PhDSeeker: Pheromone-Directed Seeker for Metabolic Pathways

Matias F. Gerard[a,*], Raúl N. Comelli[b]

[a] *Research Institute for Signals, Systems and Computational Intelligence (CONICET-UNL), Ciudad Universitaria, Santa Fe, Argentina.*
[b] *Departamento de Medio Ambiente, Fac. de Ingeniería y Ciencias Hídricas (FICH), Univ. Nacional del Litoral (UNL), Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET). Ciudad Universitaria CC 242 Paraje El Pozo, 3000, Santa Fe, Argentina.*

Abstract

Manually finding relationship networks among compounds can be a hard and time-consuming task. However, this process is fundamental when looking for a metabolic pathway that explains how multiple compounds are related, to identify relevant pathways in organisms, filling gaps on metabolic networks, or when new mechanisms for the synthesis of important compounds are sought.

Here, we present PhDSeeker, a new tool for the automatic search of metabolic pathways. This tool is able to relate simultaneously several compounds. Furthermore, its flexibility allows it to be easily configured for addressing a wide range of situations. Solutions found are provided not only in plain text but also as interactive representations that can be analyzed in a web browser.

Source code is available at https://github.com/sinc-lab/phdseeker. A web service is also available at https://sinc.unl.edu.ar/web-demo/phds/.

[*] Corresponding author. Tel: +54 (0342) 457 5234 int 118.
Email addresses: `mgerard@sinc.unl.edu.ar` (Matias F. Gerard), `rcomelli@fich.unl.edu.ar` (Raúl N. Comelli)

Several fully documented study cases, including their settings and solutions files, are also provided as Supplementary Material.

Keywords: ant colony optimization, metabolic network representation, Metabolic pathway searching, Pathways synthesis

1. Introduction

Nowadays, information of metabolic pathways for a large number of living beings is available in databases such as KEGG [1, 2], BioCyc [3] and BRENDA [4]. This allows the online exploration of enzymes, their catalyzed biochemical reactions, and the substrates and products involved. Despite individual rules for producing compounds are well-known, it is still a challenge to identify the adequate sequence of reactions required for the synthesis of several compounds as part of a (novel) complex metabolic network [5]. Furthermore, typical tasks such as finding relationships among metabolites, filling gaps or proposing new ways to synthesize particular compounds can be time-consuming [6, 7].

Currently, there is a wide range of helpful tools available for designing metabolic pathways [8, 9], that can be grouped in two main categories: web-based and downloadable. Most of the web-based tools are associated to well-known databases such as KEGG, BioCyc and BRENDA, and give the possibility of searching metabolic pathways on their own catalogues using web access. They also include APIs and REST services, but only for accessing to stored data. Only BioCyc provides a downloadable full suite to perform the search locally. Within this category, there are also tools such as XTMS [10], BioSynther [11], MRE [12] or ATLAS [13] which combine information of those databases, mainly from KEGG, together with information of thermodynamics, stoichiometry or chemical similarity, to improve results of the design. The remaining category comprises tools that can be downloaded for designing metabolic pathways in a local machine. These are based in variations of classical search methods, allowing a greater control of the algorithm and the data used. Some examples include ReTrace [14], LPAT [15], Metabolic Tinker [16] and AGPathFinder [17]. Despite their features, both categories have some shortcomings. While simplicity of web-based tools could make the user job easier, this mechanism of access limits the possibility to modify the search space, by adding or removing specific reactions or compounds. Furthermore, a registration step is often previously required for service use, and the source code is not available for modifications. On the other hand, despite flexibility of downloadable tools, these often lack of a graphical interface and require the installation of a database or, in many cases, of additional software. All this makes these tools difficult for non-experts to use. In addition, although some include a web server to perform the search, it is often out of service.

In this work we propose a novel tool, named PhDSeeker, for the automatic design of metabolic pathways, together with a set of auxiliary tools to facilitate the construction of the datasets to carry out the searches. This Python-based tool can be used through a web service or downloaded to run in a local machine by using command-line or a graphical interface. This not only makes it more user-friendly for non-expert users, but also gives them greater control over every aspect of the design process. The core of PhDSeeker is a bioinspired metaheuristic algorithm that efficiently explores multiple pathways to relate simultaneously two or more specified compounds. Comparisons with classical searching algorithms and other recent metaheuristic approaches shown clear advantages of our bioinspired algorithm, which was able to recover complete well-known pathways [18].

The way pathways are modeled enables this tool to synthesize branched solutions. Furthermore, pathways are found taking into account the initial conditions specified. Synthesized pathways are displayed through an interactive representation that can be visualized in a web browser. Those features make PhDSeeker a valuable tool, suitable for a wide range of applications, such as synthetic biology, interpretation of metabolomics experiments and gap-filling in metabolic reconstructions.

2. Software description

PhDSeeker is an automatic tool for synthesis and design of metabolic pathways. It is implemented in Python language using standard libraries, thus allowing it to be multiplatform. This tool can be used either through a graphical user interface (GUI) or by command line. In both cases, users can specify the parameters and the initial conditions for the search, as well as the reactions dataset in a simple way. The typical PhDSeeker workflow is shown in Figure 1.

2.1. Datasets and settings

Reactions datasets consist of plain text files with one biochemical equation per line. Each one is composed by an identifier, its substrates and products, and the stoichiometric coefficients (for example, R00658: 1 C00631 → 1 C00074 + 1 C00001). Substrates and products can be indicated using their names or database identifiers, but they should not include white spaces. Datasets can be built manually, by integrating information of several databases or also including user-defined reactions. In order to help with this task, three auxiliary tools to automatize this process are provided with PhDSeeker. They can be used through a graphical interface or by command-line, and they allow to build datasets from KEGG[1], BioCyc[3] or BiGG[19] databases. These tools also automatically build two files, one with the information to translate database identifiers to compound names (for example, C00001 corresponds to water in KEGG database), and the other one with information regarding the enzymes catalyzing every reaction. This information is then used in the construction of the solutions' graphical representation. An example of those files is shown in Figure1.

PhDSeeker can be configured using its GUI, or by mean of a settings file if the tool is run in a command line. Required information and available options are the same for both cases. Building the settings file is straightforward and only requires specification of the reactions dataset, the compounds to relate, and a list of compounds to be considered as freely available (such as water, ATP, etc). This file also allows to indicate which compounds to relate should be taken into account as a source compound for pathway design. Additionally, the paths for the file with information to translate database specific identifiers to compound names, and for the file with the enzymes using each reaction in the dataset can be included. Although it is also possible to define some algorithm parameters in this file, the default values provided in the examples are adequate for most cases. When GUI-based interface is run, a set of pop-up windows allows to select in an intuitive way all files and required compounds.

2.2. Algorithm

The core of this tool is an ant colony optimization-based algorithm designed for metabolic pathways synthesis [18]. This emulates the behavior of real ants while seeking a path between their colony and a source of food. On each iteration, ants explore the set of reactions searching for possible fully connected pathways to link the compounds. After that, they share information about solutions found by each one before performing a new search. This process is guided by a cost function that evaluates the availability of substrates for each reaction (feasibility), the existence of a connection between source and target, and the number of steps involved in the pathway. As a result, solutions always relate the specified compounds through sequences of feasible reactions with as few reactions as it is possible. Furthermore, better solutions are proposed along the search process as a consequence of the knowledge shared by ants. A graphical summary of the algorithm is shown at the middle of Figure 1.

2.3. Reports and visualization

The result of the design is reported by PhDSeeker by means of three files. One corresponds to an HTML file where all the solutions found during the design can be visualized through an interactive representation. This allows to identify, at a glance, the general structure of the pathway and how reactions are interconnected. This visualization also includes hyperlinks to PubChem [20] and BioCyc for every compound in the pathway. Furthermore, hyperlinks to BRENDA database for each enzyme involved in the pathways are available, if the information of enzymes catalyzing every reaction is provided. The other two files store, in plain text, the full information of the design process. One comprises the history of the pathways found during the search of a solution, and the other holds a detailed summary of the sequence of reactions used to build the pathway together with the experimental setup.

3. Applications

PhDSeeker allows to address a wide variety of problems, ranging from simply finding relationships among compounds, to designing novel pathways with applications in metabolic engineering. To illustrate the utility and features of this tool, we applied it for the design of a pathway to produce isobutanol from pyruvate. This alcohol has important applications in the biofuel field. In consequence, its synthesis by microorganisms such as *Escherichia coli* (*E. coli*) could provide an economical way to obtain it. Reactions from MetaCyc v21.5 database [21] where used to design a pathway that synthesizes isobutanol. This allowed PhDSeeker to explore around 16000 reactions from approximately 3000 different organisms to find a solution.

The pathway proposed by PhDSeeker for this task is shown at the bottom of Figure 2a. It is comprised of five reactions, from which only the two furthest to the right in the figure are not in *E. coli*. Interestingly, this solution is identical to the one manually designed by *Atsumi* and coworkers for the same purpose [22], that is shown in Figure 2b. This example shows the usefullness of the tool to assist in the process of finding relationships and designing pathways for compound production. Experimental details about the search configuration and the analysis for this example and for other more complex study cases are available in Supplementary Material.

4. Conclusion

PhDSeeker provides a simple way to automatically search and design metabolic pathways in a desktop PC or using a web service. Configuration results very intuitive, either using its graphical interface or a settings file. Its flexibility allows for it to cover a wide range of applications, from finding relationships among specific compounds to the design of novel pathways for metabolic engineering. The proposed pathways can be analyzed easily through an interactive representation that can be displayed in any web browser. Those features make PhDSeeker a useful bioinformatic tool for systems biology and metabolic engineering, capable of simplifying the analysis and design of new proposals.

Acknowledgements

This work was supported by UNL (CAI+D 2016 #42).

[1] M. Kanehisa, S. Goto, KEGG: Kyoto Encyclopedia of Genes and Genomes, Nucleic Acids Res. 28 (1) (2000) 27–30. doi:10.1093/nar/28.1.27.

[2] M. Kanehisa, Y. Sato, M. Furumichi, K. Morishima, M. Tanabe, New approach for understanding genome variations in KEGG, Nucleic Acids Res. 47 (D1) (2019) D590–D595. doi:10.1093/nar/gky962.

[3] P. D. Karp, R. Billington, R. Caspi, C. A. Fulcher, M. Latendresse, A. Kothari, I. M. Keseler, M. Krummenacker, P. E. Midford, Q. Ong, W. K. Ong, S. M. Paley, P. Subhraveti, The BioCyc collection of microbial genomes and metabolic pathways, Brief. Bioinform. (2017) bbx085doi:10.1093/bib/bbx085.

[4] L. Jeske, S. Placzek, I. Schomburg, A. Chang, D. Schomburg, BRENDA in 2019: a European ELIXIR core data resource, Nucleic Acids Res. 47 (D1) (2019) D542–D549. doi:10.1093/nar/gky1048.

[5] Z. Algfoor, M. Sunar, A. Abdullah, H. Kolivand, Identification of metabolic pathways using pathfinding approaches: a systematic review, Briefings in Functional Genomics 16 (2) (2017) 87–98. doi:10.1093/bfgp/elw002.

[6] A. S. Karim, M. C. Jewett, Chapter two - cell-free synthetic biology for pathway prototyping, in: N. Scrutton (Ed.), Enzymes in Synthetic Biology, Vol. 608 of Methods in Enzymology, Academic Press, 2018, pp. 31 – 57. doi:10.1016/bs.mie.2018.04.029.

[7] Z. Tan, X. Zheng, Y. Wu, X. Jian, X. Xing, C. Zhang, In vivo continuous evolution of metabolic pathways for chemical production, Microbial Cell Factories 18 (2019) 82. doi:10.1186/s12934-019-1132-y.

[8] L. Wang, S. Dash, C. Ng, C. Maranas, A review of computational tools for design and reconstruction of metabolic pathways, Synthetic and Systems Biotechnology 2 (4) (2017) 243–252. doi:10.1016/j.synbio.2017.11.002.

[9] S. Kim, M. Peña, M. Moll, G. Bennett, L. Kavraki, A review of parameters and heuristics for guiding metabolic pathfinding, Journal of Cheminformatics 9 (51). doi:10.1186/s13321-017-0239-6.

[10] P. Carbonell, P. Parutto, J. Herisson, S. Pandit, J. Faulon, XTMS: pathway design in an eXTended metabolic space, Nucleic Acids Research 42 (2014) W389–W394. doi:10.1093/nar/gku362.

[11] W. Tu, H. Zhang, J. Liu, Q. Hu, BioSynther: a customized biosynthetic potential explorer, Bioinformatics 32 (3) (2016) 472–473. doi:10.1093/bioinformatics/btv599.

[12] H. Kuwahara, M. Alazmi, X. Cui, X. Gao, MRE: a web tool to suggest foreign enzymes for the biosynthesis pathway design with competing endogenous reactions in mind, Nucleic Acids Research 44 (2016) W217–W225. doi:10.1093/nar/gkw342.

[13] N. Hadadi, J. Hafner, A. Shajkofci, A. Zisaki, V. Hatzimanikatis, ATLAS of Biochemistry: A Repository of All Possible Biochemical Reactions for Synthetic Biology and Metabolic Engineering Studies, ACS Synth. Biol. 5 (2016) 1155–1166. doi:10.1021/acssynbio.6b00054.

[14] E. Pitkänen, P. Jouhten, J. Rousu, Inferring branching pathways in genome-scale metabolic networks, BMC Systems Biology 3 (2009) 103. doi:10.1186/1752-0509-3-103.

[15] A. Heath, G. Bennett, L. Kavraki, Finding metabolic pathways using atom tracking, Bioinformatics 26 (2010) 1548–1555.

[16] K. McClymont, O. Soyer, Metabolic tinker: an online tool for guiding the design of synthetic metabolic pathways, Nucleic Acids Research 41 (11) (2013) e113. doi:10.1093/nar/gkt234.

[17] Y. Huang, C. Zhong, H. Lin, J. Wang, A Method for Finding Metabolic Pathways Using Atomic Group Tracking, PLOS ONE 12 (1) (2017) e0168725. doi:10.1371/journal.pone.0168725.

[18] M. Gerard, G. Stegmayer, D. Milone, Metabolic pathways synthesis based on ant colony optimization, Sci. Rep. 8 (2018) 16398. doi:10.1038/s41598-018-34454-z.

[19] Z. A. King, J. Lu, A. Dräger, P. Miller, S. Federowicz, J. A. Lerman, A. Ebrahim, B. O. Palsson, N. E. Lewis, BiGG Models: A platform for integrating, standardizing and sharing genome-scale models, Nucleic Acids Res. 44 (D1) (2016) D515–D522. doi:10.1093/nar/gkv1049.

[20] S. Kim, J. Chen, T. Cheng, A. Gindulyte, J. He, S. He, Q. Li, B. Shoemaker, P. Thiessen, B. Yu, L. Zaslavsky, J. Zhang, E. Bolton, PubChem 2019 update: improved access to chemical data, Nucleic Acids Res. 47 (D1) (2019) D1102–D1109. doi:10.1093/nar/gky1033.

[21] R. Caspi, T. Altman, K. Dreher, C. Fulcher, P. Subhraveti, I. Keseler, A. Kothari, M. Krummenacker, M. Latendresse, L. Mueller, Q. Ong, S. Paley, A. Pujar, A. Shearer, M. Travers, D. Weerasinghe, P. Zhang, P. Karp, The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases, Nucleic Acids Res. 40 (2011) D742–53.

[22] S. Atsumi, T.-Y. Wu, E.-M. Eckl, S. D. Hawkins, T. Buelter, J. C. Liao, Engineering the isobutanol biosynthetic pathway in *Escherichia coli* by comparison of three aldehyde reductase/alcohol dehydrogenase genes, Appl. Microbiol. Biotechnol. 85 (2010) 651–657. doi:10.1007/s00253-009-2085-6.

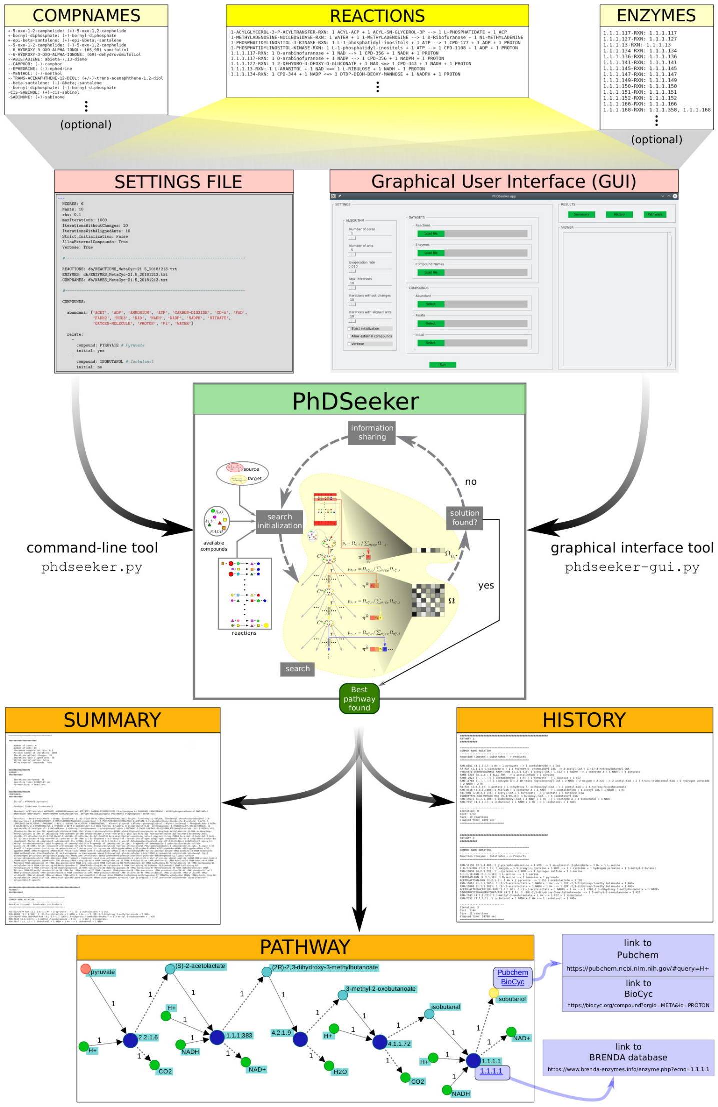


Figure 1: PhDSeeker workflow. Data files, sets of compounds and algorithm parameters' can be specified using the GUI or through a *settings file* if command-line interface is used. This information is used by an ant-colony based algorithm to iteratively build better metabolic pathways to relate compounds according to initial conditions specified. Results are returned as two plain text files (summary and history) and also as an interactive visualization (pathway) with the best pathways found along the search. Compounds and reactions include links to some important databases. Blue squares show and example of links for enzymes (EC: 1.1.1.1) and compounds (isobutanol).

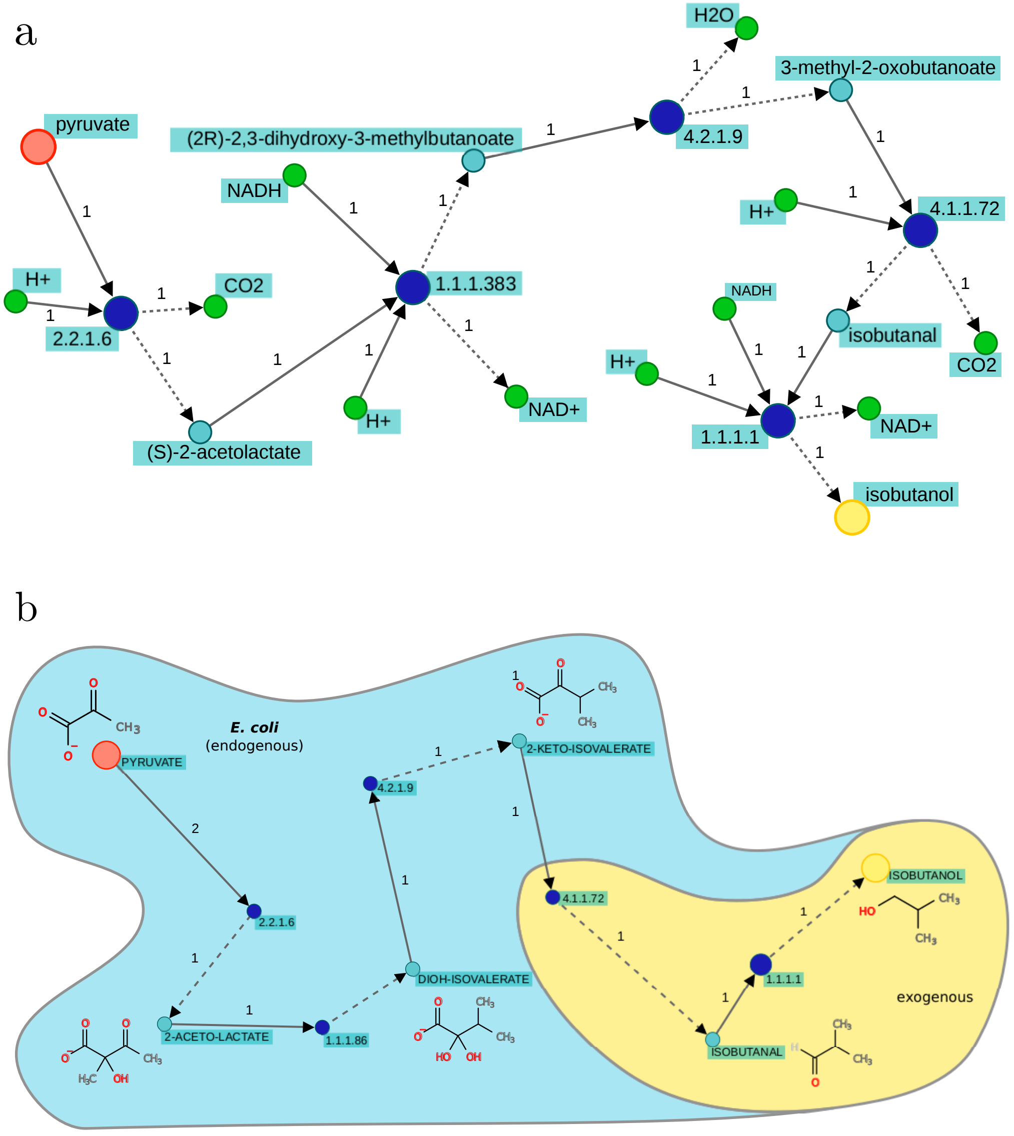


Figure 2: Synthesis of isobutanol from pyruvate in *E. coli*. a) Pathway proposed by PhDSeeker. b) Pathway designed and experimentally validated by Atsumi *et al.*

Supplementary Material

PhDSeeker: Pheromone-Directed Seeker for Metabolic Pathways

Matias F. Gerard[1], Raúl N. Comelli[2]

[1]Research Institute for Signals, Systems and Computational Intelligence (CONICET-UNL), Ciudad Universitaria, Santa Fe, Argentina.

[2]Depto. de Medio Ambiente, Fac. de Ingeniería y Ciencias Hídricas (FICH), Univ. Nacional del Litoral (UNL), Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET). Ciudad Universitaria CC 242 Paraje El Pozo, 3000, Santa Fe, Argentina.

1. Organization

This document aims to present some illustrative examples of how to configure and use PhDSeeker. A detailed explanation of the algorithm used in this tool is available in [1]. The examples shown here do not attempt to make an exhaustive explanation of the available options. They should only be considered as a guide to understand how the tool works. Whilst the following examples can be run using command-line or GUI-based tool, we preferred to explain how to build settings files and to use auxiliary tools by command-line since it could be a more complex task.

Each Section presents one possible scenario where PhDSeeker could be used. A brief description of the example and the database used are first presented. Then, we explain how to build the dataset for the selected database using the complementary tools provided with PhDSeeker. Next, the configuration file is setup and its main components are explained. Finally, it is explained how to perform the search, and found solutions are analyzed.

This document is organized as follows: Section 2 presents a simple example of searching for a pathway linking two compounds using a small set of reactions. Section 3 describes a harder task, where the goal is to find a pathway linking 4 compounds, but using the full set of reactions for *E. coli*. Here, a way to incorporate information of the problem is also discussed. Section 4 introduces two examples on how PhDSeeker could be used to design metabolic pathways with applications to metabolic engineering. Two designs reported in literature are analyzed, and PhDSeeker is setup to search for pathways linking the same compounds. The configuration files and those generated as a solution for each example presented below can be found in the `SupMat` folder in the source code of PhDSeeker.

2. Simple pathway linking 2 compounds

This Section introduces the use of PhDSeeker through two simple examples, involving only a source and a target compound. The first case explores the use of the Kyoto Encyclopedia of Genes and Genomes (KEGG) database for metabolic pathways searching. The second one explains how to search for pathways using this tool and the BiGG database of biological models.

2.1. Pathway searching using KEGG database

Currently, KEGG is an important database that stores information of around 6000 organisms, involving more than 11000 reactions and 18000 compounds[1]. This makes it an important resource when searching for metabolic pathways. For example, the well-known Glycolysis[2] pathway is available in

[1] Information extracted from `http://rest.kegg.jp/info/kegg` on 02/2019.

[2] `https://www.genome.jp/kegg-bin/show_pathway?rn00010`

KEGG under the code rn00010. This pathway is made up from the available information of all the organisms in the database and is comprised by 53 reactions, whereof only 27 are in the forward direction. Figure S1 shows a scheme of the KEGG the Glycolysis pathway.

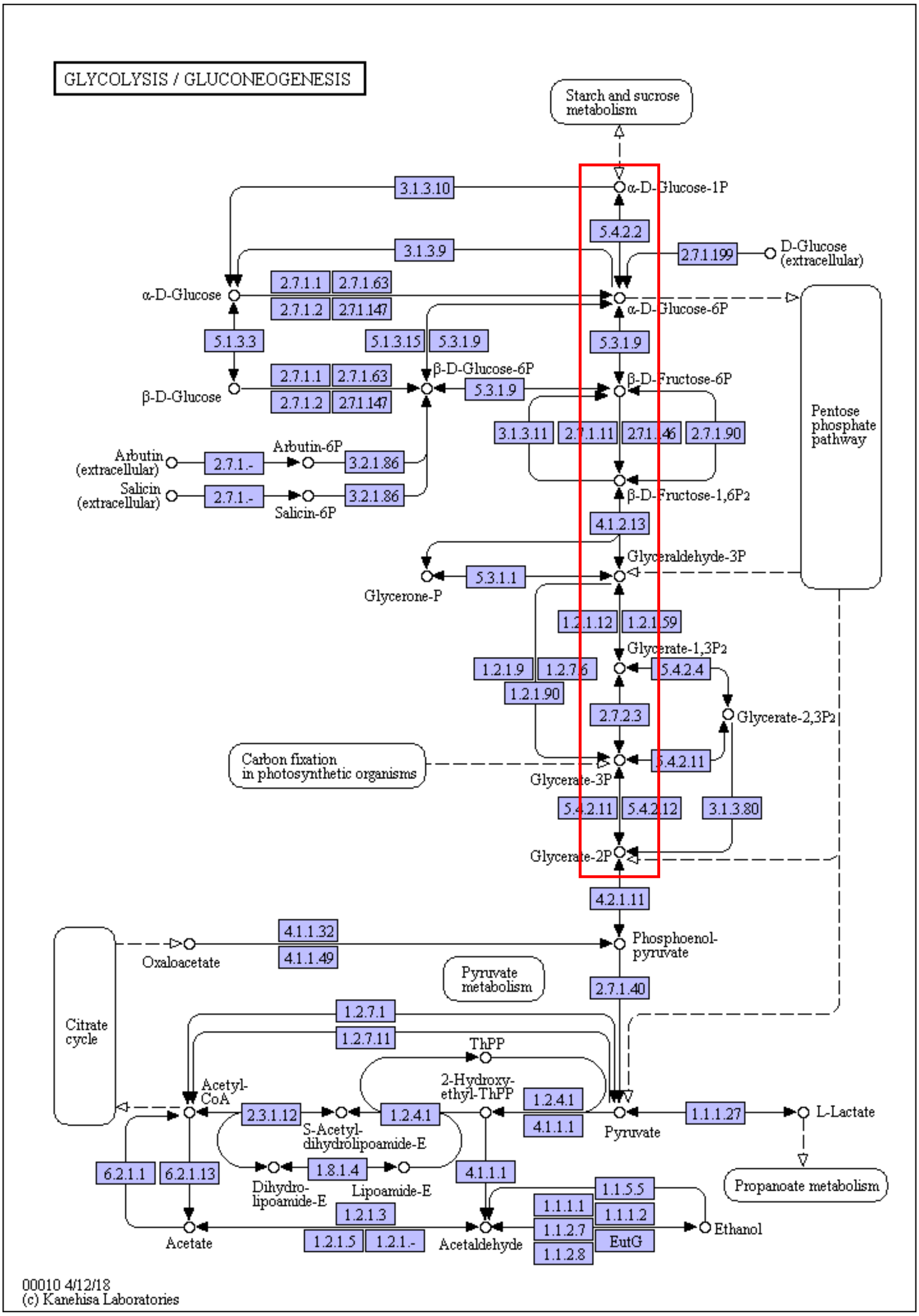
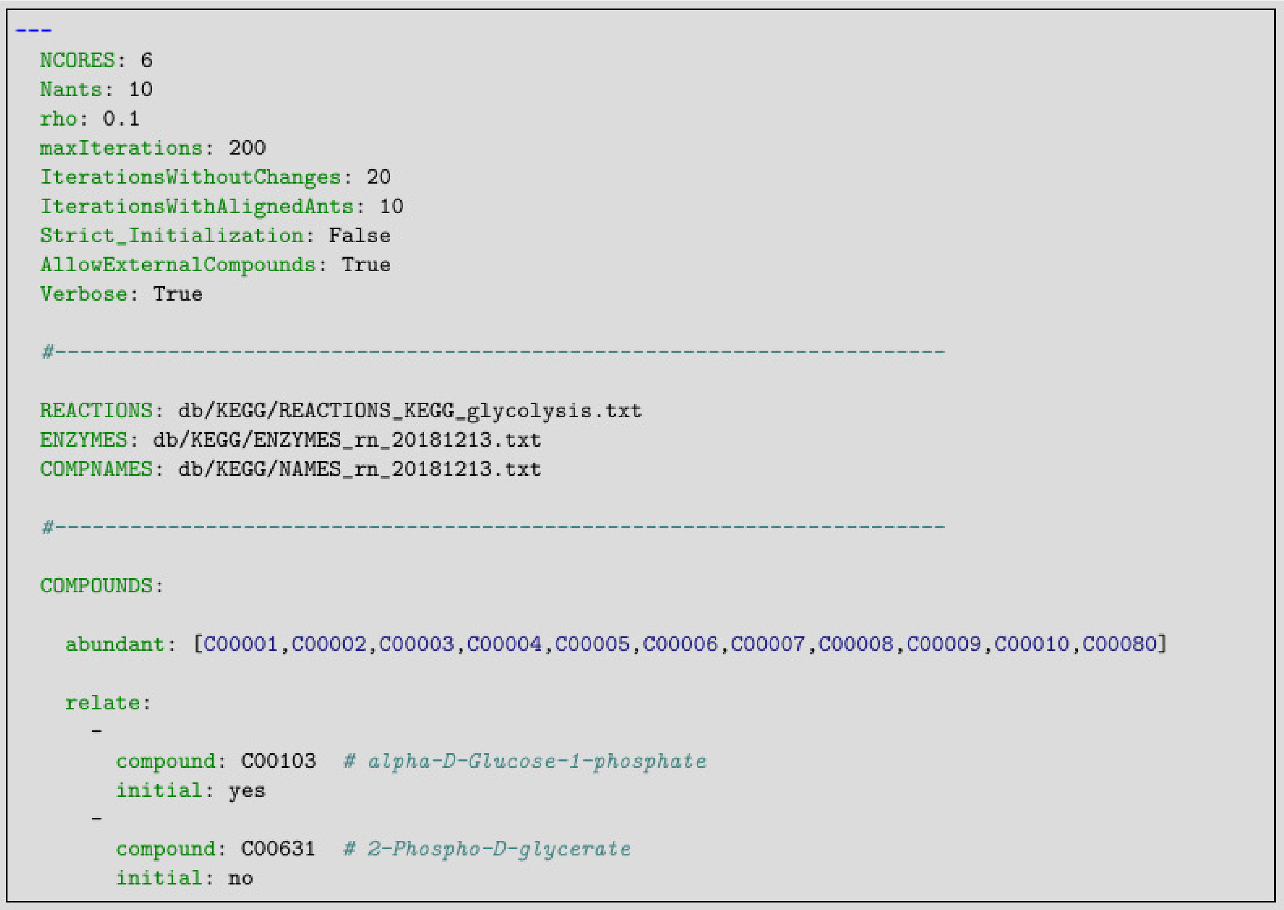


Figure S1: KEGG glycolysis pathway and connections to other known pathways. The red box shows an example of a pathway to produce phosphoenolpyruvate from α-D-glucose-1P. Image adapted from `https://www.genome.jp/kegg-bin/show_pathway?rn00010`.

Before carrying out the search with PhDSeeker it is necessary to construct the dataset of reactions to use. Construction of the required dataset can be easily performed with the KEGG2PHDS.py command-line tool provided with PhDSeeker (see PhDSeeker documentation for details). Thus, the



```
---
  NCORES: 6
  Nants: 10
  rho: 0.1
  maxIterations: 200
  IterationsWithoutChanges: 20
  IterationsWithAlignedAnts: 10
  Strict_Initialization: False
  AllowExternalCompounds: True
  Verbose: True

  #-----------------------------------------------------------------------

  REACTIONS: db/KEGG/REACTIONS_KEGG_glycolysis.txt
  ENZYMES: db/KEGG/ENZYMES_rn_20181213.txt
  COMPNAMES: db/KEGG/NAMES_rn_20181213.txt

  #-----------------------------------------------------------------------

  COMPOUNDS:

    abundant: [C00001,C00002,C00003,C00004,C00005,C00006,C00007,C00008,C00009,C00010,C00080]

    relate:
      -
        compound: C00103  # alpha-D-Glucose-1-phosphate
        initial: yes
      -
        compound: C00631  # 2-Phospho-D-glycerate
        initial: no
```

Figure S2: Experimental setup for searching a pathway linking D-Glucose-1-phosphate and 2-Phospho-D-glycerate. Information extracted from `SETTINGS_KEGG_glycolysis.yaml` file (`SupMat` folder).

following command must be executed to construct the dataset with reactions involved in Glycolysis.

```
>> python KEGG2PHDSFiles.py --organism rn --pathways rn00010
```

For simplicity, the code **rn** is chosen for this example, since this code integrates all reactions in KEGG, regardless the source organism. In order to work with a small search space and keep this as a simple example, only reactions from pathway rn00010 (Glycolysis) are used. However, if one wants to work with a given organism and specific pathways, it is only necessary to specify the code of the organism (”eco”, for *E. coli*), and the pathways of interest (ath00010 for Glycolysis in *A. thaliana*).

Resulting dataset is stored as three separated files:

- `REACTIONS_rn_1pathway_<yyyymmdd>.txt` (reactions that can be used to build a pathway)
- `ENZYMES_rn_1pathway_<yyyymmdd>.txt` (enzymes catalyzing each reaction)
- `NAMES_rn_1pathway_<yyyymmdd>.txt` (compound names for each compound code)

where <yyyymmdd> indicates <`year,month,day`> when the dataset was built.

The next step is setting up the search by properly defining parameters in the configuration file[3]. Figure S2 shows the setup used for this example. As it can be seen, the search is configured to find a pathway producing 2-Phospho-D-Glycerate from α-D-Glucose-1P (red box in Figure S1 shows a solution example).

It is of importance to note that all compounds are indicated with their KEGG codes, because substrates and products of reactions in the dataset are also specified with this notation. Also, the

[3] Detailed information of configuration parameters is available in the PhDSeeker documentation.

Table S1: Pathway proposed by PhDSeeker to produce 2-Phospho-D-glycerate from D-Glucose-1P.

R00959 (5.4.2.2 +1): 1 **D-Glucose-1P** → 1 **α-D-Glucose-6P**
R02740 (5.3.1.9): 1 **α-D-Glucose-6P** → 1 **β-D-Fructose-6P**
R04779 (2.7.1.11): 1 **ATP** + 1 **β-D-Fructose-6P** → 1 **ADP** + 1 **β-D-Fructose 1,6-P$_2$**
R01070 (4.1.2.13): 1 **β-D-Fructose 1,6-P$_2$** → 1 **Glycerone-1P** + 1 **D-Glyceraldehyde-3P**
R01058 (1.2.1.9 +1): 1 **H_2O** + 1 **NADP$^+$** + 1 **D-Glyceraldehyde-3P** → 1 **NADPH** + 1 **H$^+$** + 1 **3-Phospho-D-glycerate**
R01518 (5.4.2.11 +1): 1 **3-Phospho-D-glycerate** → 1 **2-Phospho-D-glycerate**

three files for our dataset build from Glycolysis are specified. This configuration file is provided with PhDSeeker, in the *SupMat* folder in the source code. Once the experimental setup is defined, PhDSeeker search is started with the following command:

```
>> python phdseeker.py --settings config/SETTINGS_KEGG_glycolysis.yaml
```

As a result, PhDSeeker returns three files:

- <yyyymmdd-hhmmss>__REACTIONS_KEGG_glycolysis__SUMMARY.txt
- <yyyymmdd-hhmmss>__REACTIONS_KEGG_glycolysis__PATHWAY.html
- <yyyymmdd-hhmmss>__REACTIONS_KEGG_glycolysis__HISTORY.txt

where <yyyymmdd-hhmmss> indicates <year,month,day-hour,minute,second> from when the execution was over. SUMMARY file contains a description of the parameters used in the search, some measures related to algorithm execution, compounds involved in the search, and the sequence of reactions that make up the pathway. PATHWAY file contains the interactive representation of the pathway found. HISTORY file contains a detail of the sequence of reactions for the different different pathways found all through out the search.

The SUMMARY file shows that the search required 31 iterations (1.7 seconds) to converge for a solution. As a reminder, in each iteration, each ant builds a pathway to produce 2-phosphoglycerate from α-D-glucose-1P. At the end, ants share information to improve solutions in the next iteration. Clearly, that helps in finding the solutions faster. Interestingly, HISTORY file shows that 4 pathways were proposed until the convergence to a solution. This indicates there is more than one way to establish the link between both compounds. Regarding the solution, Figure S1 shows the sequence of reactions for the pathway found. It is important to remark that since PhDSeeker is based on a metaheuristic algorithm, solutions may vary between different runs. Thus, a new run could produce a different solution to the one shown here. In order to facilitate the analysis of the pathway, the compounds are colored according to their function. Reactions are ordered from top to bottom according to the order they are carried out. This pathway contains 6 reactions, and consumes ATP but produces NADPH. Furthermore, note reactions R00959, R01058 and R01518 are catalyzed for more than one enzyme, as indicated by the symbol +. Figure S3 shows the schematic representation built by PhDSeeker for the same pathway. As explained in the README file provided with this tool, a contextual menu with links to the BRENDA database for each enzyme can be displayed by double clicking with the left button on the enzymes that have + in their description while pressing SHIFT key.

It is important to note that accessing the description of the enzymes (BRENDA database) provides the researcher with additional tools to select better alternatives. It is possible to select an enzyme (or metabolic branch) considering aspects not included in the current version of the program, such us the enzyme's affinity for the substrates or needing for cofactors. In addition, in the case of exogenous enzymes (absent in the host microorganism), preference for enzymes available in closely related organisms or with similar codon usage may be the key. For example, in the reaction R00959 (α-D-glucose-1P → D-glucose-6P catalyzed by the phosphoglucomutase), the program suggested two alternatives: EC 5.4.2.2 and EC 5.4.2.5. The difference between these isoenzymes is the distribution

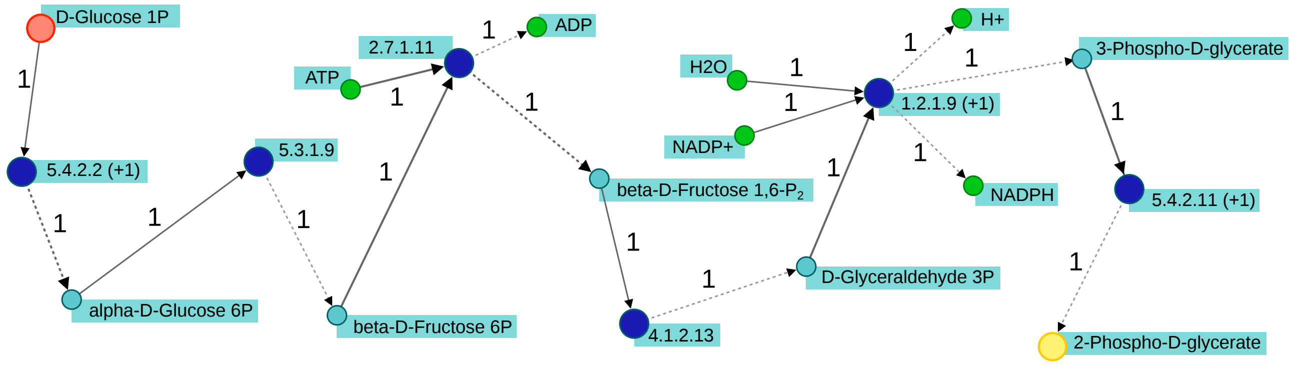


Figure S3: Pathway to produce 2-Phospho-D-Glycerate from α-D-Glucose-1P. Dark blue circles indicate reactions. Red and yellow circles correspond to source and target compounds, respectively. Green circles indicate freely available compounds. Light blue circles identify intermediary compounds.

around the kingdoms and the dependence by cofactors. The first is widely distributed among the kingdoms (Plantae, Animalia, Bacteria, Protozoa and Fungi) and it shown maximum activity only in the presence of α-D-glucose 1,6-bisphosphate (cofactor), while the EC 5.4.2.5 is activated by D-glucose (cofactor) and it is only present in several bacteria of the genus *Escherichia*, *Klebsiella* and *Mycobacterium*. The same applies to reaction R01058 (D-glyceraldehyde 3-phosphate + NADP+ + H_2O → 3-phospho-D-glycerate + NADPH + 2 H^+, catalyzed by glyceraldehyde-3-phosphate dehydrogenase), where the mentioned enzymes are widely distributed among the species (EC 1.2.1.9), are specific to a microorganism (EC 1.2.1.90 from *Thermoproteus tenax*), are broad spectrum (EC 1.2.7.5 catalyses the oxidation of several aldehydes to their corresponding acids) or have differential affinity for the cofactors (EC 1.2.1.12 and EC 1.2.1.13 are NAD^+- and $NADP^+$-dependent, respectively). Finally, for the reaction R01518 (2-phospho-D-glycerate → 3-phospho-D-glycerate catalyzed by phosphoglycerate mutase), the isoenzymes shown differential affinity for the same cofactors: EC 5.4.2.11 (from vertebrates, insects, yeast and some Gram-negative bacteria) require 2,3-bisphospho-D-glycerate as a cofactor and it has no requirement for metal ions; whereas the EC 5.4.2.12 (from higher plants, arachnids, archaea and some Gram-positive bacteria) contains two Mn^{2+} (or in some species two Co^{2+} ions) and it have maximum activity in the absence of 2,3-bisphospho-D-glycerate.

2.2. Pathway searching using BiGG database

The BiGG Models database is another interesting resource for metabolic pathways searching. It includes models for more than 60 organisms, comprising over 23000 reactions and 7000 compounds. While those model are commonly used to simulate and analyze the organism behavior under several conditions, they can also be used to provide a reactions base over which to perform the search of metabolic pathways.

Currently, the best BiGG model available for *E. coli* is named iJO1366[4]. This is made up of 2583 reactions and 1805 compounds. Construction of a dataset from this model can be easily done with the BiGG2PHDS.py command-line tool provided with PhDSeeker (see PhDSeeker documentation for details), simply by running the following command:

```
>> python BiGG2PHDSFiles.py --file BiGG/iJO1366.json
```

The dataset is made up of 2251 reactions, and only 1640 are in the forward direction. This contains 332 reactions less than the model because some irrelevant reactions for pathways searching were removed (324 exchange reactions, 6 intracellular demand reactions, and 2 biomass reactions). By default, this dataset is stored in three separated files:

- `REACTIONS_iJO1366.txt` (reactions that can be used to build a pathway)
- `ENZYMES_iJO1366.txt` (enzymes catalyzing each reaction)

[4] `http://bigg.ucsd.edu/models/iJO1366`

- `NAMES_iJO1366.txt` (compound names for each compound code)

It is important to remark that no information of enzymes is available in this model, so the `ENZYMES_iJO1366.txt` file is generated but is empty. However, this information can be extracted from the *universal_model* provided by BiGG. This model combines the information of the compounds, reactions and enzymes for every organism available in the database. Thus, it is possible to apply the BiGG2PHDSFiles.py tool to this model for building the reactions, enzymes and names files and use them to complete the missing information in the iJO1366 model. To avoid confusions with names (compartment information is not included in their names), we only use information of enzymes for this example.

The goal of this example is to find a pathway that produces citosolyc citrate (cit_c) from extracellular D-Glucose (glc__D_e), illustrating the use of compartments in metabolic pathways searching. According to this, the search was configured as shown in Figure S4, taking the recently generated dataset of reactions as search space. Parameter *external compounds* was set as *True*, forcing PhDSeeker to find a solution using only the compounds initially provided. If this was specified as *False*, PhDSeeker would automatically include in the list of freely available compounds all substrates of each reaction used in the pathway that requires D_glucose as substrate (see the documentation for more details of this parameter). In turn, the parameter *AllowExternalCompounds* was also set as *False*, in order to reduce the number of solutions. In order to keep the compartment notation in the compounds codes, the NAMES_iJO1366.txt file was not included in the settings. Compounds are indicated using BiGG notation because substrates and products of reactions in the dataset are also specified in the same way. Configuration file is provided with PhDSeeker, in the *SupMat* folder. Once the experimental setup is defined, PhDSeeker is executed with the following command:

```
>> python phdseeker.py --settings config/SETTINGS_BiGG_glc__D_e-cit_c.yaml
```

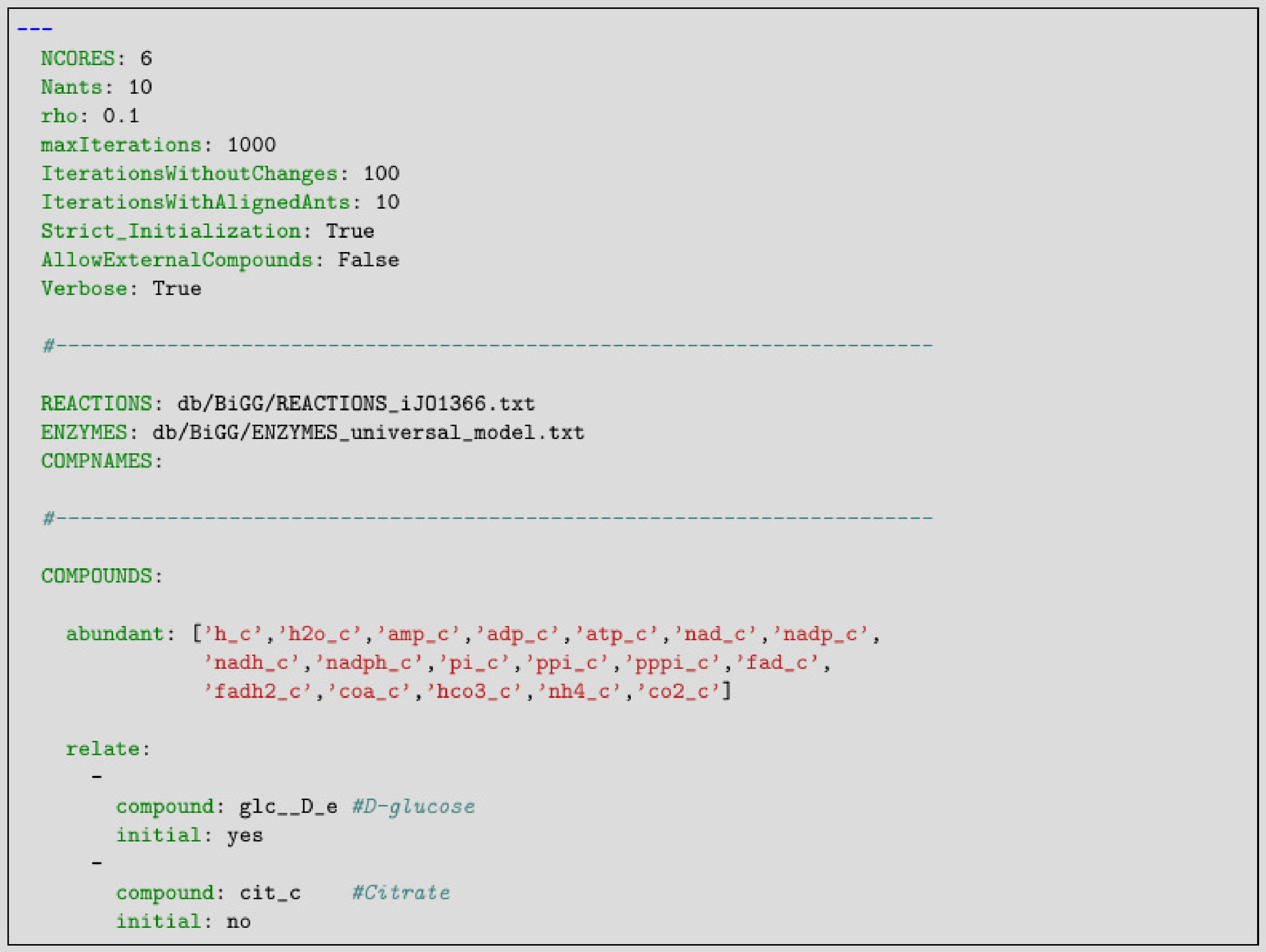
As a result, PhDSeeker returns three files:

- `<yyyymmdd-hhmmss>_BiGG_SUMMARY.txt`
- `<yyyymmdd-hhmmss>_BiGG_PATHWAY.html`
- `<yyyymmdd-hhmmss>_BiGG_HISTORY.txt`

where `<yyyymmdd-hhmmss>` indicates `<year,month,day-hour,minute,second>` from when the execution was over. SUMMARY file contains a description of the parameters used in the search, some measures related to algorithm execution, compounds involved in the search, and the sequence of reactions that make up the pathway. PATHWAY file contains the interactive representation of the pathway found. HISTORY file contains a detail of the sequence of reactions for the different different pathways found all through out the search.

The SUMMARY file shows that the search required 100 iterations (91 seconds) to converge for a solution. This reflects a good performance of the tool, mainly considering that more than 2000 reactions are available to build a solution, and that only 2 pathways were proposed until the convergence to a solution. In reference to the pathway found, Figure S2 shows the sequence of reactions comprising the solution. This pathway contains 8 reactions and two transporters to move glucose from the (outer) medium to periplasm (GLCtex_copy1) and the citoplasm (GLCabcpp). Figure S5 shows a schematic representation built by PhDSeeker for the same pathway. Note separation in layers to remark the transporting processes.

The analysis of the proposed pathway is of note. The glucose follows an oxidative metabolism through 6–phosphogluconate mediated by enzymes from the Entner-Doudoroff (ED) pathway. This route uses 6–phosphogluconate dehydratase (EC, coded by edd gene) and 2-keto-3-deoxyphosphogluconate aldolase (EC) to create pyruvate from glucose [2]. The same destiny for glucose is obtained in the classic glycolysis by the Embden-Meyerhof-Parnas (EMP) pathway. Because the ED pathway demands a minor set of reactions, the software will prefer it over the EMP pathway. As aforementioned, the last selection criteria is of the researcher.



```
---
  NCORES: 6
  Nants: 10
  rho: 0.1
  maxIterations: 1000
  IterationsWithoutChanges: 100
  IterationsWithAlignedAnts: 10
  Strict_Initialization: True
  AllowExternalCompounds: False
  Verbose: True

  #-----------------------------------------------------------------------

  REACTIONS: db/BiGG/REACTIONS_iJO1366.txt
  ENZYMES: db/BiGG/ENZYMES_universal_model.txt
  COMPNAMES:

  #-----------------------------------------------------------------------

  COMPOUNDS:

    abundant: ['h_c','h2o_c','amp_c','adp_c','atp_c','nad_c','nadp_c',
               'nadh_c','nadph_c','pi_c','ppi_c','pppi_c','fad_c',
               'fadh2_c','coa_c','hco3_c','nh4_c','co2_c']

    relate:
      -
        compound: glc__D_e #D-glucose
        initial: yes
      -
        compound: cit_c    #Citrate
        initial: no
```

Figure S4: Experimental setup for searching pathways to produce citrate from D-Glucose. Information extracted from `SETTINGS_BiGG_glc__D_e-cit_c.yaml` file (`SupMat` folder).

Table S2: Pathway proposed by PhDSeeker to produce cit_c (citosolyc Citrate) from glc__D_e (extracellular D-glucose). Compounds are colored according to their functions into the pathway. Reactions are ordered from top to bottom according to the order they are carried out.

GLCtex_copy1 (-.-.-.-): 1 **glc__D_e** → 1 **glc__D_p**
GLCabcpp (3.6.3.17): 1 **atp_c** + 1 **glc__D_p** + 1 **h2o_c** → 1 **adp_c** + 1 **glc__D_c** + 1 **h_c** + 1 **pi_c**
HEX1 (2.7.1.1 +1): 1 **atp_c** + 1 **glc__D_c** → 1 **adp_c** + 1 **g6p_c** + 1 **h_c**
G6PDH2r (1.1.1.363 +1): 1 **g6p_c** + 1 **nadp_c** → 1 **6pgl_c** + 1 **h_c** + 1 **nadph_c**
PGL (3.1.1.31): 1 **6pgl_c** + 1 **h2o_c** → 1 **6pgc_c** + 1 **h_c**
EDD (4.2.1.12): 1 **6pgc_c** → 1 **2ddg6p_c** + 1 **h2o_c**
EDA (4.1.2.14 +1): 1 **2ddg6p_c** → 1 **g3p_c** + 1 **pyr_c**
PPS (2.7.9.2): 1 **atp_c** + 1 **h2o_c** + 1 **pyr_c** → 1 **amp_c** + 2 **h_c** + 1 **pep_c** + 1 **pi_c**
PFL (2.3.1.54): 1 **coa_c** + 1 **pyr_c** → 1 **accoa_c** + 1 **for_c**
PPC (4.1.1.31): 1 **co2_c** + 1 **h2o_c** + 1 **pep_c** → 1 **h_c** + 1 **oaa_c** + 1 **pi_c**
CS (2.3.3.1 +2): 1 **accoa_c** + 1 **h2o_c** + 1 **oaa_c** → 1 **cit_c** + 1 **coa_c** + 1 **h_c**

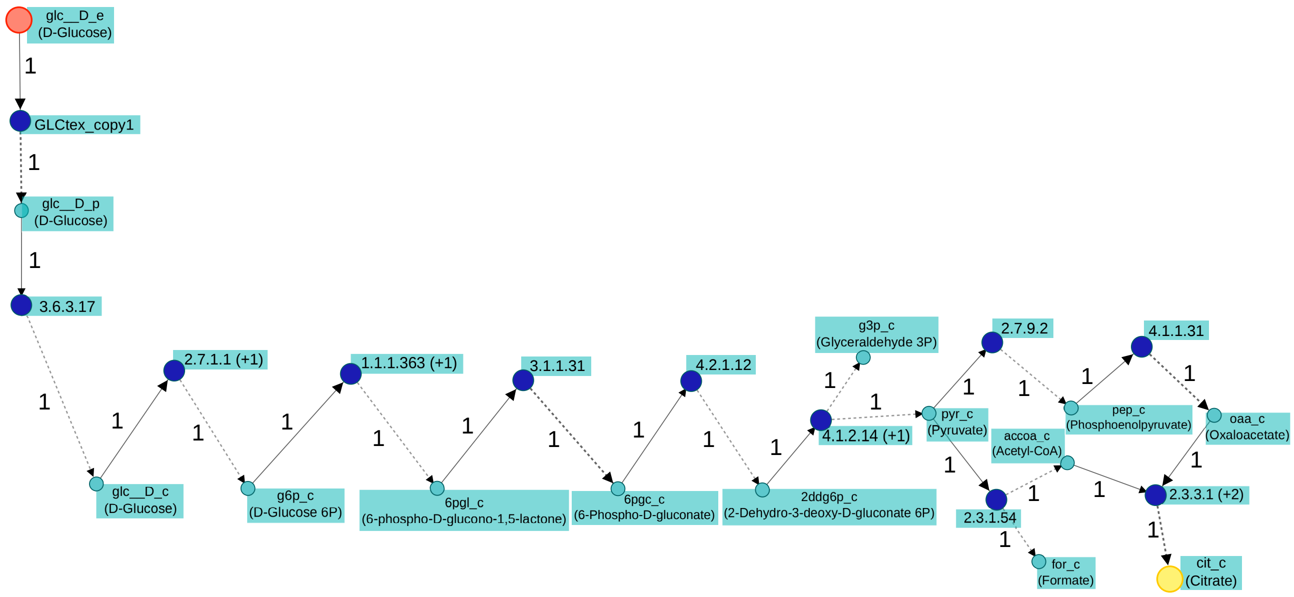


Figure S5: Pathway proposed by PhDSeeker to produce cit_c (citosolyc Citrate) from glc__D_e (extracellular D-glucose). Dark blue circles indicate reactions. Red and yellow circles correspond to source and target compounds, respectively. Green circles indicate freely available compounds. Light blue circles identify intermediary compounds for the synthesis of the target. Image extracted from the PhDSeeker pathway representation.

As a final remark, an interesting application for this is the evaluation of organisms capabilities to synthesize and excrete a given compound from a specific substrate. For example, considering the *E. coli* case studied, we can assess if this organism is capable of synthesizing extracellular D-xylose (xyl_D_e) from glucose (glc__D_e), simply changing cit_c to xyl_D_e in the settings file. By running this search, we will note that PhDSeeker is not able to find any solution, because D-xylose can only be consumed from the medium (according to the restrictions of the model). However, if the objective is to find a pathway regardless of the direction of the reactions (this information could be taken into account at some later stage), it is possible to eliminate this restriction simply building a dataset without the information of the sense of the reactions and performing the search with this dataset. While the pathway found may be unfeasible, it could serve as a basis for proposing a new pathway for the synthesis of this compound.

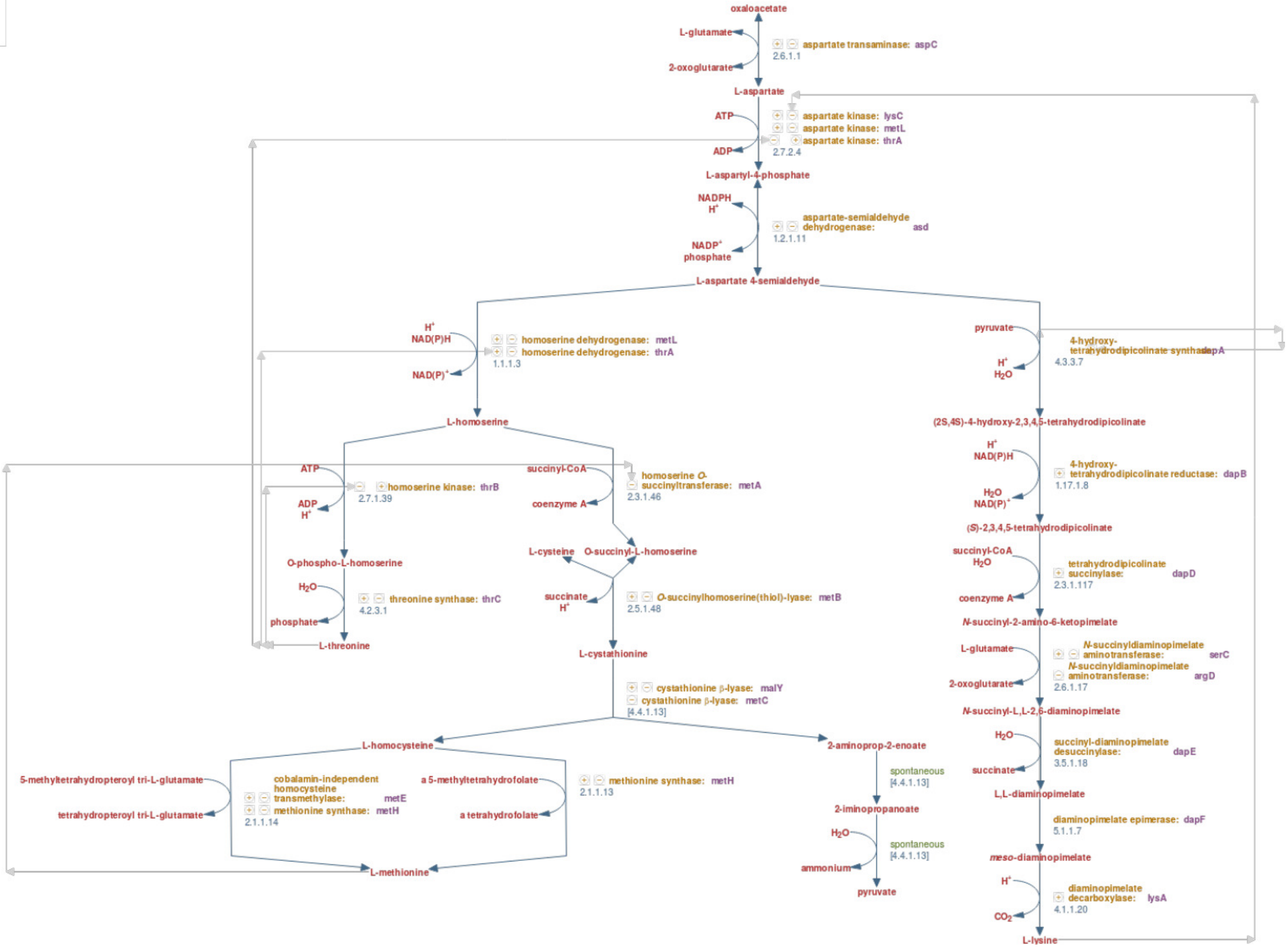

oxaloacetate
L-glutamate
2-oxoglutarate
aspartate transaminase: aspC
2.6.1.1
L-aspartate
ATP
ADP
aspartate kinase: lysC
aspartate kinase: metL
aspartate kinase: thrA
2.7.2.4
L-aspartyl-4-phosphate
NADPH
H+
NADP+
phosphate
aspartate-semialdehyde dehydrogenase: asd
1.2.1.11
L-aspartate 4-semialdehyde
H+
NAD(P)H
NAD(P)+
homoserine dehydrogenase: metL
homoserine dehydrogenase: thrA
1.1.1.3
L-homoserine
ATP
ADP
H+
homoserine kinase: thrB
2.7.1.39
O-phospho-L-homoserine
H2O
phosphate
threonine synthase: thrC
4.2.3.1
L-threonine
succinyl-CoA
coenzyme A
homoserine O-succinyltransferase: metA
2.3.1.46
L-cysteine
O-succinyl-L-homoserine
succinate
H+
O-succinylhomoserine(thiol)-lyase: metB
2.5.1.48
L-cystathionine
cystathionine β-lyase: malY
cystathionine β-lyase: metC
[4.4.1.13]
L-homocysteine
5-methyltetrahydropteroyl tri-L-glutamate
tetrahydropteroyl tri-L-glutamate
cobalamin-independent homocysteine transmethylase: metE
methionine synthase: metH
2.1.1.14
a 5-methyltetrahydrofolate
a tetrahydrofolate
methionine synthase: metH
2.1.1.13
L-methionine
2-aminoprop-2-enoate
spontaneous
[4.4.1.13]
2-iminopropanoate
H2O
ammonium
spontaneous
[4.4.1.13]
pyruvate
pyruvate
H+
H2O
4-hydroxy-tetrahydrodipicolinate synthase: dapA
4.3.3.7
(2S,4S)-4-hydroxy-2,3,4,5-tetrahydrodipicolinate
H+
NAD(P)H
H2O
NAD(P)+
4-hydroxy-tetrahydrodipicolinate reductase: dapB
1.17.1.8
(S)-2,3,4,5-tetrahydrodipicolinate
succinyl-CoA
H2O
coenzyme A
tetrahydrodipicolinate succinylase: dapD
2.3.1.117
N-succinyl-2-amino-6-ketopimelate
L-glutamate
2-oxoglutarate
N-succinyldiaminopimelate aminotransferase: serC
N-succinyldiaminopimelate aminotransferase: argD
2.6.1.17
N-succinyl-L,L-2,6-diaminopimelate
H2O
succinate
succinyl-diaminopimelate desuccinylase: dapE
3.5.1.18
L,L-diaminopimelate
diaminopimelate epimerase: dapF
5.1.1.7
meso-diaminopimelate
H+
CO2
diaminopimelate decarboxylase: lysA
4.1.1.20
L-lysine

Figure S6: Superpathway of L-lysine, L-threonine and L-methionine biosynthesis I. Image extracted from `https://biocyc.org/ECOLI/new-image?object=P4-PWY`

3. Complex pathway linking several compounds

While commonly the goal in pathway searching is finding a pathway between 2 compounds, there are situations which require to relate a larger number of them. For example, the *Superpathway of L-lysine, L-threonine and L-methionine biosynthesis I* in *E. coli*[5] describe the synthesis of three aminoacids from oxaloacetate. Figure S6 shows a scheme of this pathway, extracted from EcoCyc. This starts from *oxaloacetate* and synthesizes the aminoacids through three branches that share some reactions. Furthermore, *L-methionine* can be synthesized by two independent reactions. For the remaining of this section, we will consider this pathway as our reference.

Despite a search between pairs of compounds could be used to built this pathway, it is still necessary to know in advance that oxaloacetate is the initial substrate. Furthermore, a disadvantage of this strategy is that the search would not take advantage of intermediate compounds that can be shared between branches of the same pathway, and so it would be necessary for some mechanism to merge solutions.

In general, three different situations may arise when looking for a pathway that relates multiple compounds:

1. the initial substrate and some key compounds for the synthesis are known,
2. only the initial substrate is known,
3. there is no information on how the compounds relate.

[5] `https://biocyc.org/ECOLI/new-image?object=P4-PWY`

PhDSeeker can tackles these issues for the same dataset, only specifying different search conditions. To demonstrate this, we will look for pathways that relate oxaloacetate, L-lysine, L-threonine and L-methionine (see Figure S6), using the reactions for *E. coli* in the EcoCyc v21.5 database, considering the different situations described.

The dataset was built using the `compounds.dat` and `reactions.dat` files of the EcoCyc v21.5 database, and the command-line tools[6] provided with PhDSeeker. It was performed by running the following command:

```
>> python BioCyc2PHDSFiles.py --cpdDB BioCyc/compounds.dat
                              --rxnDB BioCyc/reactions.dat
                              --allreversible False
```

Note that the command was separated in three lines to keep it from exceeding the margins of the sheet. However, it must be typed on a single line, separated by single space. The option `--allreversible` is set as `False`, in order to take into account the available information of reactions reversibility. This dataset contains a total of 2380 reactions of which only 467 are in the forward direction. Information is stored in three separated files:

- `REACTIONS_<organism_name>_<yyyymmdd>.txt` (reactions that can be used to build a pathway)
- `ENZYMES_<organism_name>_<yyyymmdd>.txt` (enzymes catalyzing each reaction)
- `NAMES_<organism_name>_<yyyymmdd>.txt` (compound names for each compound code)

where <`organism_name`> indicates the name of the organism for this database (*Escherichia coli K-12 substr. MG1655* for EcoCyc 21.5, automatically extracted from `compounds.dat`), and <`yyyymmdd`> indicates <`year,month,day`> when the dataset was built. Using these files, the three scenarios are analyzed in the following subsections.

3.1. Search using information on required compounds and initial substrate

Available information about compounds can be easily used by PhDSeeker, simply adding it into the configuration file. Figure S7 shows the setup for searching the pathway of Figure S6. Oxaloacetate is specified as the initial substrate by setting *yes* in the corresponding `initial` label. As it is observed, some particular compounds are required to carry out specific reactions. For example, the first reaction (catalyzed by EC 2.6.1.1) requires *L-glutamate*, while synthesis of *L-methionine* needs *5-methyltetrahydropteroyl tri-L-glutamate* or *5-methyltetrahydrofolate*. Since this information is known in advance, this can be included into the list of freely available compounds in the configuration file, in order to avoid PhDSeeker to synthesize them. After analyzing the pathway, compounds *GLT*, *SUC-COA*, *CYS*, *CPD-1302* are added into the list of `abundant` compounds. Note that this list also includes cofactors and other compounds which are typically available in the cell. Once the experimental setup is defined, PhDSeeker is run with the following command:

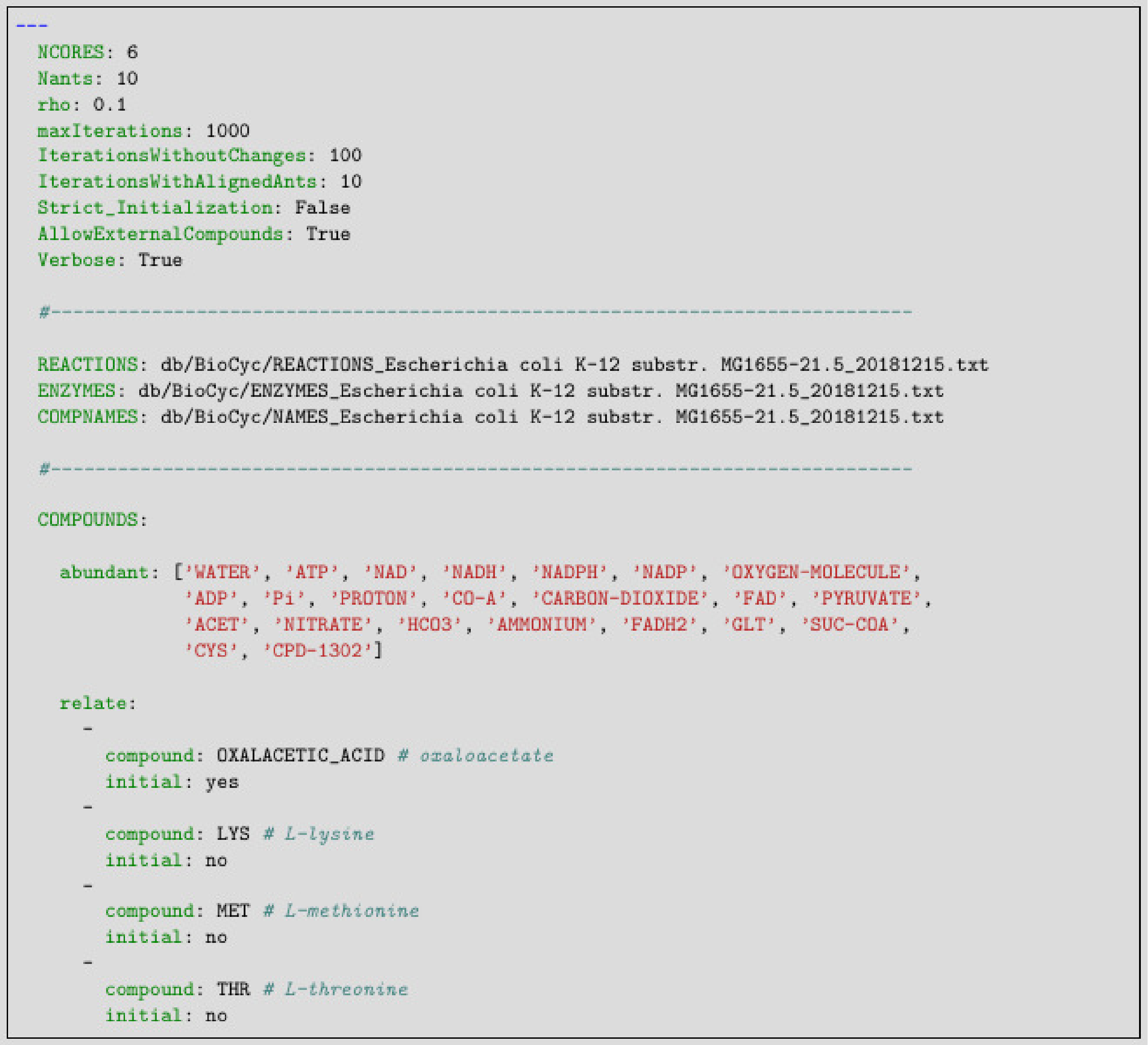
```
>> python phdseeker.py --settings config/SETTINGS_EcoCyc_3AA_from_oxaloacetate_1.yaml
```

As a result, PhDSeeker returns three files:

- <`yyyymmdd-hhmmss`>`_BioCyc_SUMMARY.txt`
- <`yyyymmdd-hhmmss`>`_BioCyc_PATHWAY.html`
- <`yyyymmdd-hhmmss`>`_BioCyc_HISTORY.txt`

where <`yyyymmdd-hhmmss`> indicates <`year,month,day-hour,minute,second`> from when the execution was over. SUMMARY file contains a description of the parameters used in the search, some

[6] See PhDSeeker documentation for a detailed explanation of how build the dataset from BioCyc files.



```
---
  NCORES: 6
  Nants: 10
  rho: 0.1
  maxIterations: 1000
  IterationsWithoutChanges: 100
  IterationsWithAlignedAnts: 10
  Strict_Initialization: False
  AllowExternalCompounds: True
  Verbose: True

  #-----------------------------------------------------------------------------

  REACTIONS: db/BioCyc/REACTIONS_Escherichia coli K-12 substr. MG1655-21.5_20181215.txt
  ENZYMES: db/BioCyc/ENZYMES_Escherichia coli K-12 substr. MG1655-21.5_20181215.txt
  COMPNAMES: db/BioCyc/NAMES_Escherichia coli K-12 substr. MG1655-21.5_20181215.txt

  #-----------------------------------------------------------------------------

  COMPOUNDS:

    abundant: ['WATER', 'ATP', 'NAD', 'NADH', 'NADPH', 'NADP', 'OXYGEN-MOLECULE',
               'ADP', 'Pi', 'PROTON', 'CO-A', 'CARBON-DIOXIDE', 'FAD', 'PYRUVATE',
               'ACET', 'NITRATE', 'HCO3', 'AMMONIUM', 'FADH2', 'GLT', 'SUC-COA',
               'CYS', 'CPD-1302']

    relate:
      -
        compound: OXALACETIC_ACID # oxaloacetate
        initial: yes
      -
        compound: LYS # L-lysine
        initial: no
      -
        compound: MET # L-methionine
        initial: no
      -
        compound: THR # L-threonine
        initial: no
```

Figure S7: Experimental setup for searching pathways to synthesize L-lysine, L-threonine and L-methionine from oxalacetate. Information extracted from SETTINGS_EcoCyc_3AA_from_oxaloacetate_1.yaml file (SupMat folder).

measures related to algorithm execution, compounds involved in the search, and the sequence of reactions that make up the pathway. PATHWAY file contains the interactive representation of the pathway found. HISTORY file contains a detail of the sequence of reactions for the different different pathways found all through out the search.

The SUMMARY file shows that around 200 iterations and approximately 43 minutes are required by PhDSeeker to provide automatically the final solution. However, the best solution is improved 6 times along search, as it can be seen in the HISTORY file. Furthermore, the first solution is obtained only after 22 seconds of execution (more than 2300 reactions must be taken into account for pathway building). Concerning the pathway found, Figure S3 shows the best solution reported. This pathway contains 17 reactions, from which 16 (in boldface) belong to pathway of Figure S6. In general terms, most of the pathway is recovered by this search, and the solution allows the synthesis of all aminoacids. The only apparent difference is the reaction CYSTATHIONINE-BETA-LYASE-RXN (listed in 9th position in Figure S3). In the pathway extracted from EcoCyC (Figure S6), the cleavage of the L-cystathionine is catalyzed by the cysteine-S-conjugate β-lyase (EC 4.4.1.13), a pyridoxal-phosphate protein that can act on a broad range of L-cysteine-S-conjugates, releasing a thiol (L-homocysteine in the case of L-cystathionine) and an unstable product that spontaneously yield pyruvate and ammonia. In the pathway proposed by PhDSeeker, the same reaction is catalyzed by the cystathionine β-lyase

Table S3: Pathway designed by PhDSeeker for synthesizing L-methionine, L-threonine and L-lysine from oxalacetate. Compounds are colored according to their functions in the pathway. Reactions are ordered from top to bottom according to the order they are carried out.

ASPAMINOTRANS-RXN (2.6.1.1): 1 L-glutamate + 1 oxaloacetate → 1 2-oxoglutarate + 1 L-aspartate

ASPARTATEKIN-RXN (2.7.2.4): 1 ATP + 1 L-aspartate → 1 ADP + 1 L-aspartyl-4P

ASPARTATE-SEMIALDEHYDE-DEHYDROGENASE-RXN (1.2.1.11): 1 L-aspartyl-4P + 1 NADPH + 1 H^+ → 1 L-aspartate-semialdehyde + 1 $NADP^+$ + 1 Pi

HOMOSERDEHYDROG-RXN (1.1.1.3): 1 L-aspartate-semialdehyde + 1 NADPH + 1 H^+ → 1 L-homoserine + 1 $NADP^+$

HOMOSERKIN-RXN (2.7.1.39): 1 ATP + 1 L-homoserine → 1 ADP + 1 O-phospho-L-homoserine + 1 H^+

THRESYN-RXN (4.2.3.1): 1 O-phospho-L-homoserine + 1 H_2O → 1 Pi + 1 L-threonine

HOMSUCTRAN-RXN (2.3.1.46): 1 L-homoserine + 1 succinyl-CoA → 1 CoA + 1 O-succinyl-L-homoserine

O-SUCCHOMOSERLYASE-RXN (2.5.1.48): 1 L-cysteine + 1 O-succinyl-L-homoserine → 1 L-cystathionine + 1 H^+ + 1 succinate

CYSTATHIONINE-BETA-LYASE-RXN (4.4.1.8): 1 L-cystathionine + 1 H_2O → 1 NH_4^+ + 1 L-homocysteine + 1 pyruvate

HOMOCYSMET-RXN (2.1.1.14): 1 5-methyltetrahydropteroyl tri-L-glutamate + 1 L-homocysteine → 1 tetrahydropteroyl tri-L-glutamate + 1 L-methionine

DIHYDRODIPICSYN-RXN (4.3.3.7): 1 L-aspartate-semialdehyde + 1 pyruvate → 1 (2S,4S)-4-hydroxy-2,3,4,5-tetrahydrodipicolinate + 1 H^+ + 1 H_2O

RXN-14014 (1.17.1.8): 1 (2S,4S)-4-hydroxy-2,3,4,5-tetrahydrodipicolinate + 1 NADH + 1 H^+ → 1 (S)-2,3,4,5-tetrahydrodipicolinate + 1 NAD^+ + 1 H_2O

TETHYDPICSUCC-RXN (2.3.1.117): 1 (S)-2,3,4,5-tetrahydrodipicolinate + 1 succinyl-CoA + 1 H_2O → 1 CoA + 1 N-succinyl-2-amino-6-ketopimelate

SUCCINYLDIAMINOPIMTRANS-RXN (2.6.1.17): 1 L-glutamate + 1 N-succinyl-2-amino-6-ketopimelate → 1 2-oxoglutarate + 1 N-succinyl-L,L-2,6-diaminopimelate

SUCCDIAMINOPIMDESUCC-RXN (3.5.1.18): 1 N-succinyl-L,L-2,6-diaminopimelate + 1 H_2O → 1 L,L-diaminopimelate + 1 succinate

DIAMINOPIMEPIM-RXN (5.1.1.7): 1 L,L-diaminopimelate → 1 meso-diaminopimelate

DIAMINOPIMDECARB-RXN (4.1.1.20): 1 meso-diaminopimelate + 1 H^+ → 1 CO_2 + 1 L-lysine

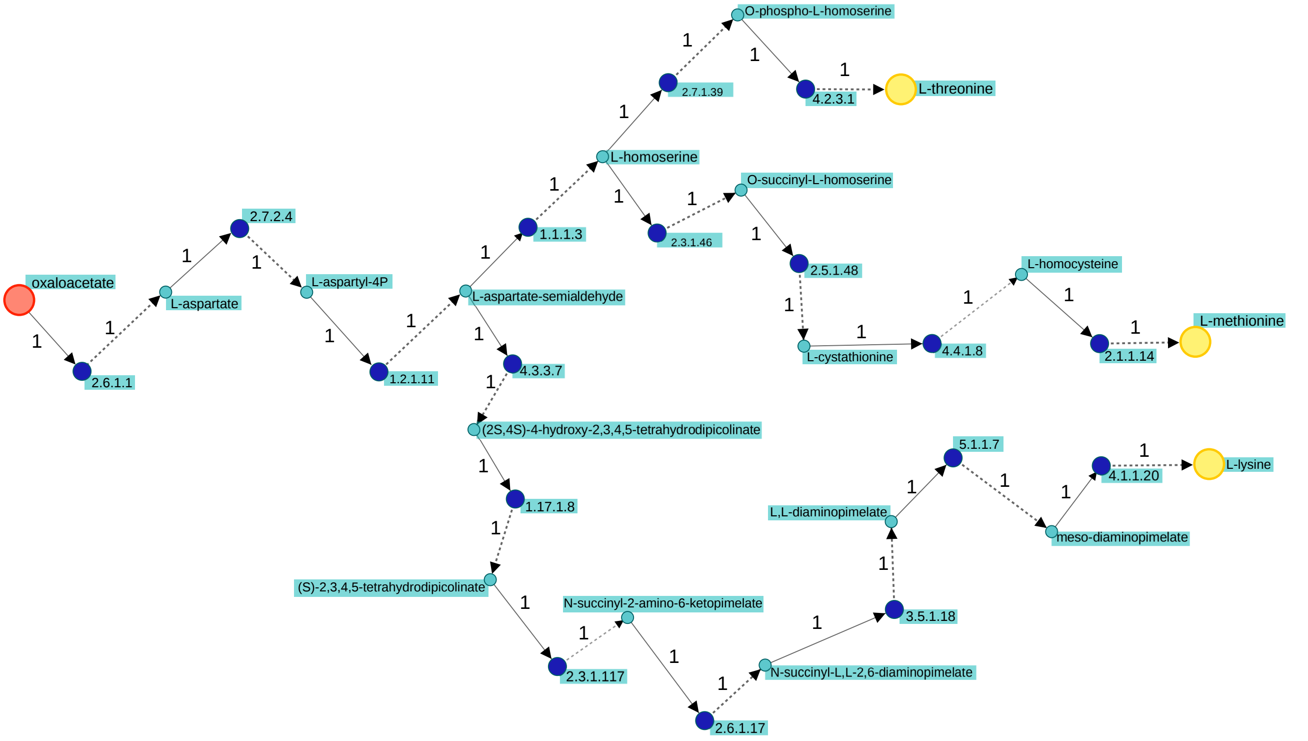
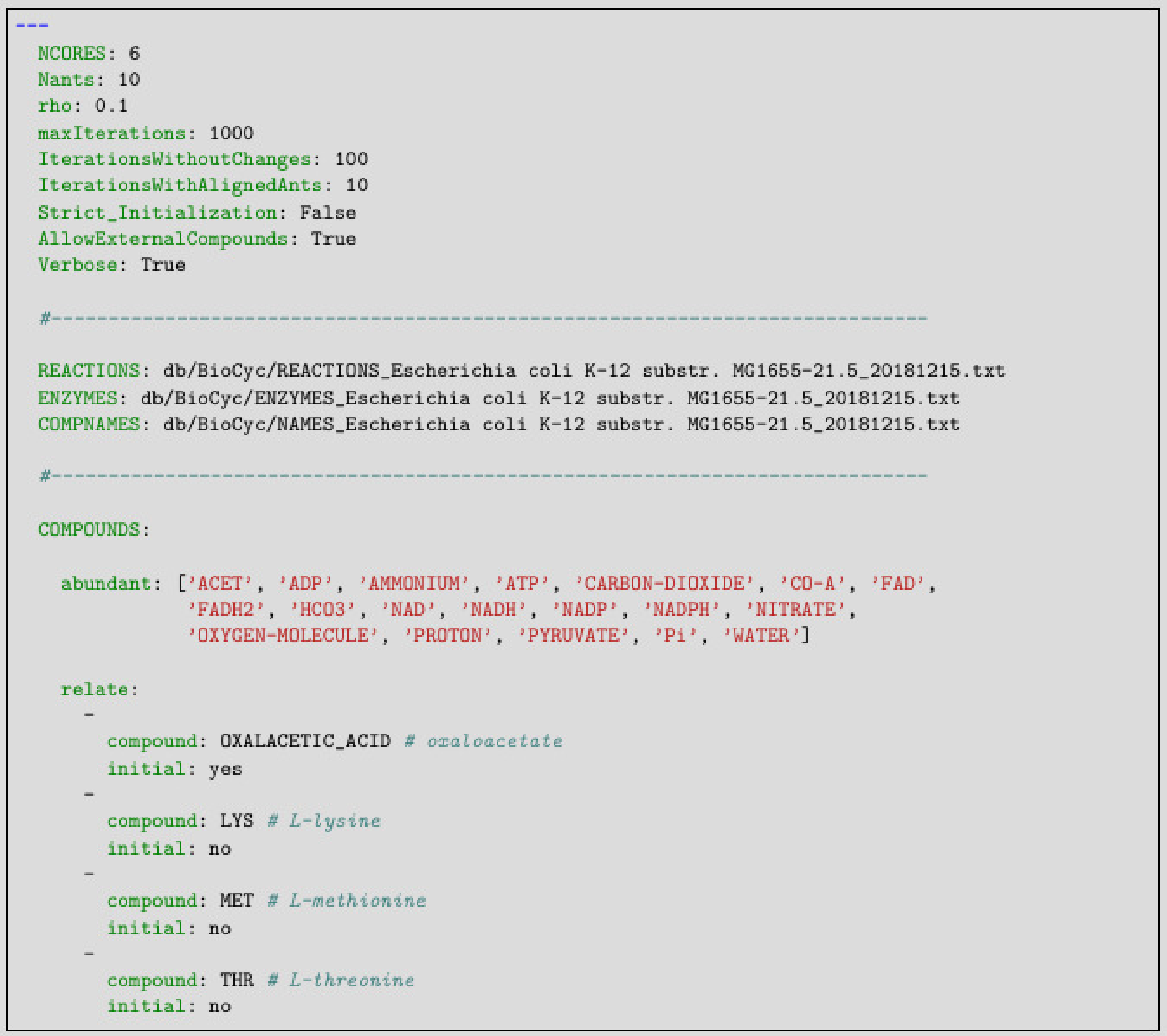


Figure S8: Pathway designed by PhDSeeker for synthesizing L-methionine, L-threonine and L-lysine from oxalacetate. For better interpretation, freely available compounds are not shown in the figure.

(EC 4.4.1.8), another pyridoxal-phosphate protein but which seem to be more specific to cleave the carbon-sulfur bond present in the L-cystathionine. In conclusion, both enzymes catalyze the same reaction. So, the pathway proposed by PhDSeeker is in agreement with the reported in the literature. Figure S8 shows the schematic representation built by PhDSeeker for this pathway.



```
---
  NCORES: 6
  Nants: 10
  rho: 0.1
  maxIterations: 1000
  IterationsWithoutChanges: 100
  IterationsWithAlignedAnts: 10
  Strict_Initialization: False
  AllowExternalCompounds: True
  Verbose: True

  #------------------------------------------------------------------------------

  REACTIONS: db/BioCyc/REACTIONS_Escherichia coli K-12 substr. MG1655-21.5_20181215.txt
  ENZYMES: db/BioCyc/ENZYMES_Escherichia coli K-12 substr. MG1655-21.5_20181215.txt
  COMPNAMES: db/BioCyc/NAMES_Escherichia coli K-12 substr. MG1655-21.5_20181215.txt

  #------------------------------------------------------------------------------

  COMPOUNDS:

    abundant: ['ACET', 'ADP', 'AMMONIUM', 'ATP', 'CARBON-DIOXIDE', 'CO-A', 'FAD',
               'FADH2', 'HCO3', 'NAD', 'NADH', 'NADP', 'NADPH', 'NITRATE',
               'OXYGEN-MOLECULE', 'PROTON', 'PYRUVATE', 'Pi', 'WATER']

    relate:
      -
        compound: OXALACETIC_ACID # oxaloacetate
        initial: yes
      -
        compound: LYS # L-lysine
        initial: no
      -
        compound: MET # L-methionine
        initial: no
      -
        compound: THR # L-threonine
        initial: no
```

Figure S9: Experimental setup for searching pathways to synthesize L-lysine, L-threonine and L-methionine from oxalacetate. All compounds are indicated with the same encoding used in the reactions dataset. Information extracted from SETTINGS_EcoCyc_3AA_from_oxaloacetate_2.yaml file (SupMat folder).

3.2. Search using only information of initial substrate

One important feature of PhDSeeker is the capability of finding ways to synthesize compounds that could be required but which are not available in the initial list of freely available ones. To this end, it incorporates as many feasible reactions (from the available compounds) as it would be required to synthesize the missing compounds needed to the pathway under construction.

Supposing the information of substrates for some reactions is not available (the opposite situation to that shown in the previous section), PhDSeeker still might be able to find a pathway to synthesize the three aminoacids from oxaloacetate. Taking this into account, configuration file was setup according to information in Figure S9. As it can be seen, compounds *GLT*, *SUC-COA*, *CYS*, and *CPD-1302* were removed from the list of abundant ones. Once the experimental setup is defined, PhDSeeker is run with the following command:

```
>> python phdseeker.py --settings config/SETTINGS_EcoCyc_3AA_from_oxaloacetate_2.yaml
```

As a result, PhDSeeker returns three files:

Table S4: Pathway designed by PhDSeeker for synthesizing *L-methionine*, *L-threonine* and *L-lysine* from *oxalacetate*. Compounds are colored according to their function. Reactions are ordered from top to bottom according to the order they are carried out.

RXN-2962 (1.1.1.284): 1 NAD^+ + 1 S-(hydroxymethyl)glutathione → 1 S-formylglutathione + 1 NADH + 1 H^+
S-FORMYLGLUTATHIONE-HYDROLASE-RXN (3.1.2.12): 1 S-formylglutathione + 1 H_2O → 1 formate + 1 glutathione + 1 H^+
RXN-12618 (3.4.19.13): 1 glutathione + 1 H_2O → 1 L-cysteinyl-glycine + 1 L-glutamate
RXN-6622 (3.4.13.18): 1 L-cysteinyl-glycine + 1 H_2O → 1 L-cysteine + 1 glycine
RXN0-6984 (3.4.13.18): 1 glycyl-L-glutamate + 1 H_2O → 1 L-glutamate + 1 glycine
ASPAMINOTRANS-RXN (2.6.1.1): 1 L-glutamate + 1 oxaloacetate → 1 2-oxoglutarate + 1 L-aspartate
ASPARTATEKIN-RXN (2.7.2.4): 1 ATP + 1 L-aspartate → 1 ADP + 1 L-aspartyl-4P
ASPARTATE-SEMIALDEHYDE-DEHYDROGENASE-RXN (1.2.1.11): 1 L-aspartyl-4P + 1 NADPH + 1 H^+ → 1 L-aspartate-semialdehyde + 1 $NADP^+$ + 1 Pi
HOMOSERDEHYDROG-RXN (1.1.1.3): 1 L-aspartate-semialdehyde + 1 NADH + 1 H^+ → 1 L-homoserine + 1 NAD^+
HOMOSERKIN-RXN (2.7.1.39): 1 ATP + 1 L-homoserine → 1 ADP + 1 O-phospho-L-homoserine + 1 H^+
THRESYN-RXN (4.2.3.1): 1 O-phospho-L-homoserine + 1 H_2O → 1 Pi + 1 L-threonine
DIHYDRODIPICSYN-RXN (4.3.3.7): 1 L-aspartate-semialdehyde + 1 pyruvate → 1 (2S,4S)-4-hydroxy-2,3,4,5-tetrahydrodipicolinate + 1 H^+ + 1 H_2O
RXN-14014 (1.17.1.8): 1 (2S,4S)-4-hydroxy-2,3,4,5-tetrahydrodipicolinate + 1 NADPH + 1 H^+ → 1 (S)-2,3,4,5-tetrahydrodipicolinate + 1 $NADP^+$ + 1 H_2O
2OXOGLUTARATEDEH-RXN (-.-.-.-): 1 2-oxoglutarate + 1 CoA + 1 NAD^+ → 1 CO^2 + 1 NADH + 1 succinyl-CoA
TETHYDPICSUCC-RXN (2.3.1.117): 1 (S)-2,3,4,5-tetrahydrodipicolinate + 1 succinyl-CoA + 1 H_2O → 1 CoA + 1 N-succinyl-2-amino-6-ketopimelate
SUCCINYLDIAMINOPIMTRANS-RXN (2.6.1.17): 1 L-glutamate + 1 N-succinyl-2-amino-6-ketopimelate → 1 2-oxoglutarate + 1 N-succinyl-L,L-2,6-diaminopimelate
SUCCDIAMINOPIMDESUCC-RXN (3.5.1.18): 1 N-succinyl-L,L-2,6-diaminopimelate + 1 H_2O → 1 L,L-diaminopimelate + 1 succinate
DIAMINOPIMEPIM-RXN (5.1.1.7): 1 L,L-diaminopimelate → 1 meso-diaminopimelate
TRANS-RXN-291 (3.6.3): 1 ATP + 1 meso-diaminopimelate + 1 H_2O → 1 ADP + 1 meso-diaminopimelate + 1 H^+ + 1 Pi
DIAMINOPIMDECARB-RXN (4.1.1.20): 1 meso-diaminopimelate + 1 H^+ → 1 CO_2 + 1 L-lysine
ABC-3-RXN (3.6.3.21): 1 ATP + 1 L-lysine + 1 H_2O → 1 ADP + 1 L-lysine + 1 H^+ + 1 Pi
HOMSUCTRAN-RXN (2.3.1.46): 1 L-homoserine + 1 succinyl-CoA → 1 CoA + 1 O-succinyl-L-homoserine
RXN-15147 (-.-.-.-): 1 O-succinyl-L-homoserine → 1 (2Z)-2-aminobut-2-enoate + 1 H^+ + 2 succinate
RXN-15122 (-.-.-.-): 1 (2Z)-2-aminobut-2-enoate + 1 H^+ + 1 H_2O → 1 L-threonine
O-SUCCHOMOSERLYASE-RXN (2.5.1.48): 1 L-cysteine + 1 O-succinyl-L-homoserine → 1 L-cystathionine + 1 H^+ + 1 succinate
RXN-15131 (-.-.-.-): 1 L-cystathionine → 1 2-aminoprop-2-enoate + 1 L-homocysteine
CYSTATHIONINE-BETA-LYASE-RXN (4.4.1.8): 1 L-cystathionine + 1 H_2O → 1 NH_4^+ + 1 L-homocysteine + 1 pyruvate
MMUM-RXN (2.1.1.10): 1 S-methyl-L-methionine + 1 L-homocysteine → 2 L-methionine + 1 H^+

- `<yyyymmdd-hhmmss>__Escherichia coli K-12 substr. MG1655-21.5__SUMMARY.txt`
- `<yyyymmdd-hhmmss>__Escherichia coli K-12 substr. MG1655-21.5__PATHWAY.html`
- `<yyyymmdd-hhmmss>__Escherichia coli K-12 substr. MG1655-21.5__HISTORY.txt`

where `<yyyymmdd-hhmmss>` indicates `<year,month,day-hour,minute,second>` from when the execution was over. SUMMARY file contains a description of the parameters used in the search, some measures related to algorithm execution, compounds involved in the search, and the sequence of reactions that make up the pathway. PATHWAY file contains the interactive representation of the pathway found. HISTORY file contains a detail of the sequence of reactions for the different different pathways found all through out the search.

As it is shown in the SUMMARY file, 336 iterations and around 1 hour are required to return the solution. However, note that the first solution in the HISTORY file is returned after 19 seconds, and this is then improved 14-times until converge to the final metabolic pathway. The sequence of reactions comprising the best solution is presented in Figure S4. As it can be seen, this pathway involves 2 transporters and 26 reactions, from which only 16 strictly belong to the pathway of Figure S6. Clearly, this solution contains unnecessary or redundant reactions, that probably might be filtered specifying a greater number of iterations into the configuration file. However, this also implies a longer search time. Note that while reaction RXN-15131 is not catalyzed by any enzyme in the solution found (we used EcoCyc v21.5), in EcoCyc database v22.5 it is catalyzed by enzyme 4.4.1.13. This solution proposes the same mechanisms than does the reference pathway to synthesize `L-lysine` and `L-threonine`. Moreover, this also incorporates additional mechanisms to their synthesis. For example, `L-threonine` is additionally synthesized from `L-homoserine` following the sequence `2.3.1.46`, `RXN-15147`, `RXN-15122`. Interestingly, the main difference with the reference pathway is how L-methionine is synthesized. Instead of using the reactions catalyzed by 2.1.1.13/14 (5-methyltetrahydropteroyl tri-L-glutamate is

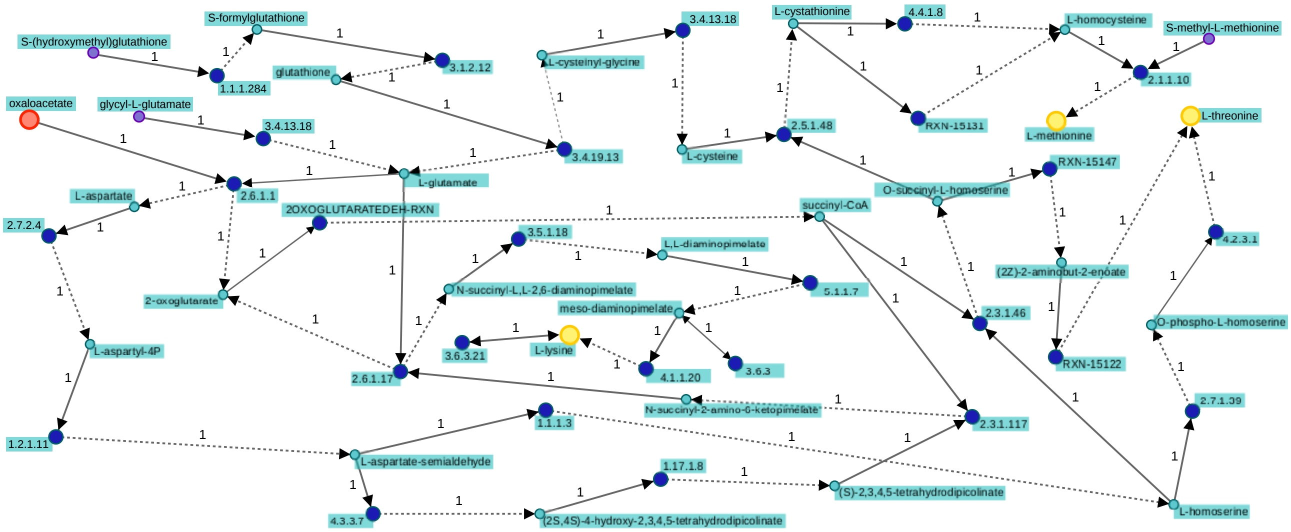
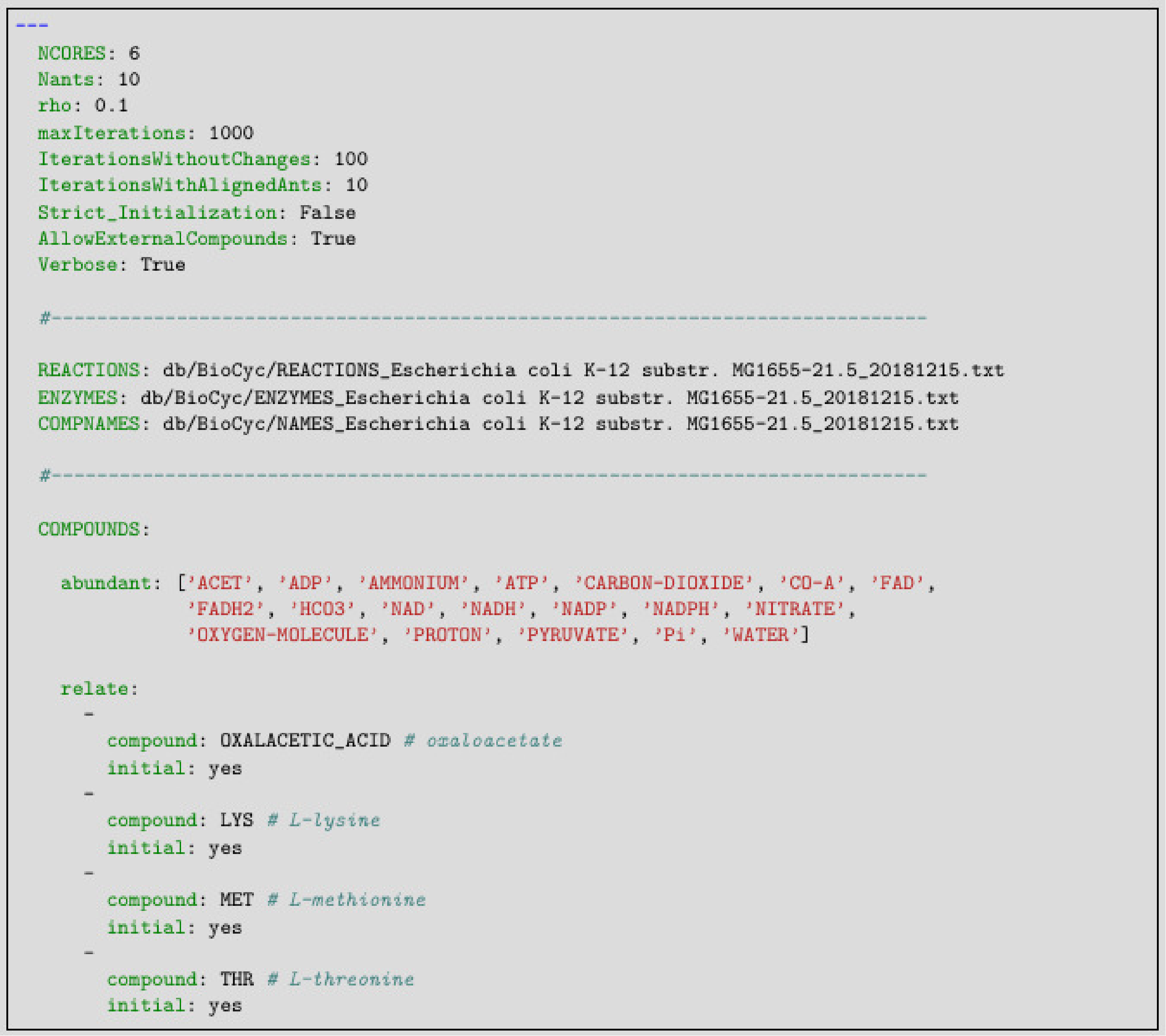


Figure S10: Pathway designed by PhDSeeker for synthesizing L-methionine, L-threonine and L-lysine from oxalacetate without information of available compounds. For better interpretation, freely available compounds are not shown in the figure.

automatically identified as a non-synthesizable compound and included by PhDSeeker into the list of freely available ones), PhDSeeker chooses `MMUM-RXN` (2.1.1.10). It is not surprising because both alternatives only require one step for synthesizing the aminoacid and therefore they are similar for the tool.

As a final remark, allowing external compounds in the search helps to expand the range of solutions to be found. Even though not including them reduces search space and exploration time, this is at the expense of clipping the set of possible solutions. Clearly, `L-glutamate` and `L-methionine` could not be synthesized without external compounds.



```
---
  NCORES: 6
  Nants: 10
  rho: 0.1
  maxIterations: 1000
  IterationsWithoutChanges: 100
  IterationsWithAlignedAnts: 10
  Strict_Initialization: False
  AllowExternalCompounds: True
  Verbose: True

  #------------------------------------------------------------------------------

  REACTIONS: db/BioCyc/REACTIONS_Escherichia coli K-12 substr. MG1655-21.5_20181215.txt
  ENZYMES: db/BioCyc/ENZYMES_Escherichia coli K-12 substr. MG1655-21.5_20181215.txt
  COMPNAMES: db/BioCyc/NAMES_Escherichia coli K-12 substr. MG1655-21.5_20181215.txt

  #------------------------------------------------------------------------------

  COMPOUNDS:

    abundant: ['ACET', 'ADP', 'AMMONIUM', 'ATP', 'CARBON-DIOXIDE', 'CO-A', 'FAD',
               'FADH2', 'HCO3', 'NAD', 'NADH', 'NADP', 'NADPH', 'NITRATE',
               'OXYGEN-MOLECULE', 'PROTON', 'PYRUVATE', 'Pi', 'WATER']

    relate:
      -
        compound: OXALACETIC_ACID # oxaloacetate
        initial: yes
      -
        compound: LYS # L-lysine
        initial: yes
      -
        compound: MET # L-methionine
        initial: yes
      -
        compound: THR # L-threonine
        initial: yes
```

Figure S11: Experimental setup for searching pathways to synthesize L-lysine, L-threonine and L-methionine from oxalacetate. All compounds are indicated with the same encoding used in the reactions dataset. Information extracted from `SETTINGS_EcoCyc_3AA_from_oxaloacetate_3.yaml` file (`SupMat` folder).

3.3. Search without previous information of required compounds or initial substrate

One extreme situation arises when there is no knowledge in advance about what compound is the initial substrate or what substrates are required for particular reactions in the pathway. To simulate this situation, it will be indicated in the PhDSeeker configuration file that all compounds to be related (4 for this example) can be used as the initial substrate of the pathway. This is done by simply specifying `yes` in the `initial` label of each compound to relate. Furthermore, since there is no information of specific compounds to include into the list of abundant ones, a list with commonly used compounds is considered. Figure S11 shows the setup for this search. Using this configuration, PhDSeeker is run with the following command:

```
>> python phdseeker.py --settings config/SETTINGS_EcoCyc_3AA_from_oxaloacetate_3.yaml
```

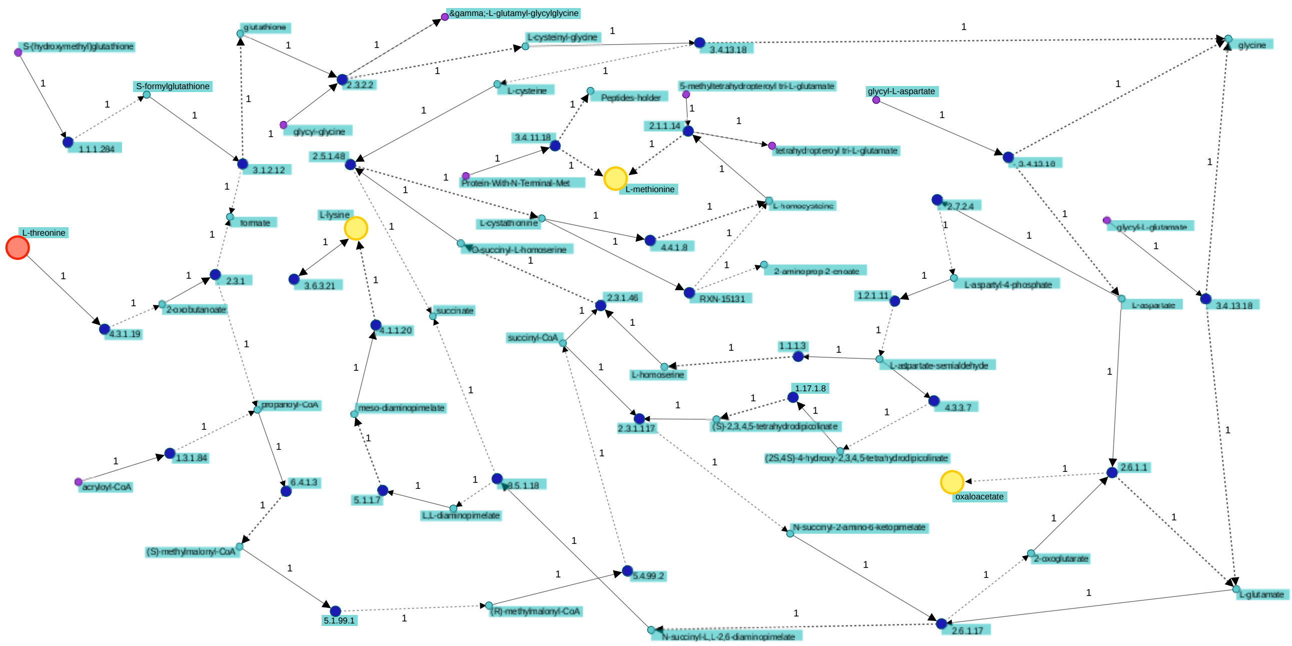


Figure S12: Pathway designed by PhDSeeker to relate L-methionine, L-threonine, L-lysine and oxalacetate. For better interpretation, freely available compounds are not shown in the figure.

As a result, PhDSeeker returns three files:

- `<yyyymmdd-hhmmss>__Escherichia coli K-12 substr. MG1655-21.5__SUMMARY.txt`
- `<yyyymmdd-hhmmss>__Escherichia coli K-12 substr. MG1655-21.5__PATHWAY.html`
- `<yyyymmdd-hhmmss>__Escherichia coli K-12 substr. MG1655-21.5__HISTORY.txt`

where `<yyyymmdd-hhmmss>` indicates `<year,month,day-hour,minute,second>` from when the execution was over. SUMMARY file contains a description of the parameters used in the search, some measures related to algorithm execution, compounds involved in the search, and the sequence of reactions that make up the pathway. PATHWAY file contains the interactive representation of the pathway found. HISTORY file contains a detail of the sequence of reactions for the different different pathways found all through out the search.

The tool required around 1 hour and 741 iterations to find the proposed solution provided in the SUMMARY file. Note that while PhDSeeker searches for the best pathway, this was improved 18-times, indicating that the process of building the solution was challenging. We must point out that while the solution presented here uses L-threonine as the initial substrate, this is only one of many solutions that PhDSeeker is able to find. In fact, most of the solutions found with this configuration use oxaloacetate as the initial substrate and only in some cases other compounds are chosen.

In reference to the pathway found, it is of note that PhDSeeker is able to link all the initially specified compouds using only the available genetic information about the host (*E. coli* K12). Despite this great feature, the considerations on the feasibility of the proposed solution is out of the scope of this work. It is interesting to highlight that this information (a pathway only using the available enzymes of the host) could be compared with a ”freely” run, i.e., using all available enzymes (from different hosts). This comparison could provides addtitional data about unwished interferences of the endogenous enzymes respect to heterologous pathways which are trying to introduce in a selected host as part of a synthetic biology experiment.

sinc(*i*) Research Institute for Signals, Systems and Computational Intelligence (sinc.unl.edu.ar)
M. Gerard & R.N. Comelli; "PhDSeeker: Pheromone-Directed Seeker for metabolic pathways"
BioSystems, Vol. 198, 2020.

Table S5: Pathway designed by PhDSeeker to relate *L-methionine*, *L-threonine*, *L-lysine* and *oxalacetate* without specifying the initial one. Compounds are colored according to their function. Reactions are ordered from top to bottom according to the order they are carried out.

RXN-2962 (1.1.1.284): 1 **$NADP^+$** + 1 **S-(hydroxymethyl)glutathione** → 1 **S-formylglutathione** + 1 **NADPH** + 1 **H^+**

S-FORMYLGLUTATHIONE-HYDROLASE-RXN (3.1.2.12): 1 **S-formylglutathione** + 1 **H_2O** → 1 **formate** + 1 **glutathione** + 1 **H^+**

RXN-18092 (2.3.2.2): 1 **glutathione** + 1 **glycyl-glycine** → 1 **&gamma;-L-glutamyl-glycylglycine** + 1 **L-cysteinyl-glycine**

RXN-6622 (3.4.13.18): 1 **L-cysteinyl-glycine** + 1 **H_2O** → 1 **L-cysteine** + 1 **glycine**

THREDEHYD-RXN (4.3.1.19): 1 **L-threonine** → 1 **2-oxobutanoate** + 1 **ammonium**

KETOBUTFORMLY-RXN (2.3.1): 1 **2-oxobutanoate** + 1 **coenzyme A** → 1 **formate** + 1 **propanoyl-CoA**

RXN-9087 (1.3.1.84): 1 **acryloyl-CoA** + 1 **NADPH** + 1 **H^+** → 1 **$NADP^+$** + 1 **propanoyl-CoA**

PROPIONYL-COA-CARBOXY-RXN (6.4.1.3): 1 **ATP** + 1 **hydrogencarbonate** + 1 **propanoyl-CoA** → 1 **ADP** + 1 **(S)-methylmalonyl-CoA** + 1 **H^+** + 1 **phosphate**

METHYLMALONYL-COA-EPIM-RXN (5.1.99.1): 1 **(S)-methylmalonyl-CoA** → 1 **(R)-methylmalonyl-CoA**

METHYLMALONYL-COA-MUT-RXN (5.4.99.2): 1 **(R)-methylmalonyl-CoA** → 1 **succinyl-CoA**

RXN0-6984 (3.4.13.18): 1 **glycyl-L-glutamate** + 1 **H_2O** → 1 **L-glutamate** + 1 **glycine**

3.4.11.18-RXN (3.4.11.18): 1 **Protein-With-N-Terminal-Met** + 1 **H_2O** → 1 **L-methionine** + 1 **H^+** + 1 **Peptides-holder**

RXN0-6987 (3.4.13.18): 1 **glycyl-L-aspartate** + 1 **H_2O** → 1 **glycine** + 1 **L-aspartate**

ASPARTATEKIN-RXN (2.7.2.4): 1 **ATP** + 1 **L-aspartate** → 1 **ADP** + 1 **L-aspartyl-4-phosphate**

ASPARTATE-SEMIALDEHYDE-DEHYDROGENASE-RXN (1.2.1.11): 1 **L-aspartyl-4-phosphate** + 1 **NADPH** + 1 **H^+** → 1 **L-aspartate-semialdehyde** + 1 **$NADP^+$** + 1 **phosphate**

DIHYDRODIPICSYN-RXN (4.3.3.7): 1 **L-aspartate-semialdehyde** + 1 **pyruvate** → 1 **(2S,4S)-4-hydroxy-2,3,4,5-tetrahydrodipicolinate** + 1 **H^+** + 1 **H_2O**

RXN-14014 (1.17.1.8): 1 **(2S,4S)-4-hydroxy-2,3,4,5-tetrahydrodipicolinate** + 1 **NADH** + 1 **H^+** → 1 **(S)-2,3,4,5-tetrahydrodipicolinate** + 1 **NAD^+** + 1 **H_2O**

TETHYDPICSUCC-RXN (2.3.1.117): 1 **(S)-2,3,4,5-tetrahydrodipicolinate** + 1 **succinyl-CoA** + 1 **H_2O** → 1 **coenzyme A** + 1 **N-succinyl-2-amino-6-ketopimelate**

SUCCINYLDIAMINOPIMTRANS-RXN (2.6.1.17): 1 **L-glutamate** + 1 **N-succinyl-2-amino-6-ketopimelate** → 1 **2-oxoglutarate** + 1 **N-succinyl-L,L-2,6-diaminopimelate**

SUCCDIAMINOPIMDESUCC-RXN (3.5.1.18): 1 **N-succinyl-L,L-2,6-diaminopimelate** + 1 **H_2O** → 1 **L,L-diaminopimelate** + 1 **succinate**

DIAMINOPIMEPIM-RXN (5.1.1.7): 1 **L,L-diaminopimelate** → 1 **meso-diaminopimelate**

DIAMINOPIMDECARB-RXN (4.1.1.20): 1 **meso-diaminopimelate** + 1 **H^+** → 1 **CO_2** + 1 **L-lysine**

ABC-3-RXN (3.6.3.21): 1 **ATP** + 1 **L-lysine** + 1 **H_2O** → 1 **ADP** + 1 **L-lysine** + 1 **H^+** + 1 **phosphate**

HOMOSERDEHYDROG-RXN (1.1.1.3): 1 **L-aspartate-semialdehyde** + 1 **NADPH** + 1 **H^+** → 1 **L-homoserine** + 1 **$NADP^+$**

HOMSUCTRAN-RXN (2.3.1.46): 1 **L-homoserine** + 1 **succinyl-CoA** → 1 **coenzyme A** + 1 **O-succinyl-L-homoserine**

O-SUCCHOMOSERLYASE-RXN (2.5.1.48): 1 **L-cysteine** + 1 **O-succinyl-L-homoserine** → 1 **L-cystathionine** + 1 **H^+** + 1 **succinate**

CYSTATHIONINE-BETA-LYASE-RXN (4.4.1.8): 1 **L-cystathionine** + 1 **H_2O** → 1 **ammonium** + 1 **L-homocysteine** + 1 **pyruvate**

RXN-15131 (-.-.-.-): 1 **L-cystathionine** → 1 **2-aminoprop-2-enoate** + 1 **L-homocysteine**

HOMOCYSMET-RXN (2.1.1.14): 1 **5-methyltetrahydropteroyl tri-L-glutamate** + 1 **L-homocysteine** → 1 **tetrahydropteroyl tri-L-glutamate** + 1 **L-methionine**

ASPAMINOTRANS-RXN (2.6.1.1): 1 **2-oxoglutarate** + 1 **L-aspartate** → 1 **L-glutamate** + 1 **oxaloacetate**

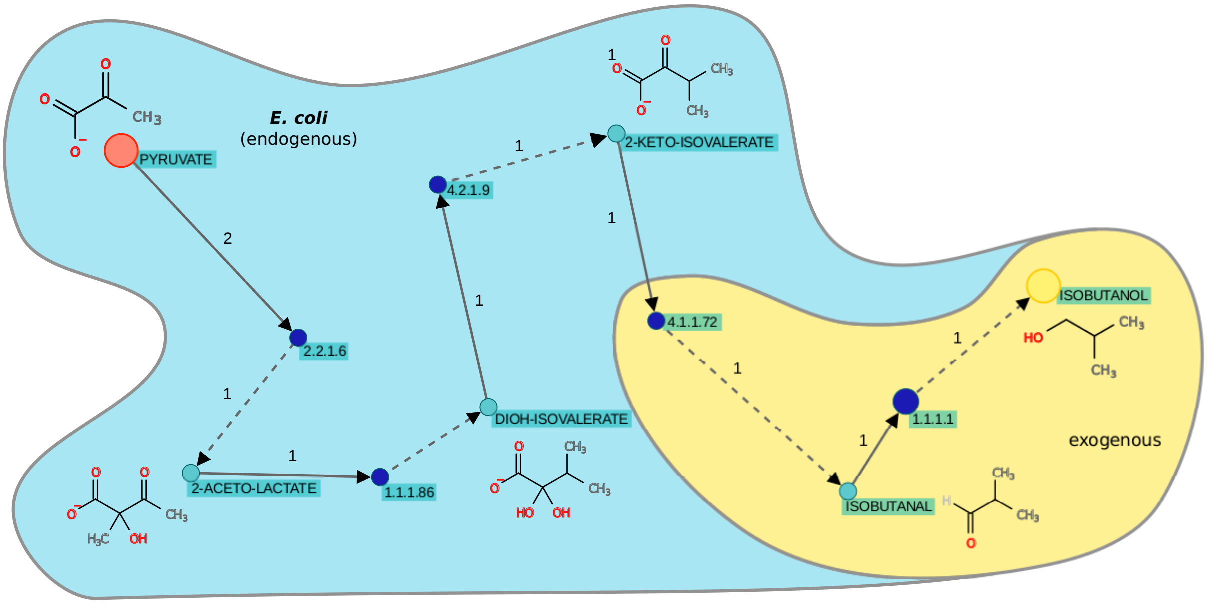


Figure S13: Pathway proposed by Atsumi *et al.* to produce isobutanol from pyruvate. Figure reconstructed from data available in [3]. Reactions and molecular structures were extracted from MetaCyc.

4. Designing metabolic pathways for industrial applications

This section analyzes the use of PhDSeeker for metabolic pathways design. For this purpose, two pathways successfully designed and tested in *E. coli* are taken as examples. As it could be expected, the wild-type organism is not able to produce those compounds. This makes it necessary to find additional reactions to those that comprise its metabolism to provide it with the new capabilities required. Taking this into account, the MetaCyc v21.5 database is used in the experiments below. Therefore, the dataset is built with the command-line tools provided with PhDSeeker from the `compounds.dat` and `reactions.dat` files of the MetaCyc v21.5 database by running the following command:

```
>> python BioCyc2PHDSFiles.py --cpdDB compounds.dat --rxnDB reactions.dat --allreversible False
```

This dataset contains a total of 15882 reactions from 2960 organisms. From these, only 2818 reactions are reversible. As in all cases, information of the built dataset is stored in three separated files:

- `REACTIONS_MetaCyc-21.5_<yyyymmdd>.txt` (reactions that can be used to build a pathway)
- `ENZYMES_MetaCyc-21.5_<yyyymmdd>.txt` (enzymes catalyzing each reaction)
- `NAMES_MetaCyc-21.5_<yyyymmdd>.txt` (compound names for each compound code)

where <yyyymmdd> indicates <year,month,day> when the dataset was built. The configuration file to be used in this search is setup according to Figure S14. The list of freely available compounds is the same used in Section 3.2..

4.1. Synthesis of isobutanol from pyruvate

Higher alcohols offer advantages over ethanol as gasoline substitutes because of their higher energy density and lower hygroscopicity. For this reason, compounds such as isobutanol are important in biofuel production. However, chemical synthesis of these compounds is often difficult and expensive. Thus, the first example corresponds to a pathway designed by Atsumi and co-workers to produce isobutanol from pyruvate in *E. coli* [4]. Their design is build up of three endogenous reactions belonging to *E. coli* and two exogenous ones [3]. Figure S13 shows the proposed pathway, highlighting the exogenous and endogenous reactions involved in the synthesis.

```
---
  NCORES: 6
  Nants: 10
  rho: 0.1
  maxIterations: 1000
  IterationsWithoutChanges: 20
  IterationsWithAlignedAnts: 10
  Strict_Initialization: False
  AllowExternalCompounds: True
  Verbose: True

  #----------------------------------------------------------------------------

  REACTIONS: db/BioCyc/REACTIONS_MetaCyc-21.5_20181213.txt
  ENZYMES: db/BioCyc/ENZYMES_MetaCyc-21.5_20181213.txt
  COMPNAMES: db/BioCyc/NAMES_MetaCyc-21.5_20181213.txt

  #----------------------------------------------------------------------------

  COMPOUNDS:

    abundant: ['ACET', 'ADP', 'AMMONIUM', 'ATP', 'CARBON-DIOXIDE', 'CO-A', 'FAD',
               'FADH2', 'HCO3', 'NAD', 'NADH', 'NADP', 'NADPH', 'NITRATE',
               'OXYGEN-MOLECULE', 'PROTON', 'Pi', 'WATER']

    relate:
      -
        compound: PYRUVATE # Pyruvate
        initial: yes
      -
        compound: ISOBUTANOL # Isobutanol
        initial: no
```

Figure S14: Experimental setup for searching pathways to synthesize isobutanol from pyruvate. Information extracted from `SETTINGS_MetaCyc-21.5_pyruvate_isobutanol.yaml` file (`SupMat` folder).

For this task, PhDSeeker was run using the parameters in Figure S14 with the following command:

```
>> python phdseeker.py --settings config/SETTINGS_MetaCyc-21.5_pyruvate_isobutanol.yaml
```

As a result, PhDSeeker returns three files:

- `<yyyymmdd-hhmmss>__MetaCyc-21.5__SUMMARY.txt`
- `<yyyymmdd-hhmmss>__MetaCyc-21.5__PATHWAY.html`
- `<yyyymmdd-hhmmss>__MetaCyc-21.5__HISTORY.txt`

where `<yyyymmdd-hhmmss>` indicates `<year,month,day-hour,minute,second>` from when the execution was over. SUMMARY file contains a description of the parameters used in the search, some measures related to algorithm execution, compounds involved in the search, and the sequence of reactions that make up the pathway. PATHWAY file contains the interactive representation of the pathway found. HISTORY file contains a detail of the sequence of reactions for the different different pathways found all through out the search.

The first thing observed when analyzing the SUMMARY file is that the search took about 44 hours (1d:20h:17m:9s), using a total of 741 iterations. While it may seem like a considerable time for searching, there are several factors to consider. The search space for this problem is really extensive, having more than 15000 reactions available to build the solution. In addition, as it is shown in the HISTORY file, the tool returns the first solution (containing 13 reactions) in about 1 hour. Then, it is improved 6-times until the final pathway proposed by PhDSeeker is obtained. Note that although

Table S6: Pathway proposed by PhDSeeker to produce isobutanol from pyruvate.

ACETOLACTSYN-RXN (2.2.1.6): 1 **H$^+$** + 2 **pyruvate** → 1 **(S)-2-acetolactate** + 1 **CO$_2$**
RXN-16061 (1.1.1.383): 1 **(S)-2-acetolactate** + 1 **NADH** + 1 H$^+$ → 1 **(2R)-2,3-dihydroxy-3-methylbutanoate** + 1 **NAD$^+$**
DIHYDROXYISOVALDEHYDRAT-RXN (4.2.1.9): 1 **(2R)-2,3-dihydroxy-3-methylbutanoate** → 1 **3-methyl-2-oxobutanoate** + 1 **H$_2$O**
RXN-7643 (4.1.1.72): 1 **3-methyl-2-oxobutanoate** + 1 **H$^+$** → 1 **CO$_2$** + 1 **isobutanal**
RXN-7657 (1.1.1.1): 1 **isobutanal** + 1 **NADH** + 1 **H$^+$** → 1 **isobutanol** + 1 **NAD$^+$**

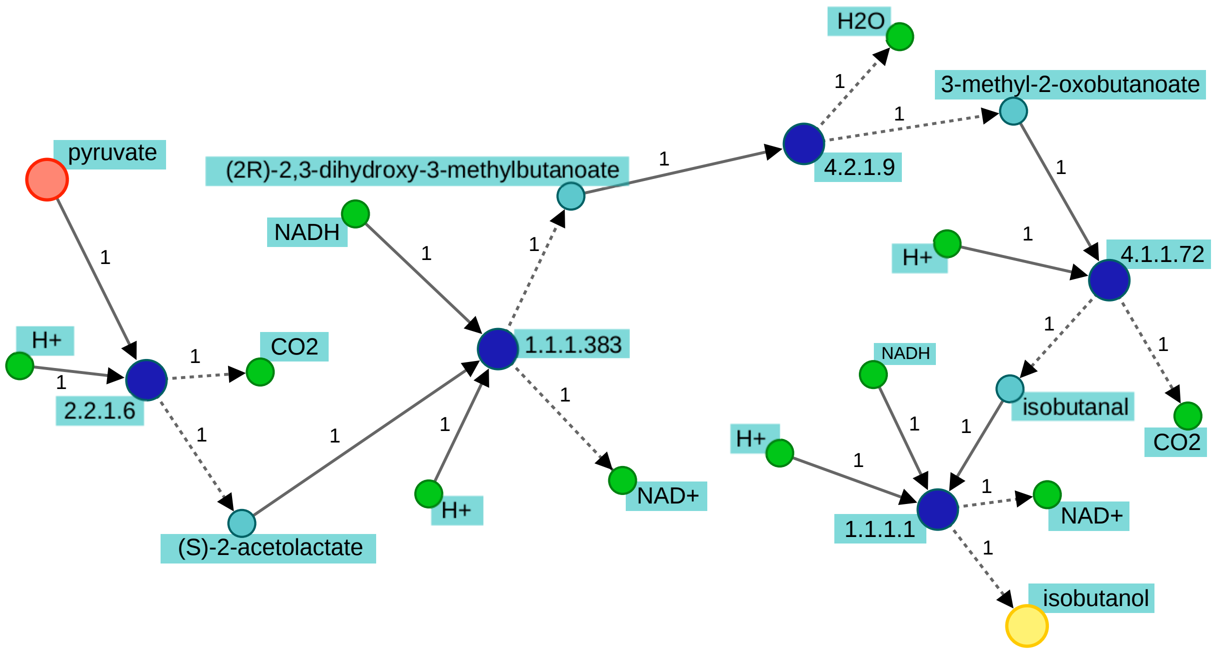


Figure S15: Pathway proposed by PhDSeeker to produce isobutanol from pyruvate.

the algorithm requires almost 2 days to return the solution, the final pathway is proposed by the tool after 18.5 hours of execution and remains unchanged until the stop criterion is reached (20 iterations without change in the best solution, as defined in the Figure S14). Finally, it is important to remark that the search is automatically performed only requiring the specification of the initial conditions.

As shown in Figure S6 and Figure S15, the metabolic pathway designed by PhDSeeker is build up by 5 reactions (of those, only the second reaction (RXN-16061) is catalyzed by a different enzyme than the original design of Atsumi *et al.*). Interestingly, enzymes 1.1.1.86 (Atsumi's pathway) and 1.1.1.383 (PhDSeeker's pathway) share substrates and products according to BRENDA database, thus catalyzing the same reaction.

The ketol-acid reductoisomerase (EC 1.1.1.86) using in the Atsumi's pathway is coded in the *E. coli* genome and their activity is NADPH-dependent. In the other hand, the enzyme proposed by PhDSeeker (EC 1.1.1.383) is closely related to the first but it can use both NADH and NADPH as cofactors with similar efficiency. This enzyme was characterized from specific microorganism different from *E. coli*. Because NADP is synthesized from NAD at expenses of ATP consumption, an enzyme NAD-dependent could be more efficiently in energy terms for the cell metabolism. However, the final choice depends upon the researcher's criteria as mentioned along this job. Redox balance, competition for intermediated compounds, cofactors usage and affinity by them and/or preference by endogenous enzymes over the heterologous pathways are key factor that should be defined by the PhDSeeker user.

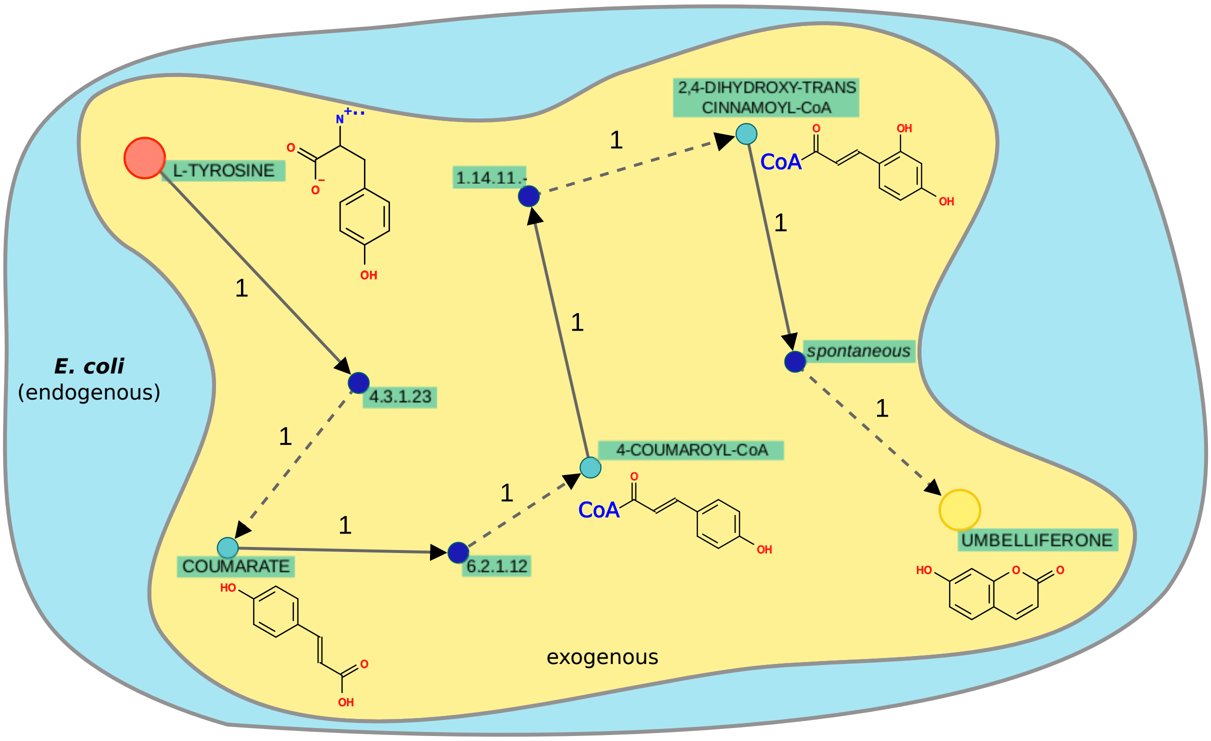


Figure S16: Pathway engineered by Santos *et al.* to produce umbelliferone from L-tyrosine. Figure reconstructed from data available in [3]. Reactions and molecular structures were extracted from MetaCyc.

4.2. Synthesis of umbelliferone from L-tyrosine

Umbelliferone is another compound with important industrial applications. This is commonly used as a component in solar-filters and to provide brightness in textile products. In order to produce this compound in a simple and less expensive way, Santos *et al.* [5] designed a metabolic pathway able to produce umbelliferone from L-tyrosine in *E. coli*. Their design involves four reactions, three that are exogenous reactions and a spontaneous one. These must be introduced in the microorganism for the production of umbelliferone. Figure S16 shows the pathway proposed by Santos and co-workers. Taking this pathway as reference, PhDSeeker was configured to design a pathway that is also capable of producing umbelliferone from L-tyrosine. Thus, the configuration file was setup according to Figure S17. Freely available compounds were specified as in previous sections.

In order to search for a pathway to synthesize umbelliferone from L-tyrosine, PhDSeeker was configured according to parameters in Figure S14, and run with the following command:

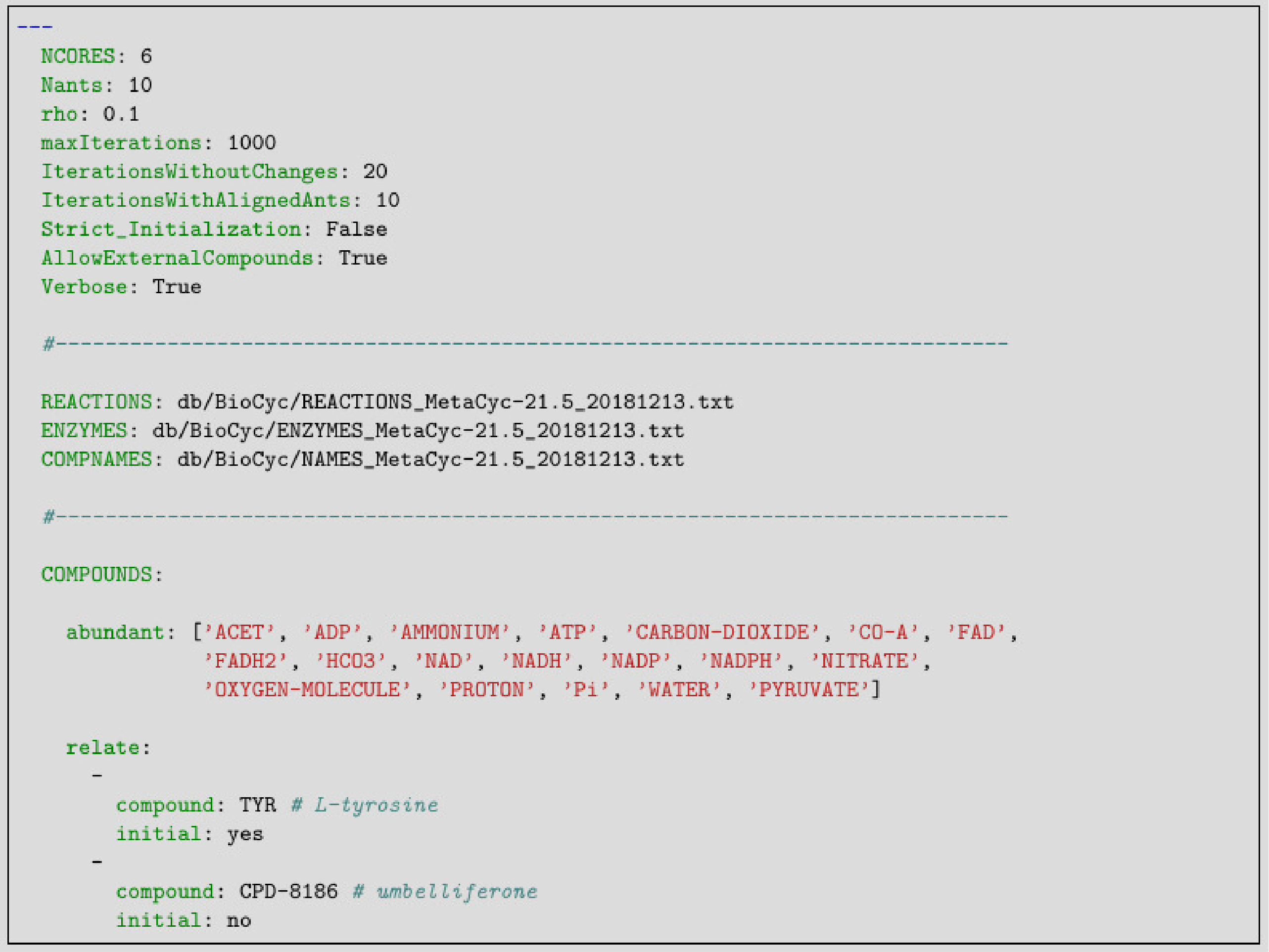
```
>> python phdseeker.py --settings config/SETTINGS_MetaCyc-21.5_L-tyrosine_umbelliferone1.yaml
```

As usually, PhDSeeker returns three files:

- <yyyymmdd-hhmmss>__MetaCyc-21.5__SUMMARY.txt
- <yyyymmdd-hhmmss>__MetaCyc-21.5__PATHWAY.html
- <yyyymmdd-hhmmss>__MetaCyc-21.5__HISTORY.txt

where <yyyymmdd-hhmmss> indicates <year,month,day-hour,minute,second> from when the execution was over. SUMMARY file contains a description of the parameters used in the search, some measures related to algorithm execution, compounds involved in the search, and the sequence of reactions that make up the pathway. PATHWAY file contains the interactive representation of the pathway found. HISTORY file contains a detail of the sequence of reactions for the different different pathways found all through out the search.

M. Gerard & R.N. Comelli; "PhDSeeker: Pheromone-Directed Seeker for metabolic pathways"



```
---
  NCORES: 6
  Nants: 10
  rho: 0.1
  maxIterations: 1000
  IterationsWithoutChanges: 20
  IterationsWithAlignedAnts: 10
  Strict_Initialization: False
  AllowExternalCompounds: True
  Verbose: True

  #----------------------------------------------------------------------------

  REACTIONS: db/BioCyc/REACTIONS_MetaCyc-21.5_20181213.txt
  ENZYMES: db/BioCyc/ENZYMES_MetaCyc-21.5_20181213.txt
  COMPNAMES: db/BioCyc/NAMES_MetaCyc-21.5_20181213.txt

  #----------------------------------------------------------------------------

  COMPOUNDS:

    abundant: [’ACET’, ’ADP’, ’AMMONIUM’, ’ATP’, ’CARBON-DIOXIDE’, ’CO-A’, ’FAD’,
               ’FADH2’, ’HCO3’, ’NAD’, ’NADH’, ’NADP’, ’NADPH’, ’NITRATE’,
               ’OXYGEN-MOLECULE’, ’PROTON’, ’Pi’, ’WATER’, ’PYRUVATE’]

    relate:
      -
        compound: TYR # L-tyrosine
        initial: yes
      -
        compound: CPD-8186 # umbelliferone
        initial: no
```

Figure S17: Experimental setup for searching pathways to synthesize umbelliferone from L-tyrosine. Information extracted from `SETTINGS_MetaCyc-21.5_L-tyrosine_umbelliferone.yaml` file (`SupMat` folder).

As expected for this kind of problems, the search takes about 2 days to complete 34 iterations of the algorithm to return a solution (see file SUMMARY). As in the previous case study, although more than 15000 reactions are available to build the solution, the first pathway is proposed after approximately 1 hour of execution. Although this solution is composed of 100 reactions, it is quickly improved to reach the final solution after 2 hours of execution (14 iterations) and 4 additional metabolic pathways (see HISTORY file for details).

The pathway designed by PhDSeeker is shown in Figure S7 and Figure S18. Remarkably, the solution proposed by the tool only shares the first reaction with the reference pathway. Moreover, the remaining reactions are part of the *superpathway of scopolin and esculin biosynthesis*[7], a pathway typically found in *Manihot esculenta* and *Nicotiana tabacum*. Another interesting aspect of the proposed solution is that CoA is not required. Despite there is not information about enzymes catalyzing those reactions, intermediate compounds are different. Although more analisys to evalauate feasibility of this solution is required, this could be an interesting alternative to the one presented by Santos and co-workers.

This solution contains the same number of reactions as the reference pathway and appears to be more specific (there are no additional intermediate compounds generated/consumed in the synthesis). Consequently, it is highly likely that a new search will lead to the same solution. In order to ensure this solution does not be found in a new search, reactions `RXN14166`, `RXN14167` and `RXN14168` were manually removed from the dataset. This new dataset, named `REACTIONS_MetaCyc-21.5_<yyyymmdd>_v2.yaml`, is then used as search space.

[7] https://biocyc.org/META/new-image?type=PATHWAY&object=PWY-7186&detail-level=2

Table S7: Pathway proposed by PhDSeeker to produce umbelliferone from L-tyrosine. Enzymes that are part of the solution proposed by Santos and co-workers are indicated in bold.

RXN-9697 (4.3.1.23 +1): 1 **L-tyrosine** → 1 **NH_4^+** + 1 **4-coumarate**
RXN-14167 (-.-.-.-): 1 **4-coumarate** → 1 **(E)-2,4-dihydroxycinnamate**
RXN-14166 (-.-.-.-): 1 **(E)-2,4-dihydroxycinnamate** → 1 **(Z)-2,4-dihydroxycinnamate**
RXN-14168 (-.-.-.-): 1 **(Z)-2,4-dihydroxycinnamate** → 1 **umbelliferone**

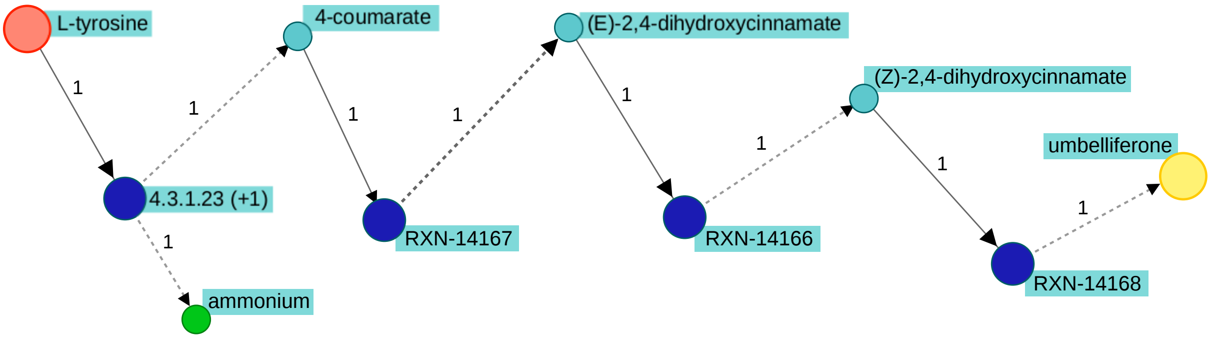


Figure S18: Pathway proposed by PhDSeeker to produce umbelliferone from L-tyrosine.

PhDSeeker is run with the same parameters in Figure S14 (only changing the REACTIONS dataset) and the following command:

```
>> python phdseeker.py --settings config/SETTINGS_MetaCyc-21.5_L-tyrosine_umbelliferone2.yaml
```

As in all cases, PhDSeeker returns three files showing the best solution as plain text and as an interactive representation, and the search history. The tool required around 1 day and 18 hours to perform 33 iterations and converge for a solution. Along the search, the best solution was improved 2-times (see HISTORY file for more details). The pathway designed is shown in Figure S8 and Figure S19. Interestingly, this pathway is composed by 2 independent pathways connected through 2-oxoglutarate. This compound is required for the enzyme 1.14.11.- to synthesize of umbelliferon but it is not freely available according to initial settings. However, PhDSeeker found a way to synthesize this compound from the available ones.

In other words, the PhDSeeker propose the same pathway as Santos *et al.* and also a second route to generate the 2-oxogutarate. This compound is necessary to the reaction mediated by the enzyme EC 1.14.11. The 2-oxoglutarate is an intermediary of the tricarboxylic acids cycle, a widely distributed pathway in the nature. So, it could be considered a freely available compound when culture media contains easily metabolized carbon sources (eg, glucose). Because this restriction was not specified at the beginning, the PhDSeeker proposed a route to obtain it and allow the synthesis of umbelliferone to be feasible.

Table S8: Pathway proposed by PhDSeeker to produce umbelliferone from L-tyrosine. Enzymes that are part of the solution proposed by Santos and co-workers are indicated in bold.

RXN-9697 (4.3.1.23 +1): 1 L-tyrosine → 1 NH_4^+ + 1 4-coumarate

4-COUMARATE–COA-LIGASE-RXN (6.2.1.12): 1 ATP + 1 CoA + 1 4-coumarate → 1 AMP + 1 4-coumaroyl-CoA + 1 PPi

3.4.13.7-RXN (3.4.13.7): 1 L-glutamyl-L-glutamate + 1 H_2O → 2 L-glutamate

RXN0-6978 (3.4.13.18): 1 L-alanyl-L-histidine + 1 H_2O → 1 L-histidine + 1 L-alanine

RXN-7571 (2.6.1.58): 1 L-histidine + 1 pyruvate → 1 imidazole-pyruvate + 1 L-alanine

HISTTRANSAM-RXN (2.6.1.38): 1 L-glutamate + 1 imidazole-pyruvate → 1 2-oxoglutarate + 1 L-histidine

GLUTAMATE-DEHYDROGENASE-RXN (1.4.1.2): 1 L-glutamate + 1 NAD^+ + 1 H_2O → 1 2-oxoglutarate + 1 NH_4^+ + 1 NADH + 1 H^+

ISOCITDEH-RXN (1.1.1.42): 1 2-oxoglutarate + 1 CO_2 + 1 NADPH → 1 $NADP^+$ + 1 D-threo-isocitrate

ISOCITRATE-DEHYDROGENASE-NAD+-RXN (1.1.1.286 +1): 1 NAD^+ + 1 D-threo-isocitrate → 1 2-oxoglutarate + 1 CO_2 + 1 NADH

RXN-12963 (1.14.11): 1 2-oxoglutarate + 1 O_2 + 1 4-coumaroyl-CoA → 1 CO_2 + 1 2,4-dihydroxycinnamoyl-CoA + 1 succinate

RXN-12965 (-.-.-.-): 1 2,4-dihydroxycinnamoyl-CoA → 1 CoA + 1 umbelliferone

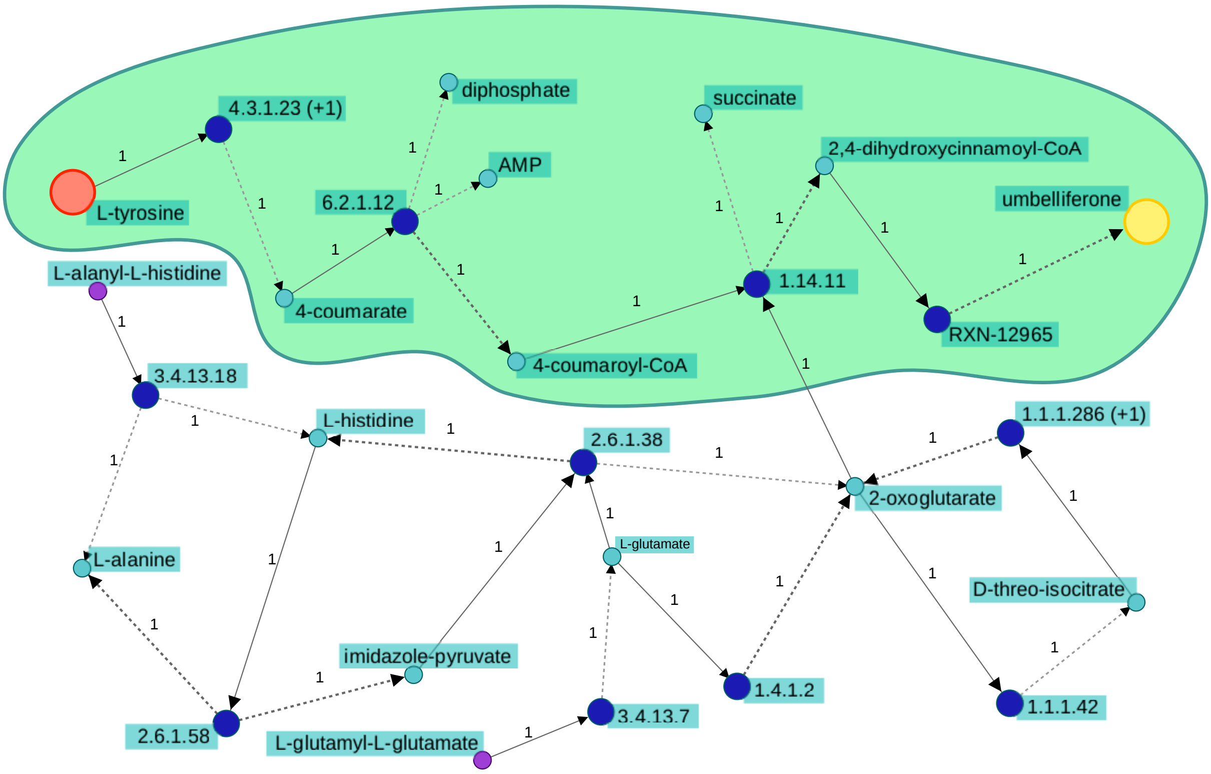


Figure S19: Pathway proposed by PhDSeeker to produce umbelliferone from L-tyrosine.

References

[1] Matias F. Gerard, Georgina Stegmayer, and Diego H. Milone. Metabolic pathways synthesis based on ant colony optimization. *Scientific Reports*, 8:16398, 2018.

[2] Tyrrell Conway. The Entner-Doudoroff pathway: history, physiology and molecular biology. *FEMS Microbiol Rev*, 9(1):1–27, 1992.

[3] Mario Latendresse, Markus Krummenacker, and Peter D. Karp. Optimal metabolic route search based on atom mappings. *Bioinformatics*, 30(14):2043–2050, 2014.

[4] Shota Atsumi, Tung-Yun Wu, Eva-Maria Eckl, Sarah D. Hawkins, Thomas Buelter, and James C. Liao. Engineering the isobutanol biosynthetic pathway in *Escherichia coli* by comparison of three aldehyde reductase/alcohol dehydrogenase genes. *Applied Microbiology and Biotechnology*, 85:651–657, 2010.

[5] Christine Nicole S. Santos, Mattheos Koffas, and Gregory Stephanopoulos. Optimization of a heterologous pathway for the production of flavonoids from glucose. *Metabolic Engineering*, 13(4):392–400, 2011.