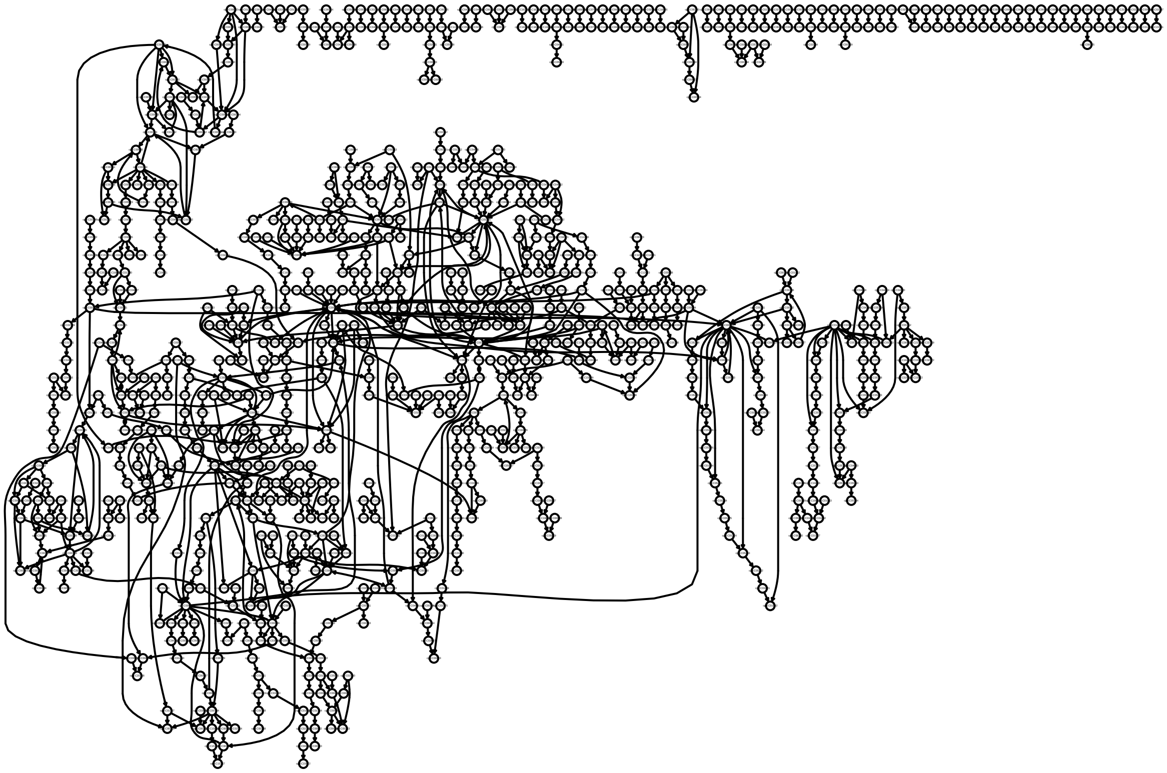


Fig. 2 The whole reaction network of E.coli. Each node is one metabolite. Each edge is a reaction. The graph is directed because reaction has substrates and products.

We can see that the network is a directed network. Some compounds could be synthesized by other compounds while others not. So the compounds, which can't be synthesized, can only be gained from outside environment. These are what we defined as seed set. To obtain the seed set, we need contract the network based on strongly connected components. For example, if the nodes shape a ring, which means every compounds could be synthesized by others, the species still need one of the ring as substrate. This is why we need to replace the ring with only one node to test if the group can be synthesized or not by other compounds.

Step 2: obtain the strongly connected components of the whole network

To simplify the metabolic network, I need to find the strongly connected components in the metabolic network of E.coli.

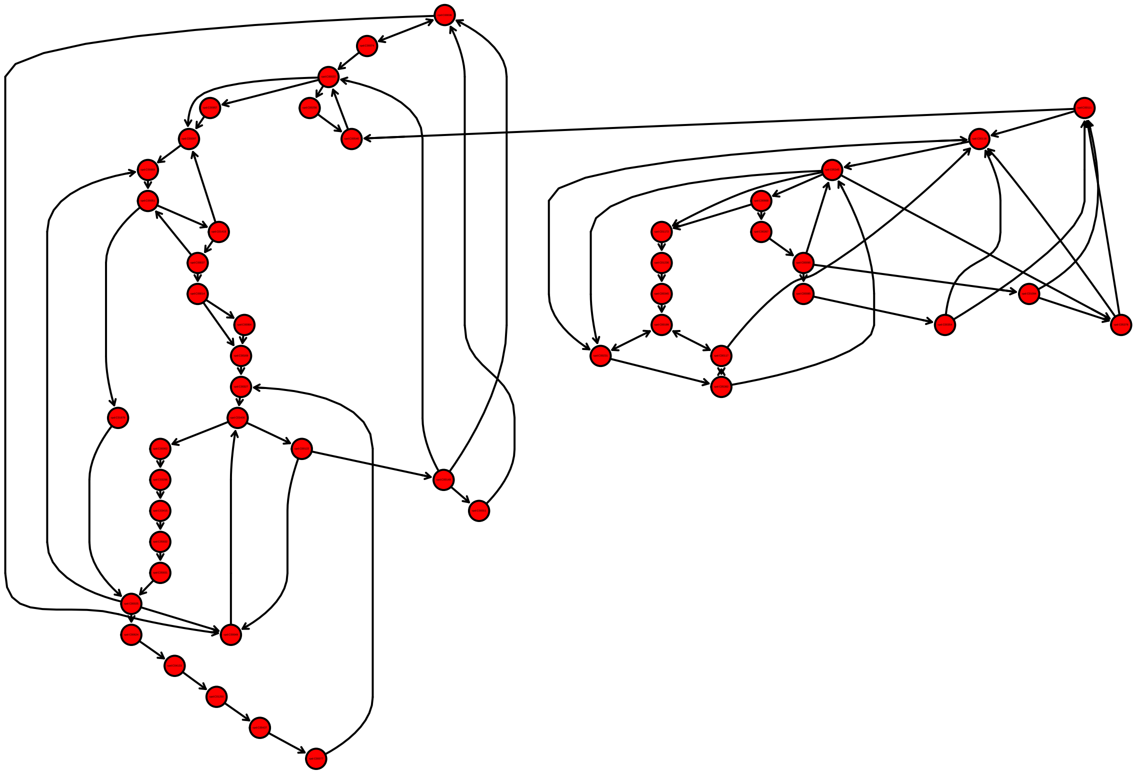
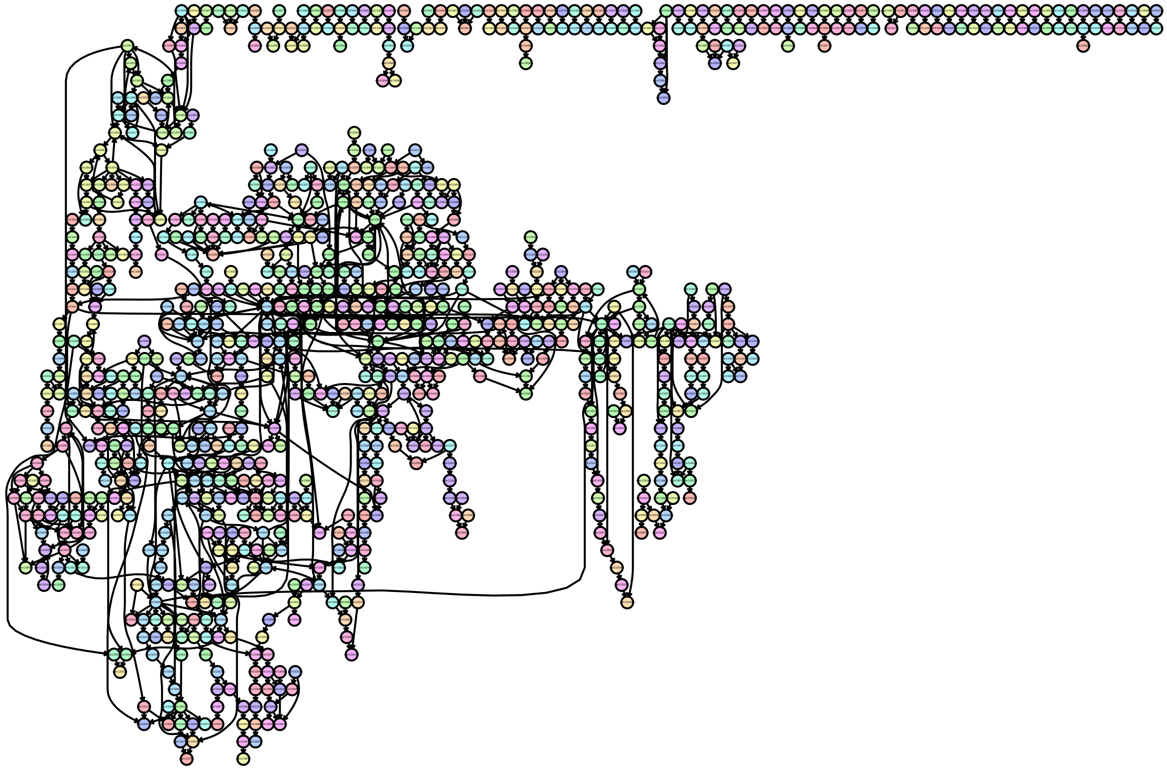


Fig. 3 The first two largest strongly connected components of the reaction network. Each node is one metabolite. Each edge is a reaction. The graph is directed because reaction has substrates and products.

So, we can replace the strongly connected components with one node in the simplified network.

Step 2: Annotate the strongly connected components in graph

Below showed that the nodes are annotated with different colors for different strongly connected components they belong to.



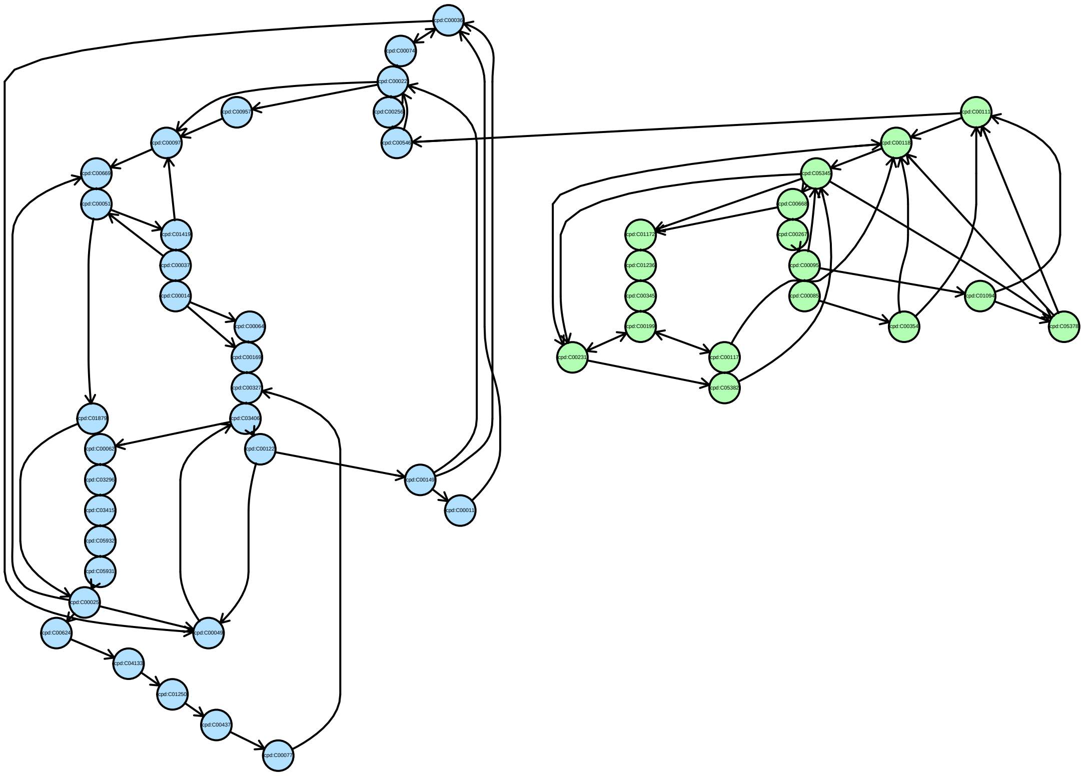


Fig. 4 Nodes of different strongly connected components are attributed with different colors.in this network, all nodes of a strongly connected components are plot with the same color.

Step 2: Contract several vertices of the same components into a single one

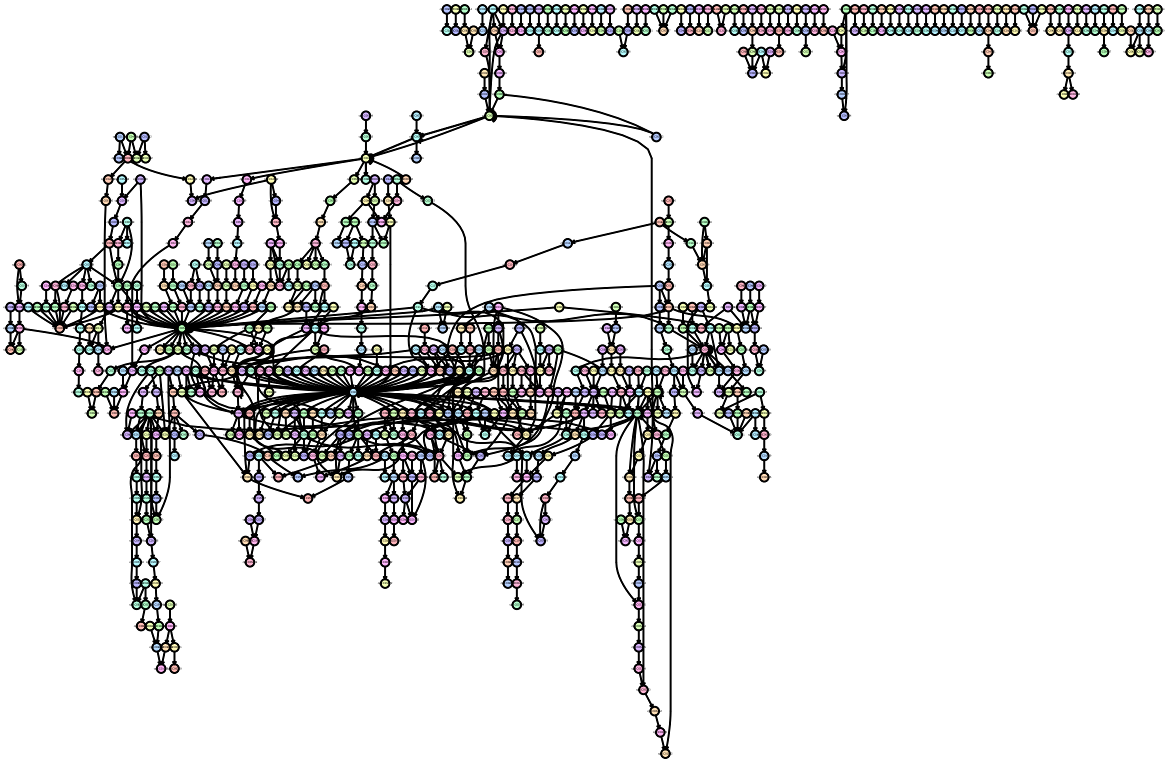


Fig. 5 The contracted reaction network.

We can see the number of nodes is reduced from 924 to 830 in the simplified network of E.Coli__. In this simplied figure, every node should have different colors.

Step 3: determine the Seed set and product set of a metabolic network

Next, I want to determine what nutrients should be obtained from the environment to produce other compounds. In the network, they are the nodes which don't have indegrees, which means these compounds can only be taken up from environment.These nodes are defined as seed set here. While the remaining nodes with indegree are the compounds which can be produced by the seed set.These nodes are defined as product sets.

seed set of metabolic network of E.Coli

```
##   [1] "cpd:C02147" "cpd:C00685" "cpd:C06714" "cpd:C02247" "cpd:C03319" "cpd:C05847" "cpd:C21994" "cpd
##  [10] "cpd:C00423" "cpd:C00127" "cpd:C01594" "cpd:C00561" "cpd:C03742" "cpd:C20396" "cpd:C00986" "cpd
##  [19] "cpd:C11457" "cpd:C05629" "cpd:C12621" "cpd:C05607" "cpd:C02265" "cpd:C00582" "cpd:C05332" "cpd
##  [28] "cpd:C05998" "cpd:C00793" "cpd:C00491" "cpd:C01888" "cpd:C00740" "cpd:C00940" "cpd:C02362" "cpd
##  [37] "cpd:C02355" "cpd:C02067" "cpd:C20254" "cpd:C00881" "cpd:C02353" "cpd:C06194" "cpd:C01260" "cpd
##  [46] "cpd:C00301" "cpd:C00580" "cpd:C11142" "cpd:C00288" "cpd:C01417" "cpd:C11537" "cpd:C00565" "cpd
##  [55] "cpd:C00121" "cpd:C00620" "cpd:C00490" "cpd:C00497" "cpd:C01412" "cpd:C00583" "cpd:C04593" "cpd
##  [64] "cpd:C00898" "cpd:C01380" "cpd:C03451" "cpd:C00937" "cpd:C01177" "cpd:C01204" "cpd:C00270" "cpd
##  [73] "cpd:C01132" "cpd:C02262" "cpd:C05402" "cpd:C05404" "cpd:C00492" "cpd:C00618" "cpd:C11516" "cpd
##  [82] "cpd:C00392" "cpd:C01019" "cpd:C00507" "cpd:C01934" "cpd:C00502" "cpd:C00312" "cpd:C04053" "cpd
##  [91] "cpd:C15930" "gl:G10610"  "cpd:C02970" "cpd:C00272" "cpd:C01007" "cpd:C05791" "cpd:C00853" "cpd
## [100] "cpd:C01935" "cpd:C01898" "cpd:C16241" "cpd:C05980" "cpd:C02356" "cpd:C00461" "cpd:C15532" "cpd
## [109] "cpd:C03150" "cpd:C15811" "cpd:C15810" "cpd:C20247" "cpd:C04294" "cpd:C01279" "cpd:C00378" "cpd
## [118] "cpd:C00989" "cpd:C16675" "cpd:C20386" "cpd:C00072" "cpd:C00114" "cpd:C00880" "cpd:C01697" "cpd
## [127] "cpd:C01847" "cpd:C00053" "cpd:C00818" "cpd:C00243" "gl:G13040"  "cpd:C03460" "cpd:C06001" "cpd
## [136] "cpd:C00198" "cpd:C06473" "cpd:C00469" "cpd:C02282" "cpd:C00798" "cpd:C03089" "cpd:C03546" "cpd
## [145] "cpd:C11638" "cpd:C07335" "cpd:C02723" "cpd:C02325" "cpd:C00590" "cpd:C02646" "cpd:C02730" "cpd
## [154] "cpd:C06508" "cpd:C06505" "cpd:C00430" "cpd:C00473" "cpd:C03479" "cpd:C00568" "cpd:C01063" "cpd
## [163] "cpd:C16476" "cpd:C02501" "cpd:C04706" "cpd:C16348" "cpd:C06613" "cpd:C12835" "cpd:C07478" "gl
## [172] "cpd:C05892" "cpd:C01212" "cpd:C06397" "cpd:C06251" "cpd:C04121" "cpd:C04652" "cpd:C00448" "cpd
## [181] "cpd:C16331" "cpd:C06427" "cpd:C01595" "cpd:C00219" "cpd:C04635" "cpd:C04317" "cpd:C00641" "cpd
## [190] "cpd:C00245" "cpd:C05688" "cpd:C00295" "cpd:C02350" "cpd:C11821" "cpd:C00002" "cpd:C03090" "cpd
## [199] "cpd:C00249" "cpd:C03939" "cpd:C04618" "cpd:C04620" "cpd:C04619" "cpd:C01209" "cpd:C04633" "cpd
## [208] "cpd:C00233" "cpd:C04411" "cpd:C02504" "cpd:C04272" "cpd:C06010" "cpd:C06007" "cpd:C01165" "cpd
## [217] "cpd:C00082" "cpd:C01267" "cpd:C01157" "cpd:C05946" "cpd:C00322" "cpd:C04462" "cpd:C00047" "cpd
## [226] "cpd:C00021" "cpd:C01077" "cpd:C05519" "cpd:C01242" "cpd:C01005" "cpd:C00152" "cpd:C00246" "cpd
## [235] "cpd:C05668" "cpd:C00988" "cpd:C00168" "cpd:C04006" "cpd:C00096" "cpd:C00159" "cpd:C04631" "cpd
## [244] "cpd:C00369" "cpd:C00714" "cpd:C00333" "cpd:C03033" "cpd:C00259" "cpd:C01101"
```

There are 249 strongly connected components as the seed set.The ID are the KEGG accession numbers. The compounds in the strongly connected components are named by one of the nodes.

Product set

Then the remained (841-249)=591 components belong to product set.

Plot the seed set and product set in graph

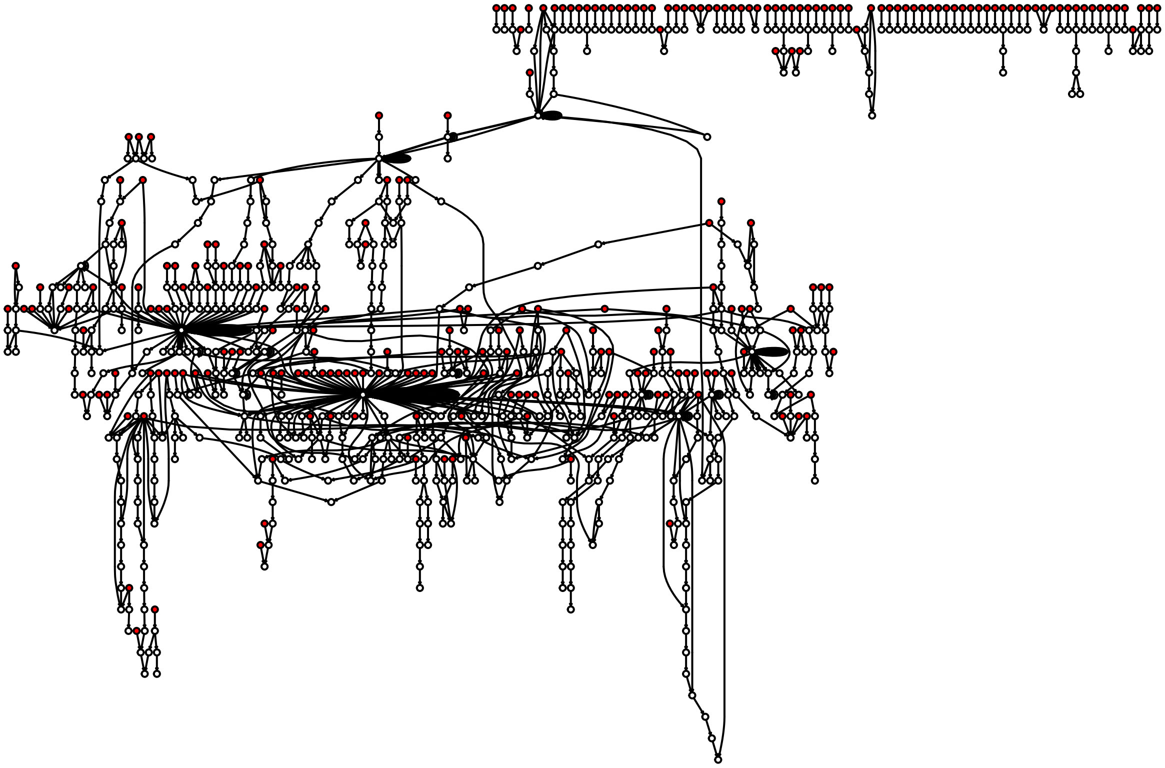


Fig. 6 The network colored by seed set (red) and product set(white).

Step 4: Metabolic competition and complementarity

Using the seed set and product set, we can define metabolic competition and complementarity indexes. The metabolic competition index represents the similarity in two species' nutritional profiles. It is calculated as the fraction of compounds of query species X's seed set that are also present in the seed set of a target Y. The metabolic complementarity index is calculated as the fraction of seed compounds of a query species X that are producible by the metabolic network of a target Y but are not a part of Y's seed set.

- It should be noted that this interaction is not symmetric. For example, A may compete with B because 90% of seed set of A is shared by B. However, B may only share 10% of seed set with A, so B would not competed with A.

Here, I will use two microbial species: E.Coli and Streptomyces coelicolor as a example and calculate the two indexes for their metabolic network.

I have shown the network for E.Coli. So I will show the simplified metabolic network of Streptomyces coelicolor colored by seed set and product set.

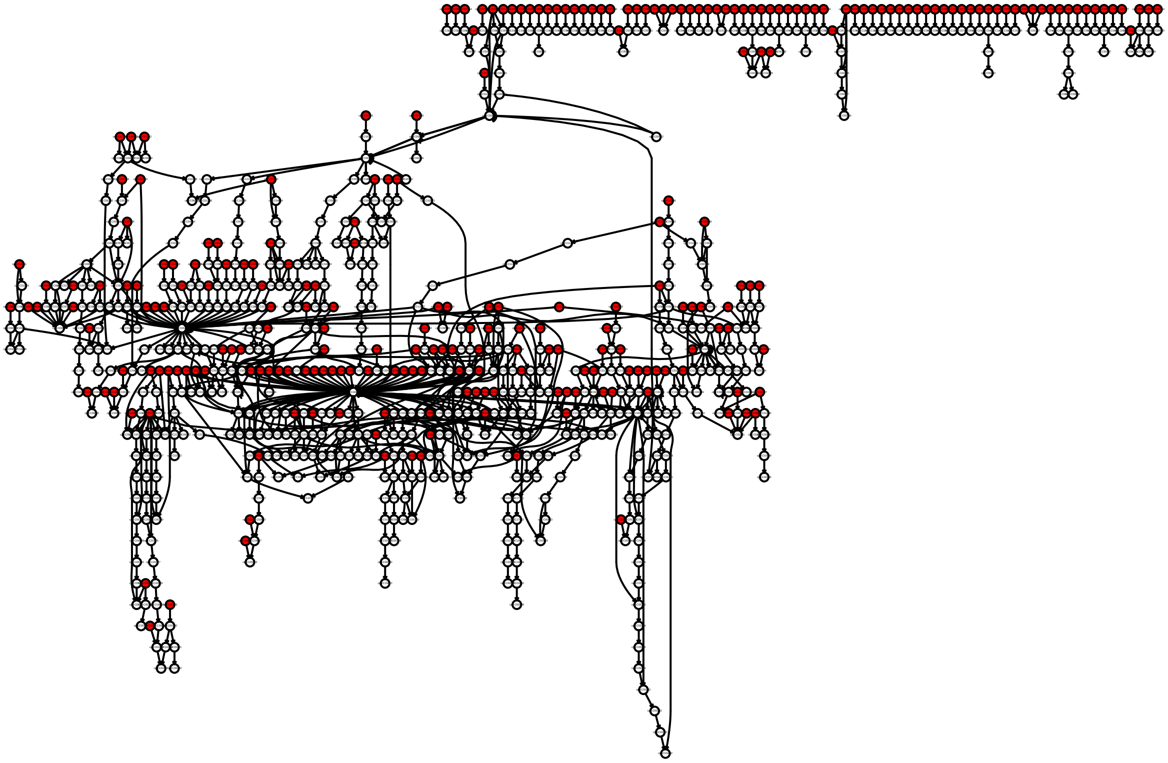
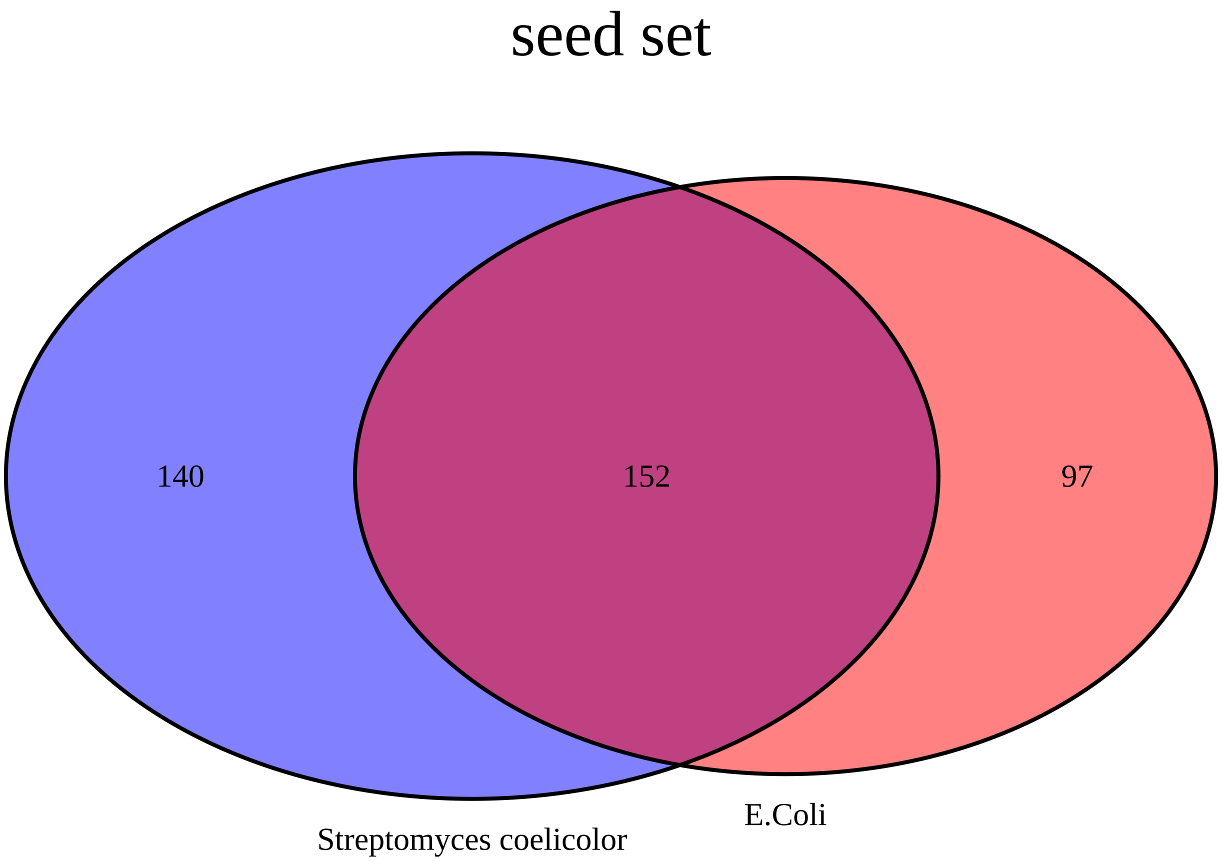


Fig. 7 The network colored by seed set (red) and product set(white) of Streptomyces coelicolor.

Step 4: The competition index

So we need to see the shared red nodes (i.e. seed set) of two species.

Step 4: The shared seed set

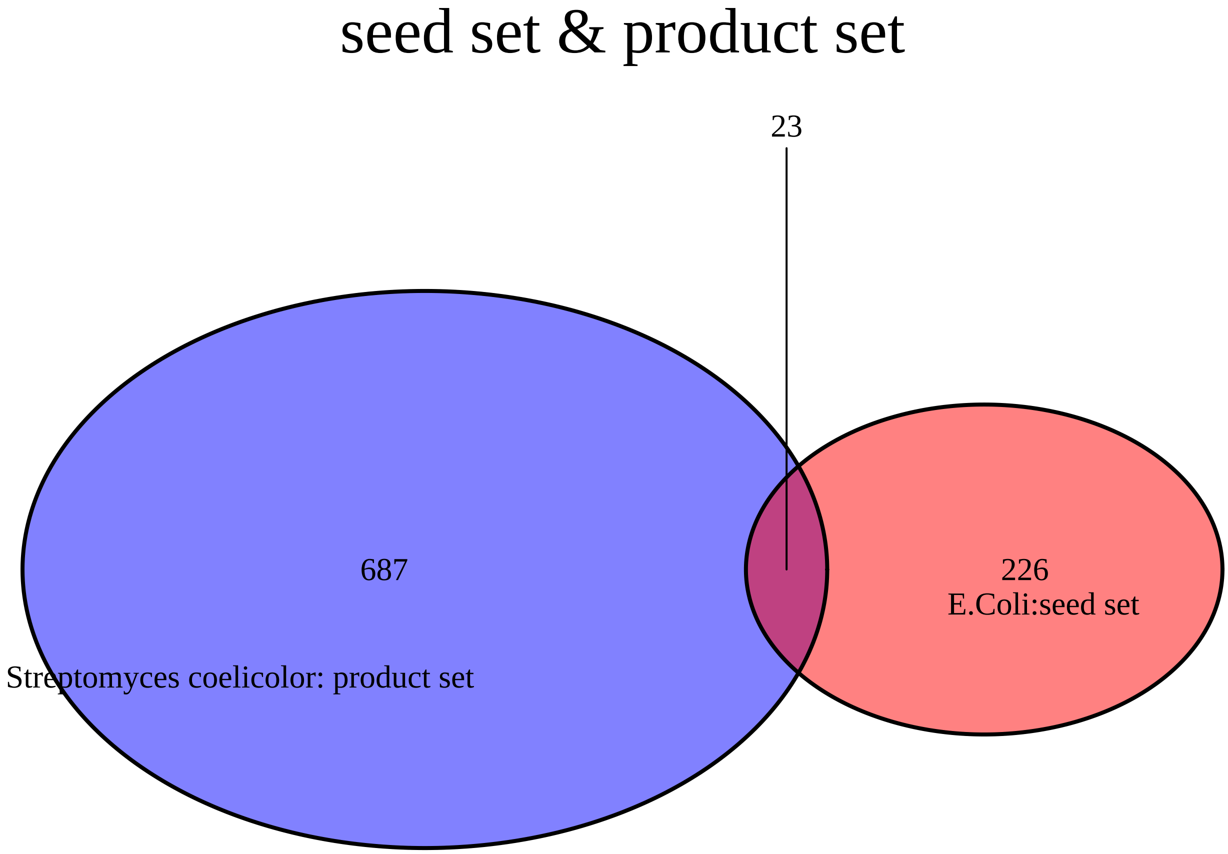


So the competition index for Streptomyces -> E.Coli is $153/(96 + 153) = 0.6144578$ while the index for E. coli -> Streptomycesis $140/(140 + 153) = 0.4778157$. There values represent how many types of nutrition they may need in common.

Step 4: the complementarity index

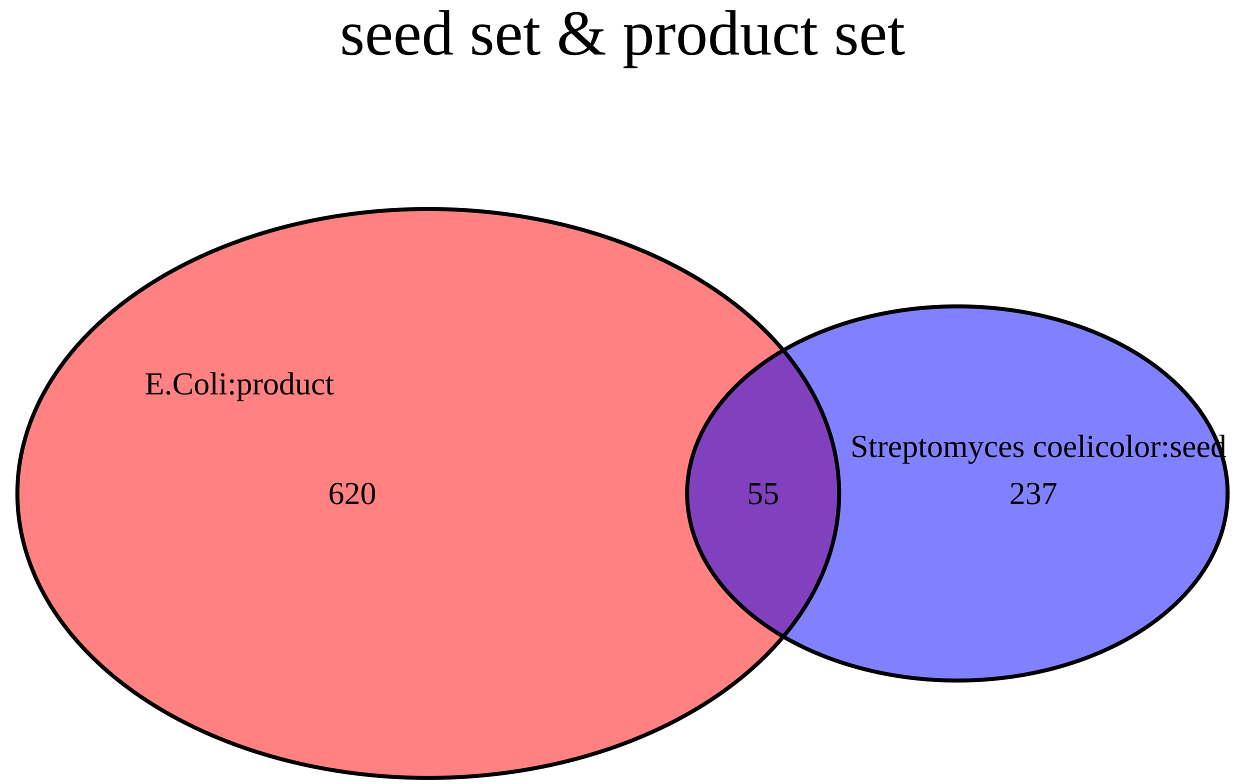
There index is directed, too. 1. The product of streptomyces could be used as seed set of E.Coli; 2.The product of E.Coli could be used as seed set of Streptomyces.

Step 4: shared seed set of E.Coli and product set of Streptomyces



So the complementarity index for Streptomyces -> E.Coli is $22/(22 + 227) = 0.0883534$,, which means 0.0883534 of nutrition requirement of E.Coli could be provided by products of Streptomyces.

Step 4: shared seed set of Streptomyces and product set of E.Coli



So the complementarity index for E.Coli -> Streptomyces is $55/(55 + 238) = 0.1877133$, which means 0.1877133 of nutrition requirement of Streptomyces could be provided by products of E.Coli.

Step 5: Calculate the two indexes for all microbes

In a real community in the nature i.e. the soil and intestine, there are over thousands species coexisting together. Their metabolic interaction is complex and remained unclear. As KEGG database is the most comprehensive database involving various species. If we can calculate the indexes for species of the database. We could further apply the calculated indexes for a real community, in which most of species has been included in KEGG database.

```
## [1] 6325
```

There are a total 1429 microbial species in KEGG database. There are several strains with slightly different network corresponding to one species, I will use the first one as a representative of the species.
To calculate the indexes for paired species, I repeated the steps for single species to calculate the seed set and product set for all species. Next, I calculated the indexes for paired species. Generally, it is a time-consuming work.

Pairwise index

In total, I downloaded 1429 metabolic network, and then determined their seed set and product set.
Thus, I could build a matrix with 1429 columns and 1429 rows that represent the directed index for metabolic competition and also one for metabolic complementarity.

The form of the matrix

Below shows the example of the competition matrix.

```
##           eco       sty       sfl       enc       esa       kpn       cro       gqu       bfl
## eco 1.0000000 0.7923077 0.7843866 0.8286853 0.8513514 0.7875458 0.8888889 0.7330827 0.3823529
## sty 0.8273092 1.0000000 0.8289963 0.7888446 0.8423423 0.7435897 0.8846154 0.6917293 0.5000000
## sfl 0.8473896 0.8576923 1.0000000 0.7689243 0.8243243 0.7289377 0.8205128 0.6804511 0.5411765
## enc 0.8353414 0.7615385 0.7174721 1.0000000 0.9324324 0.7985348 0.8418803 0.7631579 0.4000000
## esa 0.7590361 0.7192308 0.6802974 0.8247012 1.0000000 0.7142857 0.8119658 0.7030075 0.3941176
## kpn 0.8634538 0.7807692 0.7397770 0.8685259 0.8783784 1.0000000 0.8846154 0.7857143 0.3882353
## cro 0.8353414 0.7961538 0.7137546 0.7848606 0.8558559 0.7582418 1.0000000 0.7067669 0.4000000
## gqu 0.7831325 0.7076923 0.6728625 0.8087649 0.8423423 0.7655678 0.8034188 1.0000000 0.3882353
## bfl 0.2610442 0.3269231 0.3420074 0.2709163 0.3018018 0.2417582 0.2905983 0.2481203 1.0000000
```

The statistics of the two indexes

As I have gotten two matrix for indexes, some basic statistics could help us understand the possible interaction for all the isolated species. For example, distribution of the indexes could show different level of interaction.

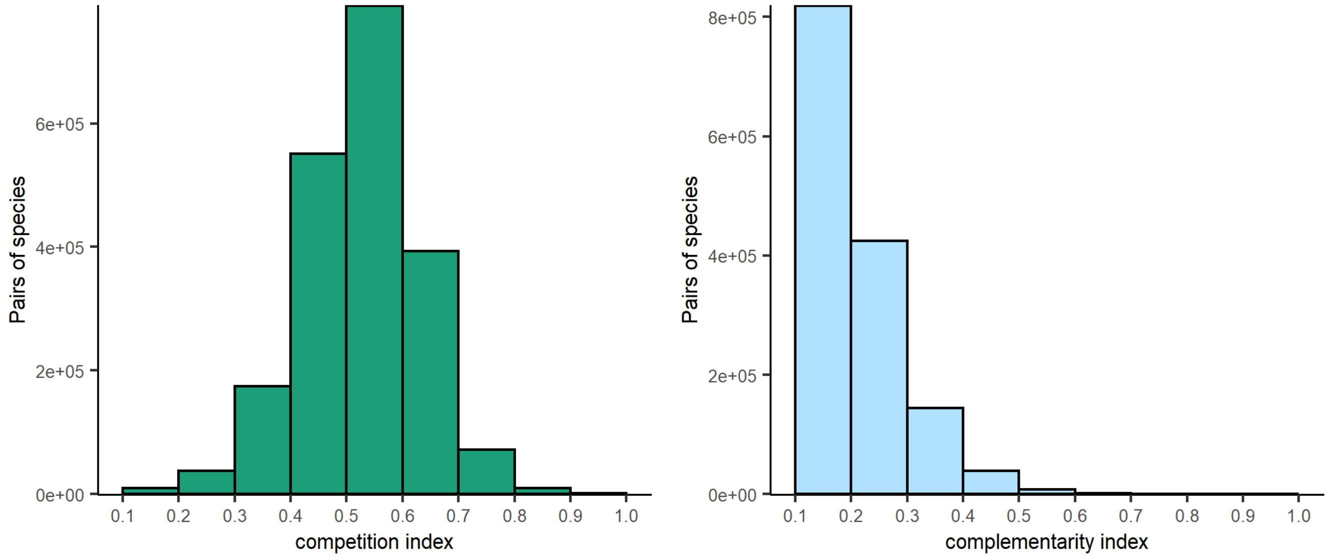


Fig. 8 The distribution of competition indexes(left) and complementariy index (right).

we can see the distribution is a normal distribution for competition indexes and a Poisson distribution for complementarity indexes. and the value is mainly between 0.3-0.8 for competition indexes and 0.1-0.5 for complementarity indexes

we can also show some clusters of species has similar interaction with other species.

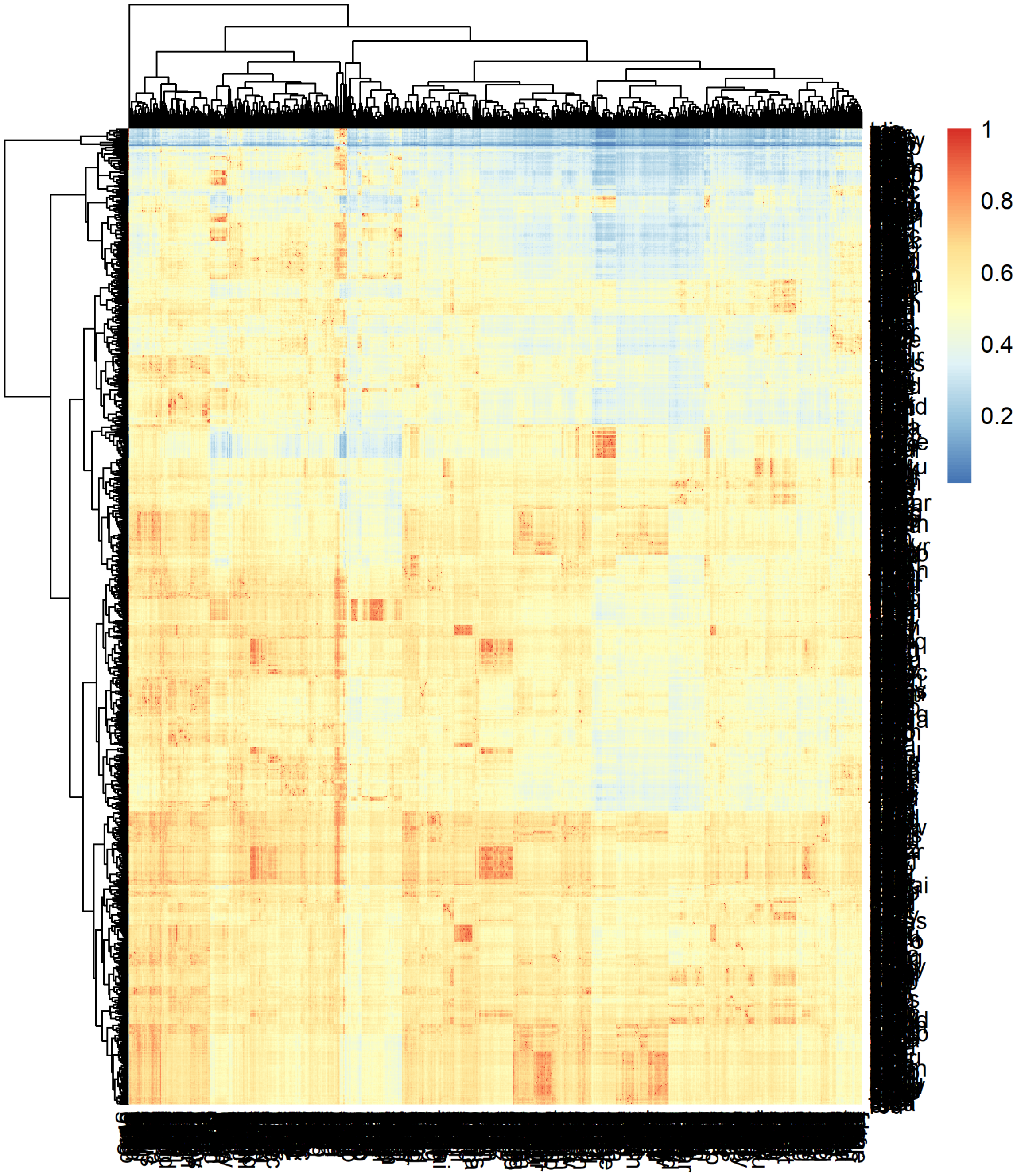


Fig. 9 The heatmap of the matrix for competition indexes.

The clustered species shows similar competition indexes with other groups. For example, the blue part is the species with low competition to some species.

Programming approach

The programming is based on R language. I mainly used the R package KEGGgraph to import the KEGG network and extract only the reaction network for later process. The plot of the metabolic network is based on the package Rgraphviz, which has good layout for the metabolic network. The process of network is done using the package igraph, which is the most powerful network processing package in R and convenient for usage. To obtain the species information from KEGG in batches, I also used the package rvest to crawl data from KEGG. The packages ggplot2 and pheatmap are also utilized for plotting figures.

What is next

I will download several microbial interaction network from the public database. Then, I will test if the metabolic interaction is associated with their interaction predicted by abundance in communities.