

**Universität Bielefeld**

Technische Fakultät  
Center for Biotechnology (CeBiTec)

# Comparing Organisms on the Level of Metabolism

Zur Erlangung des akademischen Grades eines Doktors der  
Naturwissenschaften der Universität Bielefeld vorgelegte  
Dissertation

von

Sebastian Oehm

21/07/2009





---

# Contents

---

<b>1. Introduction</b>	1
1.1. Motivation . . . . .	1
1.2. Goal . . . . .	2
1.3. Structure of this Thesis . . . . .	3
<b>2. Background</b>	5
2.1. Metabolism and Metabolic Pathways . . . . .	5
2.2. Data Sources . . . . .	7
2.2.1. BIND . . . . .	7
2.2.2. BRENDA . . . . .	8
2.2.3. EcoCyc, MetaCyc, and BioCyc . . . . .	8
2.2.4. EMP/MPW . . . . .	9
2.2.5. ERGO . . . . .	9
2.2.6. ExPASy . . . . .	10
2.2.7. KEGG . . . . .	10
2.2.8. PUMA/WIT . . . . .	10
2.3. Formal Models for Metabolic Networks . . . . .	11
2.3.1. Set Model . . . . .	11
2.3.2. Reaction Graph . . . . .	12
2.3.3. Metabolite Graph . . . . .	12
2.3.4. Bipartite Graph . . . . .	13
2.4. Concepts for Metabolic Network Comparison . . . . .	14
2.4.1. Maximum Common Subgraph-based Approaches . . . . .	14
2.4.2. Feature-based Approaches . . . . .	14
2.4.3. Edit Operation-based Approaches . . . . .	15
2.5. Clustering Methods . . . . .	15
2.6. Related Work . . . . .	19

---

<b>3. Methodology</b>	23
3.1. Assessing the Difference between Metabolic Networks . . . . .	23
3.1.1. Reactions and Metabolites . . . . .	24
3.1.2. Reaction Neighborhood . . . . .	25
3.1.3. Network Function . . . . .	26
3.2. Data Source – Decision . . . . .	26
3.3. Data Model – Decision . . . . .	27
3.4. Concept for Metabolic Network Comparison – Decision . . . . .	28
3.5. Clustering Methods – Decision . . . . .	28
<b>4. Distance Measures</b>	31
4.1. Graph Theory . . . . .	31
4.1.1. Graphs and Subgraphs . . . . .	31
4.1.2. Isomorphisms on Graphs . . . . .	34
4.1.3. Graph Edit Distance . . . . .	35
4.2. Distance Measures on Metabolic Networks . . . . .	39
4.2.1. Cost Function Requirements . . . . .	39
4.2.2. Edit Distances on Reactions and Metabolites . . . . .	40
4.2.3. Neighborhood Sensitive Reaction Edit Distances . . . . .	48
<b>5. Implementation: the CPA Web Server</b>	51
5.1. Clustering Metabolic Pathway Data . . . . .	51
5.2. Results Overview . . . . .	53
5.3. Single Pathway Clustering Results . . . . .	55
5.4. Displaying Differential Reaction Content . . . . .	57
5.5. Simultaneously Displaying Reaction Content of Several Organisms . . . . .	58
<b>6. Results</b>	59
6.1. Comparison of Different Distance Measures and Clustering Techniques . . . . .	59
6.1.1. Artificial Test Scenario . . . . .	60
6.1.2. Lysine Subpathway Test Scenario . . . . .	67
6.1.3. Choice of Distance Measure and Clustering Technique . . . . .	74
6.2. Comparative Metabolic Pathway Analysis of Five Corynebacteria . . . . .	75
6.2.1. Classification Results . . . . .	76
6.2.2. Biological Interpretation . . . . .	80
<b>7. Conclusion</b>	87
7.1. Summary . . . . .	87
7.2. Discussion . . . . .	88
7.3. Outlook . . . . .	91
<b>A. Clustering Dendrograms</b>	95
<b>Bibliography</b>	111

---

## List of Figures

---

2.1.	Set model . . . . .	11
2.2.	Reaction graph model and its ambiguities . . . . .	12
2.3.	Metabolite graph model and its ambiguities . . . . .	13
2.4.	Bipartite graph model . . . . .	13
4.1.	Directed and undirected graphs . . . . .	32
4.2.	Metabolic network modeled as bipartite directed node-labeled graph . . . . .	33
4.3.	Maximum common subgraph and minimum common supergraph . . . . .	36
5.1.	CPA web server: clustering start page . . . . .	52
5.2.	CPA web server: results overview page . . . . .	54
5.3.	CPA web server: detailed clustering results view . . . . .	56
5.4.	CPA web server: differential reaction content visualization . . . . .	57
5.5.	CPA web server: reaction content visualization . . . . .	58
6.1.	Artificial test pathway . . . . .	61
6.2.	Lysine biosynthesis subpathway . . . . .	67
A.1.	Clustering dendrograms of seven artificial organisms for artificial test pathway, distance measures $m1$ and $m2$ , based on average and complete linkage agglomerative and Ward clustering . . . . .	96
A.2.	Clustering dendrograms of seven artificial organisms for artificial test pathway, distance measures $m3$ and $m4$ , based on average and complete linkage agglomerative and Ward clustering . . . . .	97
A.3.	Clustering dendrograms of seven artificial organisms for artificial test pathway, distance measures $m5$ and $m6$ , based on average and complete linkage agglomerative and Ward clustering . . . . .	98
A.4.	Clustering dendrograms of seven artificial organisms for artificial test pathway, distance measures $m7$ and $m8$ , based on average and complete linkage agglomerative and Ward clustering . . . . .	99

A.5. Clustering dendrograms of seven artificial organisms for artificial test pathway, distance measures $m9$ and $m10$ , based on average and complete linkage agglomerative and Ward clustering . . . . .	100
A.6. Clustering dendrograms of seven artificial organisms for artificial test pathway, distance measures $m11$ and $m12$ , based on average and complete linkage agglomerative and Ward clustering . . . . .	101
A.7. Clustering dendrograms of seven artificial organisms for artificial test pathway, distance measures $m2$ and $m5$ , based on average and complete linkage agglomerative and Ward clustering . . . . .	102
A.8. Clustering dendrograms of various organisms for lysine biosynthesis sub-pathway, distance measures $m1$ and $m2$ , based on average and complete linkage agglomerative and Ward clustering . . . . .	103
A.9. Clustering dendrograms of various organisms for lysine biosynthesis sub-pathway, distance measures $m3$ and $m4$ , based on average and complete linkage agglomerative and Ward clustering . . . . .	104
A.10. Clustering dendrograms of various organisms for lysine biosynthesis sub-pathway, distance measures $m5$ and $m6$ , based on average and complete linkage agglomerative and Ward clustering . . . . .	105
A.11. Clustering dendrograms of various organisms for lysine biosynthesis sub-pathway, distance measures $m7$ and $m8$ , based on average and complete linkage agglomerative and Ward clustering . . . . .	106
A.12. Clustering dendrograms of various organisms for lysine biosynthesis sub-pathway, distance measures $m9$ and $m10$ , based on average and complete linkage agglomerative and Ward clustering . . . . .	107
A.13. Clustering dendrograms of various organisms for lysine biosynthesis sub-pathway, distance measures $m11$ and $m12$ , based on average and complete linkage agglomerative and Ward clustering . . . . .	108
A.14. Clustering dendrogram of five <i>Corynebacteria</i> for overall metabolic network, distance measure $m1$ , based on average and complete linkage agglomerative and Ward clustering . . . . .	109

---

## List of Tables

---

6.1.	Distance measures being evaluated . . . . .	60
6.2.	Reaction content of pseudo-organisms for the artificial test pathway . . .	61
6.3.	Metabolite content of pseudo-organisms for the artificial test pathway . .	61
6.4.	Manually derived classifications of the artificial organisms for the artificial test pathway . . . . .	62
6.5.	Automatically derived classifications of the artificial organisms for the artificial test pathway . . . . .	63
6.6.	Costs of edit operations for reactions of the artificial test pathway . . . .	66
6.7.	Reaction content and manual classification of analyzed organisms for the lysine biosynthesis subpathway . . . . .	68
6.8.	Metabolite content of analyzed organisms for the lysine biosynthesis subpathway . . . . .	69
6.9.	Classification results for the lysine biosynthesis subpathway . . . . .	71
6.10.	Costs of edit operations for reactions of the lysine biosynthesis subpathway	75
6.11.	Automatically derived classification of five <i>Corynebacteria</i> for top five pathways, absolute sorting . . . . .	77
6.12.	Differential reaction content for top five pathways resulting from comparative pathway analysis of five <i>Corynebacteria</i> , absolute sorting . . . . .	77
6.13.	Differential reaction content for top five pathways resulting from comparative pathway analysis of five <i>Corynebacteria</i> , relative sorting . . . . .	78
6.14.	Top eight filtered pathways resulting from comparative pathway analysis of five <i>Corynebacteria</i> , absolute sorting . . . . .	79





## Acknowledgements

Undertaking and finishing a PhD project certainly requires great effort in many respects. I am very grateful that many people supported me along the way and certainly helped me a lot to accomplish this project.

Firstly, I thank Prof. Dr. Jens Stoye and Prof. Dr. Alfred Pühler for the opportunity to do my PhD under their supervision, and furthermore I thank Prof. Dr. Jens Stoye and PD Dr. Andreas Tauch for examining this thesis.

Special thanks go to Dr. Alexander Goesmann for supporting me with great patience throughout the entire project and especially in the finishing phase. I also thank Prof. Dr. Alfred Pühler, Prof. Dr. Jens Stoye, Dr. Alexander Goesmann, Dr. Jörn Kalinowski, and PD Dr. Andreas Tauch for fruitful discussions of the research project.

Many thanks go to my friends from Mainz, who greatly supported me by repeatedly discussing the topic of research and proofreading the manuscript. Thanks a lot also to friends from Australia and Bielefeld, and to the BRF people for proofreading and discussing the manuscript.

Insbesondere möchte ich meinen Eltern ganz herzlich danken für ihre uneingeschränkte Unterstützung während meines Studiums und meiner Promotion.

I acknowledge the support of the DFG Graduiertenkolleg Bioinformatik (GK635) and the NRW Graduate School in Bioinformatics and Genome Research and thank the associated people for their support. I also acknowledge financial support by the BMBF GenoMik-Plus project as well as the EU ERA-NET PathoGenoMics SPATELIS project.

Bielefeld, July 2009

Sebastian Oehm



### 1.1. Motivation

With the accumulation of gene and protein sequence data in publicly available databases and the development of computational methods for their comparison, sequence analysis has become an extremely powerful tool to uncover functional properties of these molecules (Ogata *et al.*, 2000). In general, however, the biological function is a result of many interacting molecules forming large interaction networks such as regulatory networks or metabolic pathways. With a growing amount of data being available not only on the function of single genes, but also on these interaction networks, it becomes feasible and valuable for further extending our knowledge about life and its working principles to perform comparisons also on the level of these networks. In particular the growing amount of publicly available data on metabolic pathways as well as of functional annotation data for sequenced organisms enables the comparison of organisms based on their metabolic reaction networks on a large scale.

For example in drug target identification, a functionally oriented comparison of organisms on the level of metabolic networks is a valuable complementation of the already established gene-based comparison. Gene-based comparison can be used to compile a list of all potential gene products produced by a particular organism, and the identification of genes that are common to all organisms in a chosen group of pathogens or unique to one particular pathogen (Galperin *et al.*, 1998). However, besides gene-wise comparison it is valuable to identify the cellular process potential drug targets are involved in and to perform a comparative pathway analysis for excluding possible side effects on the host. Sharma *et al.* (2008), for example, perform a manual metabolic pathway comparison to find those pathways that are present in the pathogen while missing in the host to ensure that potential drug targets have an effect on the pathogen, but not the host.

Comparing organisms based on their metabolic pathway variants can also be used for deducing phylogenetic relationships between organisms as has already been shown, for example, by Heymans and Singh (2003) and Forst and Schulten (2001). Moreover, it can

be applied for answering questions about lifestyle and habitat of organisms. If organisms are living in the same habitat or are following the same lifestyle (e.g. as intracellular pathogens), it is likely that they have evolved similar metabolic functionality, which might then be reflected in similar metabolic pathway variants of these organisms. If, conversely, organisms with unknown habitat or lifestyle are found to have similar pathway variants, this might indicate similarities in their habitat or lifestyle, independent of their phylogenetic relationship.

While for deriving phylogenetic trees it is appropriate to rely on the sequences of genes for assessing the similarity between metabolic pathways, this is not the case if the focus lies on comparing the function of the metabolic network for elucidating lifestyle and habitat related questions. This is because there exists a growing number of examples where on the one hand genes with similar sequence have different functions and on the other hand genes with identical function are not orthologous (Galperin *et al.*, 1998). Therefore, metabolic reactions should be the basis for this application rather than their corresponding genes.

Clearly, this sort of analysis is very sensitive to the quality of annotation, since metabolic pathway variants might appear to be similar due to missing or erroneous annotations although they actually are not similar. However, this opens yet another application area: if only very few enzymes are missing in a pathway variant of one organism in comparison to a taxonomically closely related one, this might be interpreted as indicator for missing annotations. Thus, interpreting genes in their metabolic context can assist in improving existing annotations, as has already been shown, for example, by Green and Karp (2004) and Ye *et al.* (2005).

Not much work has been published on comparing organisms based on their metabolism with the goal to analyze their shared or mutually missing reaction content and to group organisms according to similar pathway variants. Exceptions are the approaches by Ye *et al.* (2005) and Forst *et al.* (2006). The drawbacks of these approaches are that the former involves manual investigation of each pathway prior to comparative analysis, whereas the latter does not automatically group organisms according to their pathway variant. It appears that there is a need for bioinformatics tools supporting automated detection and classification of pathway variants in a set of organisms. This is especially true in light of the huge amount of data: several hundred genomes are already sequenced and annotated, and due to the growing number of ongoing sequencing projects the amount of available data is expanding ever faster and soon expected to exceed a thousand published genomes (<http://www.genomesonline.org/>).

## 1.2. Goal

The goal of this thesis is to provide a new comparative view on the metabolic capabilities of a set of organisms. Therefore, an approach performing a comparative metabolic network analysis resulting in a classification of organisms into groups of organisms sharing similar pathway variants is developed. The comparative analysis can be performed for the overall metabolic reaction networks of the organisms as well as for any choice of smaller metabolic pathways. An approach like this enables the discovery of differences in metabolism across a set of organisms that may help to develop new knowledge about metabolic peculiarities of the analyzed organisms, to detect metabolic functions neces-

sary for survival in a particular habitat, to find new candidate genes for drug design, and to reveal missing or erroneous annotations. This approach is based on metabolic reactions instead of on the respective genes and their sequences, and compares networks of reactions instead of individual reactions one by one.

Several steps need to be undertaken in order to implement this comparison strategy. Firstly, information needs to be gathered on which organisms are capable of catalyzing which metabolic reactions. This type of information can be taken from pathway-genome databases, which combine definitions of reaction equations and genome annotation data. Furthermore, reactions are to be grouped together into metabolic pathways if their metabolic function is involved in the same cellular process. Pathway definitions can either be taken from pathway-genome databases or be defined manually. Secondly, distance measures are to be developed to assess how similar the pathway variants of different organisms are to each other. Thirdly, clustering methods are needed for automatically finding groups of organisms with similar pathway variants, and finally the results are to be visualized for allowing easy access and quick interpretation. The envisaged approach should be made accessible to the research community and therefore a web server is to be developed.

Distance measures on metabolic networks may be based on different types of information. They may rely on the presence or absence of either reactions or metabolites or may take both into account. Structural information on the connections between the reactions and metabolites may be included in distance calculation as well. In this thesis different distance measures will be defined and their performance evaluated. Therefore, a theoretical framework is developed to define such distance measures, and proofs are given for certain properties of these distance measures. The theoretical framework in particular simplifies the definition of further distance measures with certain properties, and thus makes this approach very flexible with regard to future extensions.

## 1.3. Structure of this Thesis

The first chapter describes the motivation of the research undertaken in this thesis, defines the goal to be achieved and presents the structure of this thesis.

In the second chapter background information is provided. First, the concepts of metabolism and metabolic pathway are introduced. Then databases that come into consideration as data source in this thesis are reviewed. Subsequently data models that can be used to model metabolic networks are described. Then different concepts for developing distance measures are presented, followed by an introduction to methods for clustering and cluster validation. The chapter ends with a review of existing related approaches.

The third chapter starts with theoretical considerations on how to best assess distances between metabolic networks. Subsequently, decisions are made as to which distance measures are to be implemented, which database to use as data source, which model to use for modeling metabolic networks, which theoretical framework to use for developing the distance measures, and finally which clustering techniques to employ for classifying organisms according to similar pathway variants.

The fourth chapter is devoted to presenting the theory. It starts with an introduction to graph theory and edit distances on graphs. Following this, the distance measures are

formally defined.

In the following chapter the Comparative Pathway Analyzer (CPA) web server is presented. CPA is a free to use web implementation of the developed comparative approach. The functionality of this web server is demonstrated by means of an application example.

The next chapter documents the validation and application of the developed approach for comparative metabolic network analysis. First, two test scenarios are defined and all implemented distance measures and clustering techniques are evaluated for their suitability to compare metabolic networks. Then the approach is applied to a set of five *Corynebacteria* and the results are discussed in light of their biological relevance.

The last chapter concludes this thesis with an overall summary and discussion of the achieved results. Furthermore, possible improvements of the developed approach are suggested and further fields of application are outlined and discussed.

---

# Background

---

This chapter provides the background information for understanding the topic of research and for following the proposed methodology. Firstly, the concepts of metabolism and metabolic pathways are introduced, as well as the notion of functional annotation of organisms. These data are the basis of the proposed metabolic network comparison approach. The following section reviews the databases from which this data can be retrieved. Since the data has to be analyzed in the computer, models are needed for representing the relevant features of the data electronically. Therefore, different possible models applicable to metabolic network data are described in the next section. Then an overview is given of different concepts for developing distance measures on graphs, which are necessary to assess how similar two metabolic networks are to each other. Subsequently, a brief introduction to methods for cluster analysis is given. These methods are needed for automatically classifying the analyzed organisms according to the distances between their metabolic network variants. The chapter concludes with a review of existing approaches that are related to the topic of this thesis.

### 2.1. Metabolism and Metabolic Pathways

Metabolism, from greek *μετάβολος* (metabolos) for variable or shifting, is the biochemical modification of chemical compounds (metabolites) in living organisms and cells. This includes the biosynthesis of complex organic molecules (anabolism) and their breakdown (catabolism). A single step of such a biochemical modification is called a metabolic reaction. Such a reaction is characterized by the metabolites that are consumed and those that are produced, known as substrates and products, respectively, and the reaction stoichiometry, which describes how much of each substrate the reaction transforms into how much of each product. This information is given by the reaction equation. Since a product of one reaction acts as a substrate of some other reaction, taken together the reactions form a metabolic network. Although some reactions take place spontaneously,

the majority of them are catalyzed by specific proteins called enzymes. Enzymes in turn are synthesized in the cell on the basis of information coded as genes in the organism's genome. Enzymes can be classified according to the reaction they catalyze using the hierarchical enzyme classification scheme published by the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (NC-IUBMB) (1992). For each enzyme this scheme provides a code consisting of four numbers which categorize the catalyzed chemical reaction. For example, EC 1 encodes oxidoreductases, EC 1.1 encodes those oxidoreductases that act on the CH-OH group of donors, EC 1.1.1 those with NAD<sup>+</sup> or NADP<sup>+</sup> as acceptor and EC 1.1.1.1 those that act on an alcohol (i.e. alcohol dehydrogenases).

Formerly, the metabolism of an organism was investigated by wet-lab experiments. Since the advent of genome sequencing techniques the DNA sequences of genes can be determined. In a process called genome annotation these sequences are assigned information on their function in the cell. Metabolic genes for example are assigned information on the enzyme they code for as well as the function this enzyme has, i.e. the metabolic reaction it catalyzes. The development of bioinformatics methods for large scale comparison of genome sequences has enabled the automated prediction of genes and their function in the genome of newly sequenced organisms. This automatic annotation is based on the paradigm that genes with similar sequence are coding for the same function. Verification of these predictions still has to be undertaken in wet-lab experiments.

Enzymes may catalyze forward or both forward and backward direction of a reaction. The direction of reactions is important when it comes to deciding whether the function of a reaction or a chain of reactions is to degrade or synthesize a particular metabolite. A well-known example is glycolysis for degrading sugars and its reverse counter-part, gluconeogenesis, for synthesizing new sugar molecules. Most of the involved reactions are reversible and common to both pathways, despite some irreversible key reactions. In glycolysis, phosphoenolpyruvate carboxylase catalyzes the irreversible carboxylation of phosphoenolpyruvate (Kai *et al.*, 2003). If no gene is present in the organism's genome whose respective enzyme catalyzes a reverse reaction, as for example phosphoenolpyruvate carboxykinase (Holyoak *et al.*, 2006), gluconeogenesis cannot take place.

However, there is more information in metabolic networks than the mere set of constituting reactions: the structure of the network, or, in other words, the information on the functional dependency, or how the reactions are interconnected. A particular reaction can only take place if in addition to the catalyzing enzyme, all substrate metabolites are available. Other reactions might have to take place before this particular one in order to produce these substrate metabolites. This recursive phenomenon leads to the functional interdependency of reactions in a metabolic network: an organism might not be able to produce a certain metabolite if a reaction responsible for producing a precursor metabolite is missing and this metabolite cannot be obtained in any other way like, for example, via uptake from the environment.

The term metabolic pathway has traditionally been used to summarize a set of such functionally dependent reactions. Biochemical experimentation has led to the discovery of knowledge about reaction stoichiometries, and reactions sharing common intermediates were grouped together to form metabolic pathways. Well-known examples of these traditionally defined pathways include glycolysis or citric acid cycle. Naturally, this way of assembling metabolic pathways implies a certain amount of arbitrariness in selecting the reactions to be included. Nevertheless, they still represent a valid functional group-



ing of reactions, since in each case all reactions involved in a particular cellular task are grouped together. This sort of pathway can, for example, be found in the KEGG database (Kanehisa and Goto, 2000).

Besides these experimentally derived pathway definitions, some mathematical approaches have been developed that can be used to deduce subnetworks composed of functionally related reactions from the overall metabolic network of an organism. The benefit of these is that they uniquely define metabolic pathways directly from network topology, i.e. the structure of the metabolic network as characterized by its reactions, metabolites, and the reactions' stoichiometry. Some of the earlier examples are given by Seressiotis and Bailey (1988) and Mavrovouniotis *et al.* (1990). They developed methods for finding all possible reaction routes from some metabolite A to another metabolite B. A more recent approach of this category is the tool PathFinder, published by Goesmann *et al.* (2002). Other approaches involve an analysis of the stoichiometry of the reaction network under the steady state assumption. Examples are Petri Net analysis (Heiner and Koch, 2004), elementary flux mode (EFM) analysis (Schuster and Hilgetag, 1994), and extreme pathway (EP) analysis (Schilling *et al.*, 2000). However, the drawback of all these automated approaches is that the quality of their results strongly depends on the correctness and completeness of information on the overall metabolic reaction network. Results have to be verified manually and computations may have to be iterated numerous times until a satisfying result is achieved. This is a huge effort which has not yet been undertaken on a large scale.

As can be seen, different approaches are possible to define metabolic pathways. In particular, defining one's own metabolic pathways and thus generating a view on the metabolism of an organism that suits one's own research interests best is considered to be a perfectly valid approach.

## 2.2. Data Sources

Information on metabolism is traditionally published in journals and textbooks, but meanwhile several databases exist that store this data electronically. These databases differ in the type(s) of data stored (pathways, reactions, enzymes, regulatory interactions, genes, organisms, etc.), the source of this data (wet-lab experiments, in-silico predictions), the quality of the data (hand-curated or automatically generated), the total amount of available data (comprehensiveness: number of pathways and genomes/organisms), and the accessibility of the data (website or flatfile download).

In this section, databases that come into consideration as data source in this thesis for reaction data, pathway data, and genome annotation data are described in alphabetical order. Details on the choice of a database as source of information for the analysis in this thesis and the reasons for this choice are given in Section 3.2.

### 2.2.1. BIND

The BIND (Biomolecular Interaction Network Database, <http://bond.unleashedinformatics.com/>, Gilbert (2005)) is a project of the Blueprint Initiative for public bio-molecular data, which was started in 1998. Today it is owned by the media company Thomson Reuters. As its name suggests, BIND's main focus is

on biomolecular interaction data between RNA, DNA, molecular complexes, small molecules, photons (light) or genes. It also archives reaction, complex and pathway information, where molecular complexes and pathways are collections of these pairwise interactions (Bader *et al.*, 2003; Alfarano *et al.*, 2005). Data comprises automatically captured data from high-throughput projects, human-curated information from the scientific literature, as well as data integrated from other biological databases (Gilbert, 2005). However, this database does not contain explicit pathway related data. Data access is free for everyone, but limited to a small section of the whole database.

### 2.2.2. BRENDA

The BRENDA (BRaunschweig ENzyme DAtabase, <http://www.brenda-enzymes.info/>, Barthelmes *et al.* (2007)) enzyme information system is a manually annotated repository for enzyme data. Originally intended and published as a series of books in 1987, it was transformed into a publicly available database in 1998. BRENDA stores information on all enzymes that have been classified into the EC classification scheme of the NC-IUBMB. The range of data stored for each enzyme includes the catalyzed reaction, detailed description of substrate, cofactor and inhibition specificity, kinetic data, structure properties, information on purification and crystallization, properties of mutant enzymes, participation in diseases and amino acid sequences. Each single entry is linked to the enzyme source: the organism(s), tissue (if applicable), protein sequence, and to the literature reference (Barthelmes *et al.*, 2007; Schomburg *et al.*, 2004). BRENDA can be accessed via web-frontend and since 2007 the database can be downloaded as text file. It is free for academic users, whereas non-academic users need a license.

### 2.2.3. EcoCyc, MetaCyc, and BioCyc

**EcoCyc** (<http://ecocyc.org/>, Keseler *et al.* (2009)) is a model organism database for *Escherichia coli*. Launched in 1992, it stores the whole genome of the reference organism *E. coli* K-12 and information on genes, proteins, chemical compounds and molecular interactions such as enzymatic, transport and binding reactions, as well as metabolic and signaling pathways and regulatory networks obtained by annotation and literature-based curation (Karp *et al.*, 2004; Keseler *et al.*, 2005). For each enzyme, information like substrate specificity, kinetic properties, activators, inhibitors, cofactor requirements, and links to sequence and structure databases are available. The EcoCyc data can be queried via a web interface or downloaded and queried locally using a software package called Pathway Tools. While web access and data download is free for all users, the software is freely available only to academic users and available for a fee to commercial users.

**MetaCyc** (<http://metacyc.org/>, Caspi *et al.* (2008)) contains the same type of information as EcoCyc, but is not specific to a particular organism. It serves as a repository for information on many organisms, mainly microorganisms and plants, and is used as reference by a software called PathoLogic for the automatic reconstruction of the metabolic network of an organism based on its genome sequence. PathoLogic is part of the Pathway Tools software package. In MetaCyc each pathway is labeled with the organism(s) in which it is known to occur, based on wet-lab experiments reported in the literature evaluated to date. Since experimentalists have demonstrated the presence of

most pathways in only a small fraction of the organisms in which they actually occur, and because MetaCyc does not cover all known literature articles, the species information in MetaCyc is incomplete (Karp *et al.*, 2002). MetaCyc can be freely accessed via the world wide web and is also available for download. The aforementioned software package Pathway Tools can also be used to access MetaCyc.

**BioCyc** (<http://www.biocyc.org/>, Karp *et al.* (2005)) is a collection of organism specific pathway and genome databases (PGDBs), which was automatically generated using the PathoLogic software. The BioCyc collection of databases currently comprises 369 mostly eukaryotic and prokaryotic species. Each PGDB describes the genome and the predicted metabolic network including the respective information from MetaCyc. The PathoLogic software also predicts operons and candidate genes that might code for enzymes that are presumably present in the metabolic pathways, but could not be inferred from the genome. Each PGDB can be accessed using the Pathway Tools software package and is available for download in several formats. As of 2009, 482 of a total of 507 databases in the BioCyc collection have neither undergone manual curation nor review, whereas the remaining 25 have been subject to at least moderate manual curation.

#### 2.2.4. EMP/MPW

The EMP database (Selkov *et al.*, 1996) started as an effort to curate literature information on enzymology and metabolism into graphical representations of metabolic pathways. It was initiated in 1984 at the Russian Academy of Sciences to support internal projects in the mathematical simulation of cell metabolism by encoding as much of the known data relating to enzymology as possible. In 1995 the pathway diagrams covering primary and secondary metabolism, membrane transport, signal transduction pathways, intracellular traffic, translation and transcription were made freely available to other researchers. Later the pathways from EMP were integrated into the PUMA system (see Section 2.2.8) and further developed into MPW, the Metabolic Pathways Database. The original pathway diagrams of the EMP database were converted into a standardized data format. The stoichiometry of reactions as well as substrate and coenzyme specificity of enzymes, their sub-cellular locations, required prosthetic groups and cofactors, and taxonomic occurrence (not organism specific) of the reactions are presented on the respective diagrams (Selkov *et al.*, 1998). The EMP pathways can be downloaded from [ftp://ftp.mcs.anl.gov/pub/compbio/PUMA2/EMP\\_DATA/](ftp://ftp.mcs.anl.gov/pub/compbio/PUMA2/EMP_DATA/). However, this version has not been updated since 2002.

#### 2.2.5. ERGO

The ERGO database and genome analysis system (<http://ergo.integratedgenomics.com/>) has been developed at Integrated Genomics on the basis of the PUMA/WIT (see Section 2.2.8) system. It stores genome sequence and annotation data as well as metabolic reconstructions. Annotations can be done both automatically and manually. The goal is to improve functional annotation by exploiting similarity between genomes. The integrated pathway database (including pathway diagrams) is derived from the EMP pathway database. Visualization of pathways can be done using either these diagrams or the KEGG pathway maps. ERGO can be accessed via web frontend by registered users. Currently, 1074 genomes (completed and gapped ones) from bacteria (792), archaea (49),

eukaryotes (136) as well as viral genomes (241) are available in ERGO. Access to this database is not free of charge for anyone.

### 2.2.6. ExPASy

ExPASy (Expert Protein Analysis System, <http://expasy.ch/tools/pathways/>, Gasteiger *et al.* (2003)) hosts the Roche Applied Science Biochemical Pathways. This is a collection of images depicting very detailed information on bacterial metabolism. It is the computerized version of the well-known Boehringer wall chart called Biochemical Pathways originally assembled by Gerhard Michal (Michal, 1999). A keyword search for maps containing particular EC numbers and metabolites is possible. Information on EC numbers is retrieved from ExPASy's enzyme database. The provided information includes reaction name, equation, cofactors, as well as cross-references to other databases including a list of entries of the manually curated protein sequence database UniProtKB/Swiss-Prot (UniProt Consortium, 2008). Each UniProtKB/Swiss-Prot entry provides additional information on the enzyme as, for example, its occurrence in a particular organism. Enzyme and protein data are free to be downloaded as flat files, but pathway data does not exist in any computerized format other than digitized images.

### 2.2.7. KEGG

Initiated in 1995 under the Human Genome Program of the then Ministry of Education, Science and Culture of Japan, KEGG (Kyoto Encyclopedia of Genes and Genomes, <http://www.kegg.com/>, Kanehisa *et al.* (2008)) is a database system for storing, analyzing and annotating information on genome sequences, genes, enzymes, and chemical compounds with a special focus on the functional connection between these entities. KEGG's goal is to link genomic information with information on cellular processes (Kanehisa and Goto, 2000) by maintaining the gene catalog of every sequenced organism and mapping each component in the catalog to the KEGG pathway diagrams (Kanehisa, 1997). The KEGG pathway diagrams are based on the diagrams of the Boehringer wall chart and those of the Japanese Biochemical Society as well as on textbooks and online databases (Goto *et al.*, 1997; Kanehisa, 1996). When a new genome is put into KEGG, first the orthologs are automatically calculated, followed by manual annotation of ortholog identifiers. Then automated pathway reconstruction is performed including the search for missing enzymes or alternative metabolic routes followed by manual annotation of the predictions. The pathway reconstruction is undertaken by automatically matching the enzymes in the gene table with the enzymes on the pathway diagrams (Goto *et al.*, 1997). The KEGG data is daily updated (Kanehisa and Goto, 2000). Currently it contains the genomes of 102 eukaryotes, 849 bacteria, and 64 archaea. The KEGG data is freely available via web interface. It can also be downloaded free of charge as flat files.

### 2.2.8. PUMA/WIT

PUMA2 (<http://compbio.mcs.anl.gov/puma2/>, Maltsev *et al.* (2006)) was developed at Argonne National Laboratory's Mathematics and Computer Science Division as the successor of the WIT2 (WIT: What Is There) system. The WIT2 system in turn is the successor of another system called PUMA (Overbeek *et al.*, 2000). While WIT2 cannot

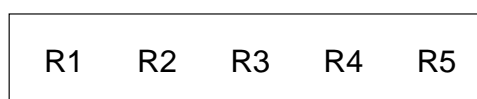
be found in the world wide web any more, PUMA2 is still online and functional. This database was designed to store genomic information along with genome annotation and metabolic reconstruction data. The goal behind this approach is to improve functional annotation by exploiting similarities between different organisms. Annotation of genomes and metabolic reconstruction is done automatically as well as manually (Maltsev *et al.*, 2006). Metabolic reconstruction is based on the genome annotation and the metabolic modules from the EMP/MPW database. Since the last release it is also possible to view the reconstructed metabolism on KEGG pathway maps. Currently, PUMA2 contains over 1,000 prokaryotic and eukaryotic genomes. It also stores data on gene annotations, enzymes, biochemical reactions and pathways, and provides links to further information in many other databases. PUMA2 is free to use for everyone via the web-based user interface. Links are provided for downloading genome sequences and annotation data as well as EMP pathway data. However, PUMA2 is not being maintained or updated by Argonne National Laboratory any more.

## 2.3. Formal Models for Metabolic Networks

Different models exist for representing metabolic networks: models relying on set theory as well as graph models of different kinds. Which model is best suited depends on the required precision needed for the desired analysis, or, in other words, how many details of the metabolic networks need to be represented in the model. Models coming into consideration are briefly introduced in the following section. In Section 3.3 the decision for one of these models is made and the reasons for this decision are explained. Although the notion of a graph is already used in the following sections, the provided information can be understood without detailed knowledge of the theoretical background. A formal introduction to graph theory is given in Section 4.

### 2.3.1. Set Model

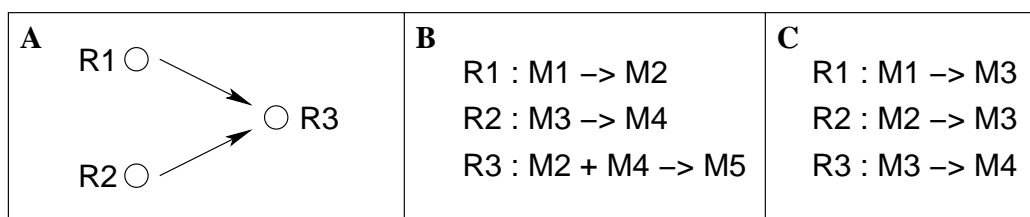
The simplest model for representing a metabolic network consists of a set containing the metabolic reactions or the metabolites or both as set elements. The elements are neither sorted nor interconnected. For identifying individual reactions and metabolites each set element can be labeled with a unique identifier. Based on these identifiers an artificial ordering can be established which accelerates searching for a particular set element. Representing networks using this model is very memory efficient. Its drawback is that it does not capture data on the interconnections between its elements and thus the structure of the metabolic network cannot be exploited for comparison. Figure 2.1 shows an example. Hong *et al.* (2004) as well as Liao *et al.* (2002) used this model for representing metabolic networks.



**Figure 2.1.:** Representation of a metabolic network as set of reactions R1 to R5.

### 2.3.2. Reaction Graph

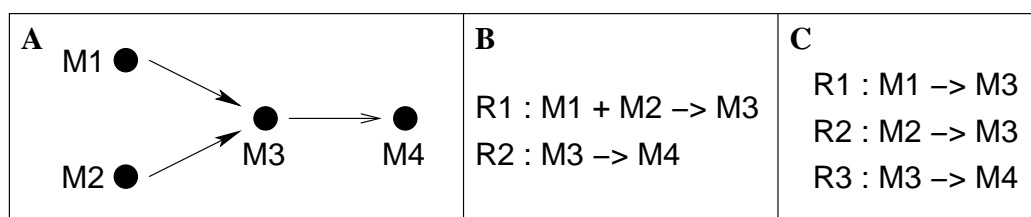
In a reaction graph reactions are modeled as nodes and metabolites as edges connecting two reaction nodes if a metabolite is produced by one of the reactions and consumed by the other. Edges may be directed from one reaction to another reaction if one wants to encode that an intermediate metabolite is produced by the former and consumed by the latter reaction. Nodes can be labeled with reaction identifiers. Figure 2.2 A shows an example. Reaction graphs can be represented by an  $N \times N$  matrix, where  $N$  denotes the number of reactions. Storing this matrix in the computer needs more memory space than the set model, but this comes with the advantage of capturing the network topology. However, the structure of the metabolic network cannot be reproduced from this model without ambiguities: from a reaction graph one cannot deduce whether products generated by different reactions and consumed by another reaction are identical. For example, the graph in Figure 2.2 A can be the model representation of each of the two differing sets of reactions in Figure 2.2 B and C. Ogata *et al.* (2000) as well as Heymans and Singh (2003) used this approach for modeling metabolic networks.



**Figure 2.2.:** A: Reaction graph consisting of three reactions R1, R2, and R3. B, C: The reaction graph in A can be the model representation of the set of reactions in B as well as of the differing set of reactions in C. From the reaction graph in A one cannot deduce whether R1 and R2 produce the same metabolite M3, which then is metabolized by R3, or whether R1 and R2 produce different metabolites which both are substrates of R3.

### 2.3.3. Metabolite Graph

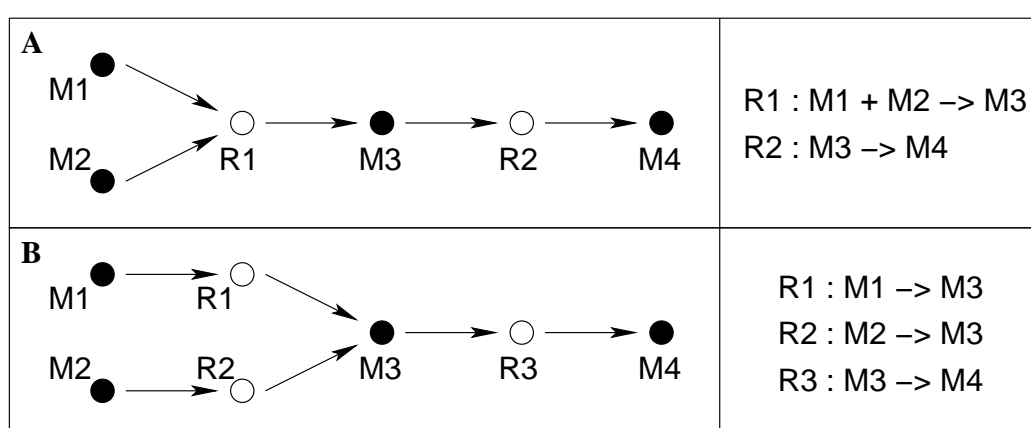
In a metabolite graph each metabolite is represented as a node, whereas reactions are encoded as edges connecting substrate metabolite nodes to product metabolite nodes. Edges may be directed to encode reaction directionality. Nodes can be labeled with metabolite identifiers. An example is shown in Figure 2.3 A. Metabolite graphs can be represented by an  $N \times N$  matrix, where  $N$  denotes the number of metabolites. As for the reaction graph, this model needs more memory space than the set model, but in exchange the network structure is captured by the model. However, as for the reaction graph, the network structure cannot be reconstructed unambiguously: one cannot distinguish whether two metabolites are involved in the same reaction. For example, the metabolite graph in Figure 2.3 A can be the model representation of the differing sets of reactions in Figure 2.3 B and C. Wagner and Fell (2001) used this approach for modeling the metabolic network of *Escherichia coli*.



**Figure 2.3.:** A: Metabolite graph consisting of four metabolites M1, M2, M3, and M4. B, C: The metabolite graph in A can be the model representation of the set of reactions in B as well as of the differing set of reactions in C. From the metabolite graph in A one cannot deduce whether a single reaction transforms both M1 and M2 into M3, or whether one reaction transforms M1 into M3 and another one transforms M2 into M3.

### 2.3.4. Bipartite Graph

A bipartite graph is a graph structure that can be employed for unambiguously modeling metabolic networks. A bipartite graph encompasses two types of nodes. Edges always connect two nodes of different type. For modeling metabolic networks one type of nodes resembles the reactions and the other type the metabolites. Edges connect substrate and product metabolites to the respective reaction. If directed edges are used then edges lead from substrate metabolites to reactions and from reactions to product metabolites. As for the previous graphs, this data structure needs more memory space than the set model. Bipartite graphs can be represented by an  $N \times M$  matrix, where  $N$  denotes the number of reactions and  $M$  the number of metabolites. Two example graphs and their reaction equations are shown in Figure 2.4. The reaction equations are the same as in Figure 2.3 B and C, however, the differing sets of reactions lead to different bipartite graphs. The bipartite graph model was used by Forst and Schulten (2001) for modeling metabolic networks.



**Figure 2.4.:** Two bipartite graphs and their corresponding reaction equations. A: Bipartite graph corresponding to Figure 2.3 A based on reaction set Figure 2.3 B. B: Bipartite graph corresponding to Figure 2.3 A based on reaction set Figure 2.3 C. For the bipartite graph model the differing sets of reactions yield different graphs.

## 2.4. Concepts for Metabolic Network Comparison

When comparing metabolic networks, methods are needed for assessing how similar two such networks are to each other. In other words, a distance between two metabolic networks needs to be calculated. Since in Section 3.3 the decision is made to model metabolic networks as graphs in this thesis, formally the problem is to assess the difference between two graphs.

Measuring the distance between graphs has been the topic of research for many years. It is usually referred to as graph matching. Applications can be found in various areas: in chemistry it is used for mapping chemical formulae in database searches, in medicine diagnoses of certain diseases may be based on the results of automatic image analyses, and in computer science methods for face recognition are developed. All have in common that the object under study is represented as graph which has to be compared to other graphs. Distance measures on graphs can be classified into maximum common subgraph-based, feature-based, and edit operation-based ones. The respective concepts are introduced in the following sections. The decision regarding which concept to use in this thesis is discussed in Section 3.4.

### 2.4.1. Maximum Common Subgraph-based Approaches

Maximum common subgraph-based distance measures rely on isomorphisms between the two graphs to be compared. Of particular interest are the maximum common subgraph and minimum common supergraph of two graphs. The maximum common subgraph of two graphs can be interpreted as the largest common part between the two graphs, where both the names of the nodes are identical as well as the connections between the nodes. The minimum common supergraph of two graphs is the smallest graph containing the two original graphs. Bunke and Shearer (1998) defined a distance measure based on the ratio between the size of the maximum common subgraph of two graphs and the size of the larger of both graphs. This distance measures how much of the larger graph can also be found in the smaller one. Fernández and Valiente (2001) introduced another distance on graphs which is based on both maximum common subgraph and minimum common supergraph. They subtract the size of the maximum common subgraph from the size of the minimum common supergraph. This distance measures the amount of differences between both graphs.

### 2.4.2. Feature-based Approaches

In feature-based methods, for each graph, one or more features (e.g. number of nodes, number of edges, shortest path, etc.) are calculated and stored in a feature vector. These feature vectors serve as representatives of the original graphs, and distances are calculated between the feature vectors instead of the original graphs themselves. Various distance measures can be applied like the Euclidean distance or the Tanimoto coefficient (Willett *et al.*, 1998), which is used, for example, in chemical database searching. It depends on the field of application as to which properties are included in the feature vector and which distance measure is appropriate to compare them.



### 2.4.3. Edit Operation-based Approaches

These approaches are based on the concept that one of the graphs is a distorted version of the other. A sequence of edit operations that transforms one graph into the other is computed. Each edit operation corresponds to one error introduced in one graph during distortion. The more errors occur, the more edit operations are needed to transform one graph into the other and the more different become the two graphs. In general there will be more than one possible sequence of edit operations. The edit distance is defined as the minimum number of edit operations needed for the transformation, or, in other words, the length of the shortest of these sequences. Furthermore, this method allows for the assignment of specific costs to each edit operation. In this case the edit distance is defined as the minimal cost over all possible sequences of edit operations.

## 2.5. Clustering Methods

In this thesis a set of organisms has to be automatically classified into previously unknown subgroups according to their metabolic network variants, where the number of subgroups also is not known in advance. Methods performing this task are subsumed under the term clustering or cluster analysis. Clustering is also referred to as unsupervised learning as opposed to supervised learning. While in supervised learning, a collection of pre-classified or labeled items is already known and the problem is to label another, yet unlabeled, item, in cluster analysis a collection of unlabeled items has to be grouped into a previously unknown number of meaningful clusters without having any prior knowledge about the labels or classes. The goal of the clustering process is to achieve a partitioning (classification) of the set of items such that items in the same cluster are similar to each other while different clusters are separated from each other. The distance between two single items and between two clusters (which can be regarded as collections of items) as well as between an item and a cluster can be defined in different ways. This strongly influences the capabilities of the particular clustering method to detect certain structures (e.g. different shapes) in the data set. The basic steps in clustering are:

- (i) selecting a distance measure for comparing two items
- (ii) choosing a clustering procedure and a clustering criterion
- (iii) estimating the number of clusters and/or validating the resulting clusters.

### Distance Measure

The choice of an appropriate distance measure (step (i)) depends on the field of application and the goal of the analysis. Suitable distance measures for the approach presented in this thesis are discussed in Section 3.1 and are formally introduced in Section 4.2.

### Clustering Procedure and Criterion

Approaches for the clustering as such (step (ii)) can be classified according to the technique used to find the clusters into hierarchical, partitional and density-based approaches (Jain *et al.*, 1999). Hierarchical methods produce a hierarchy of clusters by either of two

strategies. Firstly, continuously choosing and then subdividing a cluster into two new clusters starting from a single initial cluster containing all items (divisive clustering). Secondly, continuously merging two clusters into a single one starting from a set of singleton clusters each containing one item only (agglomerative clustering). In either case the result is a hierarchy of clusters, called a dendrogram, as well as distance levels at which clusters change. In order to yield a classification of the items, this dendrogram can be cut at a certain level, based on some criteria. Possible criteria are, for example, the number of desired clusters or the maximum distance between members of the resulting clusters or the minimum distance between different clusters. In case none of these values is known, some internal cluster validation measures can still be used to estimate the number of clusters from the data.

Partitional approaches already start with a partitioning into clusters which is then continuously refined by shuffling items between clusters depending on a criterion function that is to be optimized. The criterion can be defined either locally on a subset of the items or globally over all items. A possible criterion function to be minimized is the squared error, which works well with isolated and compact clusters (Jain *et al.*, 1999).

In density-based approaches clusters are defined depending on the local density of items. For example, the method described by Ester *et al.* (1996) requires each item in a cluster to have a minimum number of items within a particular distance. In other words: the item density in the neighborhood of each cluster member has to exceed a certain threshold. Clusters are formed according to the following two rules: each item belongs to at most one cluster and two items are in the same cluster if each item is within the minimum distance of the other.

Hastie *et al.* (2001) additionally describe mixture modeling approaches. Mixture modeling assumes that the data is a sample from a population that can be described by a probability density function, which in turn is a mixture of component density functions, where each component describes one of the clusters. The parameters of this model are fit to the data by maximum likelihood or corresponding Bayesian approaches.

Three well-known hierarchical clustering methods are now introduced:

**Complete linkage agglomerative clustering.** The complete linkage agglomerative clustering is a hierarchical approach starting with each item in a separate cluster. In each of the following steps the two clusters that are closest to each other are merged to form a new cluster. For the cases that a cluster consists of more than one item, it is necessary to define a distance between two clusters. In the complete linkage approach this distance is defined as the maximum of all distances between pairs of items, one from the first cluster and the other one from the second cluster. The algorithm finishes when a stopping criterion is reached or when only one single cluster remains. The complete linkage agglomerative clustering produces tightly bound or compact clusters (Eckes and Roßbach, 1980).

**Average linkage agglomerative clustering.** This method differs from the previous one in the way it measures the distance between two clusters. Here, the average of all pairwise distances between one item in the first and another item in the second cluster is used, instead of relying on the maximum distance. Therefore, the resulting clusters are less compact than those resulting from the complete linkage clustering technique. Using this approach spherically shaped clusters should be easily detectable. Whereas the complete linkage approach can be applied to all distance measures, the average linkage approach strictly speaking is most appropriately be applied on distance measures for

which the mean of several distances is a sensible value (Eckes and Roßbach, 1980).

**Ward clustering.** The Ward clustering method (Ward Jr, 1963) uses a special objective function called error sum of squares (*ESS*) for measuring the loss of information associated with the representation of a set of items in a cluster  $C$  by one item only, namely the centroid:

$$ESS(C) = \sum_{x_j \in C} \|x_j - \bar{x}\|^2, \quad (2.1)$$

where  $\bar{x}$  is the centroid of cluster  $C$ .

The *ESS* of a clustering is calculated as the sum of *ESS*s of the individual clusters ( $\sum_C ESS(C)$ ). The *ESS* is zero if all items are put into separate clusters, but increases if two different items are put into the same cluster. The more distinct the different items in each cluster are, the higher the *ESS*. The algorithm uses an agglomerative strategy: it successively merges those two clusters that, when merged, cause the least increase of the overall *ESS*, and stops if all items are put into a single cluster. Thus, a hierarchy of clusters is produced. The Ward method is a clustering strategy that keeps the intra-cluster distance small and thus produces compact clusters. Moreover, the resulting clusters tend to be equally sized (Eckes and Roßbach, 1980). This approach was developed under the assumption that the squared Euclidean metric is used as distance measure between items. However, it can also be applied to other distance measures (Eckes and Roßbach, 1980).

### Cluster Validation and Estimation of Number of Clusters

It is in the nature of the clustering idea that the resulting classification of the item set can be neither verified nor falsified. Nevertheless, it is important to assess how well the resulting classification represents the true structure of the data. Firstly, because different clustering algorithms have different biases due to their specific objective function. Secondly, even if there is no structure in the data at all, each algorithm will still produce a classification which in this case is meaningless. At least to some extent a quality assessment can be acquired using cluster validation measures.

Cluster validation measures can be classified into internal and external (Halkidi *et al.*, 2001; Handl *et al.*, 2005). External validation measures rely on the correct class labels. They are useful when evaluating clustering approaches on benchmark data, but are definitely not applicable if the labels are unknown. In contrast to these, internal validation measures comprise all methods that base their quality estimate on information intrinsic to the data.

Internal validation measures can be classified according to their criterion into measures assessing compactness, connectedness, separation, or combinations thereof (Handl *et al.*, 2005). Examples are within-group sum of squares for assessing compactness (Duran and Odell, 1974), k-nearest-neighbor consistency (Ding and He, 2004) and connectivity (Handl and Knowles, 2005) for connectedness, or average (weighted) inter-cluster distance and minimum separation between all pairs of individual clusters for separation (Handl *et al.*, 2005). If the validation measure either exceeds or undercuts a measure-dependent threshold, the resulting clustering is considered valid, otherwise it is considered invalid.

Most of these internal validation measures can also be used to estimate the number of clusters in a dataset. The strategy here is to compute the classifications for a range of different numbers of clusters and plot the performance under the internal validation

measure as a function of the number of clusters. The optimal number of clusters can then often be identified as a knee in the resulting performance curve if both the employed clustering algorithm and the internal measure are adequate for the dataset under consideration (Handl *et al.*, 2005). However, this method is difficult to apply, because often the knee is not easy to identify.

Another way to estimate the degree to which distance information in the original data is preserved in a partitioning is to compare the cophenetic matrix of the partitioning with the matrix holding the distance information (Romesburg, 1984; Halkidi *et al.*, 2001). The cophenetic matrix  $C$  is an  $N \times N$  matrix, where  $N$  denotes the number of items, and  $C(i, j) =$  the cophenetic distance between items  $i$  and  $j$ , which is the intergroup dissimilarity at which the two items  $i$  and  $j$  are first combined into a single cluster. The similarity between the two matrices can then be assessed, for example, using the cophenetic correlation coefficient (*cpcc*), which is closely related to the Pearson product-moment correlation coefficient (Romesburg, 1984). The clustering is considered valid if the similarity measure exceeds some threshold.

This procedure can conveniently be used in an automated approach for determining the number  $k$  of clusters in the data set: for a range of possible number of clusters  $k$ , the partitioning is represented by means of its cophenetic matrix  $C$ , and the *cpcc* is calculated between the cophenetic matrix and the original distance matrix. The cophenetic matrix  $C$  of a partitioning is defined as (Handl *et al.*, 2005; Halkidi *et al.*, 2001):

$$C(i, j) = \begin{cases} 0 & \text{if items } i \text{ and } j \text{ are in the same cluster} \\ 1 & \text{otherwise.} \end{cases} \quad (2.2)$$

Then that value for  $k$  is chosen as number of clusters in the data set for which the *cpcc* reaches its maximum value, because the higher the *cpcc* the more similar are the two matrices and thus the closer is the classification to the information contained in the original distance matrix. The *cpcc* can also be used to compare classifications resulting from different clustering techniques as long as these are based on the same distance data. The *cpcc* is defined as follows (Halkidi *et al.*, 2001):

$$cpcc(D, C) = \frac{\frac{1}{m} \sum_{\{(i,j)|1 \leq i < j \leq n\}} (D_{i,j} C_{i,j}) - \mu_D \mu_C}{\sqrt{\left[ \frac{1}{m} \sum_{\{(i,j)|1 \leq i < j \leq n\}} D_{i,j}^2 - \mu_D^2 \right] \left[ \frac{1}{m} \sum_{\{(i,j)|1 \leq i < j \leq n\}} C_{i,j}^2 - \mu_C^2 \right]}}, \quad (2.3)$$

where  $D$  denotes the distance matrix,  $C$  the cophenetic matrix,  $n$  the number of data points,  $m = n(n-1)/2$ , and  $\mu_D, \mu_C$  are the means of the matrices  $D$  and  $C$ , respectively:  $\mu_D = \frac{1}{m} \sum_{\{(i,j)|1 \leq i < j \leq n\}} D_{i,j}$ ,  $\mu_C = \frac{1}{m} \sum_{\{(i,j)|1 \leq i < j \leq n\}} C_{i,j}$ .

Cluster validation can also be performed in a more qualitative way by comparing clustering results calculated by different clustering approaches (Handl *et al.*, 2005): if clustering results are similar, this is a hint towards a good quality of the clusterings. However, if clustering results differ, this might either indicate that there is no obvious structure in the data or might be due to the inappropriateness of the applied clustering approaches or criteria.

## 2.6. Related Work

Several approaches already exist for comparing organisms based on their metabolic networks. In most cases, however, the goal is to derive a phylogenetic grouping of the analyzed organisms. Relevant approaches are summarized in the following sections, ordered by the respective date of publication.

**Manual Pathway Alignment for Medicine and Metabolic Engineering.** One of the first approaches for systematic pathway alignment, which involved a lot of manual work, was presented by Dandekar *et al.* (1999). Using glycolysis as an example, they elucidate how pathway alignment across a set of organisms can be performed and prove the usefulness of their approach by showing that the identified differences between pathway variants in different organisms are of interest for medicine and drug design as well as for metabolic engineering. In more detail, their approach is to refine given annotations and manually align the metabolic networks. Then biochemical information on the catalyzed reactions is supplemented, namely reactants and reaction stoichiometry. The drawback of this method is the significant amount of manual work. For enabling comparison of large metabolic networks or more than a few pathways and large sets of organisms, major parts of the analysis would have to be automated and this, as yet, has not been undertaken.

**Alignment of Linear Pathways for Pattern Search.** Tohsato *et al.* (2000) presented an automated approach for multiple alignment of metabolic pathways. Following Galperin *et al.* (1998), who promote the opinion that reaction similarity does not necessarily correlate with sequence similarity, they define a distance between enzymes based on the position of the respective EC numbers in the hierarchy of the EC classification scheme. The overall distance of two pathways is expressed as the information content of the pathway alignment. Their alignment algorithm is an extension of the global alignment algorithm based on dynamic programming. Only linear pathways can be aligned using this method. Branched and circular pathways need to be split beforehand. Therefore, this method is appropriate for finding patterns in linear pathways, but less suited for automatically comparing a set of pathways or the overall metabolic reaction network of a set of organisms.

**Pairwise Graph Comparison for Locally Similar Subgraph Search.** Ogata *et al.* (2000) described a graph comparison approach for detecting locally similar subgraphs. Given two graphs, a list that defines which node in the first graph corresponds to which node in the second graph is needed. In the case of enzyme graphs, nodes can be defined as corresponding if they have identical EC numbers. Then a clustering algorithm is applied that groups nodes together if in both graphs the length of the respective shortest path to any member of the group is smaller than a user defined gap distance. The authors use this method to compare reaction graphs with graphs representing the neighborhood of corresponding genes in the genome in order to construct functionally related enzyme clusters. This method could also be used to compare two reaction graphs deduced from the pathway implementations in two organisms and would then yield conserved subpathways as result. However, this method allows only pairwise comparisons and does not

produce a classification of metabolically similar organisms.

**Phylogenetic Trees based on Pathway Structure and Gene Sequence.** The goal of the method for pathway comparison published by Forst and Schulten (2001) is to derive phylogenetic trees from information in metabolic networks. They define a distance measure that combines sequence information of involved genes with structural information about the corresponding reaction networks. Substrates of the reactions that are encoded in the genome (e.g. if they are proteins) are also considered in the distance calculation. The distance is based on sequence similarity and multiplied by a special factor for weighting orthologs and paralogs differently. These individual distances are summed up to form the overall distance between two pathways. If two networks with different topology are compared, the method penalizes gaps with a special gap cost and thus takes the structure of the network into account. A gap occurs when an enzyme or substrate in one of the organisms has no match in the other organism. This approach strongly relies on gene sequence information for comparing two networks. However, though sequence similarity might be well suited for deducing phylogenies, it is not appropriate if the focus is on comparing reaction content or functionality of metabolic pathways.

**Classifying Organisms based on Pathway Profiles.** Liao *et al.* (2002) presented an approach for comparing and classifying organisms based on metabolic pathway information. They construct profiles of metabolic pathways, which are essentially strings representing presence or absence of various metabolic pathways. Pathways are taken from the WIT database. A pathway is said to be present if all involved enzymes are annotated in the organism. Pairwise similarity is calculated based on these profiles weighting each pathway attribute according to its position in a hierarchy of pathways. Similarities are then transformed into distances which are used for clustering the organisms. Results of this approach strongly depend on a proper choice of pathways and correct annotation. If only one single reaction is not annotated in a particular organism, either because it is truly missing or due to a missing annotation, the entire pathway is classified as missing. Clustering based on all pathways has the disadvantage that presence or absence of a single, but possibly significant, pathway might not be reflected in the resulting dendrogram or classification.

**Phylogenetic Trees based on Pathway Structure and EC Numbers.** Heymans and Singh (2003) presented another method for constructing phylogenetic trees by comparing metabolic pathways. In their opinion evolutionary distance is based on the divergence of the elements constituting the pathways as well as the divergence of the network structure. For this reason they define a distance measure that takes both aspects into account. They model metabolic networks as enzyme graphs. Enzyme similarity is calculated as distance in the respective EC number representation of the enzymes. Structural similarity is assessed for each enzyme node in the network graph, on the basis of the differences in adjacent nodes. The overall distance between two pathways is the sum of all individual distances. Finally, the neighbor joining clustering method from the Phylip software package (<http://evolution.genetics.washington.edu/phylip.html>) is used to construct phylogenetic trees. This approach is less suited for metabolic pathway comparison if the focus lies on functional aspects because it takes sequence information into account.

**Phylogenetic Trees based on COG Classification of Enzymes.** Hong *et al.* (2004) published an approach for constructing phylogenetic trees based on metabolic subpathway reaction content. A subpathway is one of 64 subpathways derived by subdividing the overall metabolic reaction network according to the COG classification (Tatusov *et al.*, 1997) of the metabolic reactions. The subpathway reaction content for a particular organism and subpathway is defined as the number of reactions annotated in this organism for this subpathway in relation to all reactions of this subpathway. By applying the Pearson correlation coefficient on the reaction content of all subpathways for two organisms, a pairwise distance is calculated. Subsequently, distance data are clustered using the complete linkage hierarchical clustering algorithm. However, for each COG category only the relative number of present enzymes is used for comparing two organisms. This measure does not distinguish whether the same or different reactions are missing in both organisms and thus it is less suited for comparing metabolic networks. For example, imagine a subpathway consisting of 4 reactions, one organism having 2 reactions annotated, another organism having the same two reactions annotated, and a third organism having the other two reactions annotated. All three organisms would be considered identical by this measure, although only the first two are actually identical.

**Detecting Pathway Variants by Comparing Gene Function and Pathway Structure.** Ye *et al.* (2005) developed an approach for automated detection of subsystem variants in a set of organisms. A subsystem is a group of related functional roles (such as enzymes or transporters) jointly involved in a specific aspect of the cellular machinery. It is modeled as a directed graph, where nodes represent functional roles and edges connect one such role with another one if the latter consumes a product of the former. The definition of a subsystem includes a list of functional roles as well as a table of genes assigned to these roles in the genomes of a variety of different organisms. In this approach, starting from a particular subsystem of interest, all subsystem variants (subgraphs of the original subsystem) that are present in any of the organisms are found. This approach also yields a grouping of organisms according to their particular subsystem variant. For finding the variants, the quality of candidates is assessed based on their functionality and compactness. A subgraph is called a functional pathway if it contains at least one connected path from a set of source nodes to a set of sink nodes. A subgraph is called compact if it does not include functional roles that do not contribute to functional pathways. This approach is well suited to compare metabolic networks, because, firstly, the functional aspect is assessed via functional roles regardless of the genome sequence coding for the respective enzyme and, secondly, a notion of the functionality of subsystems is taken into account. However, drawbacks are that source and sink nodes of a subsystem need to be defined manually and that functional pathways are not necessarily stoichiometrically feasible.

**Comparing Chemical Reaction Networks using Set Theory.** Forst *et al.* (2006) published an algebraic method for comparing metabolic networks. They developed the Vienna Reaction Network Library, Vienna-RNL, which implements basic set-theoretic operations on chemical reaction networks. Using these operations one can detect all metabolic innovations, i.e. reactions that occur in at least one organism from a predefined set of organisms and are missing in all organisms from another predefined

set. The approach can also be used to derive phylogenetic trees. Therefore, a pairwise distance between two networks is defined as the number of reactions occurring only in one of the networks over the total number of reactions in the joined network. For a set of organisms under study, these distances are then used to compute phylogenetic trees using the Fitch algorithm implemented in the Phylip software package (<http://evolution.genetics.washington.edu/phylip.html>) as well as using the splits-decomposition algorithm from the SplitsTree software package (<http://www.splitstree.org/>). A drawback of this approach, however, is that the classification of organisms for finding metabolic innovations needs to be defined in advance, e.g. relying on taxonomic information. It would be favourable to automatically derive a classification of organisms that reveals these metabolic innovations.



---

# Methodology

---

In this chapter decisions in relation to which methods to apply in this thesis are made. In the first section, different strategies and their data requirements for assessing the differences between metabolic networks are discussed, and decisions are made in relation to which distance measures to implement. In the following sections, an appropriate data source and data model for metabolic networks and a suitable theoretical framework for formally defining the distance measures, as well as clustering techniques for automatically classifying the analyzed organisms are chosen. These choices are made on the basis of the considerations in the first section of this chapter as well as on the information provided in the respective sections of the previous chapter, where data sources, data models, theoretical frameworks for distance measures and clustering techniques were introduced.

### 3.1. Assessing the Difference between Metabolic Networks

In this section, it is discussed which information about metabolic networks can be used to assess how similar two organisms are in terms of their metabolic capabilities, and how this information can be exploited to define distance measures.

As has been mentioned in Section 2.1, metabolic networks comprise information on constituting reactions (reaction content) and metabolites (metabolite content) as well as structural information on how reactions and metabolites are interconnected. In principle, this enables three different approaches for comparing such networks: firstly, based on the constituents of the network (reactions and metabolites), secondly, based on the structure (reaction interdependencies), and thirdly, based on the function of the network. Distance measures on metabolic networks may take one single aspect into account or may be based on combinations of different aspects.

First of all, distance measures that are based only on the constituents of the network, namely reactions and metabolites, are discussed. Then, distance measures additionally

taking into account the structure of the networks are considered. Distance measures based only on the structure are not taken into account, because a similar structure of two metabolic networks does not imply that their metabolic function is similar. The function is encoded in reaction and metabolite identifiers and the structure of the network. Finally, approaches for assessing whether a network is functional or not are discussed.

Since in this thesis different distance measures will be implemented and their performance evaluated, all distance measures will be normalized, such that the values of all distances are within the interval  $[0, 1]$ . Normalization simplifies comparing clustering dendrograms obtained from different distance measures.

Although in principle, distance measures do not have to be metrics if they are used for clustering, in this thesis they are nevertheless constructed to have this property. This is because metrics in particular fulfill the triangle inequality, which expresses the transitivity of the distance relation: the triangle inequality implies that if item A is close to item B and B is close to item C then A must be close to C as well. The triangle inequality in particular demands that the distance between items A and C is smaller than or equal to the sum of the distances between A and B and between B and C. Although this boundary could be chosen differently, its existence is of importance. Distances for which A and C are far apart, although A and B as well as B and C are close contradict the intuitive notion of a distance measure.

Distance measures will be formally defined in Section 4.2. Their performance will be evaluated in Sections 6.1.1 and 6.1.2, and the decision as to which distance measure is most appropriate for comparing metabolic networks will be made in Section 6.1.3.

### 3.1.1. Reactions and Metabolites

An intuitive distance measure between two metabolic networks is the number of reactions and metabolites that are present in one of the networks while missing in the other and vice versa. The higher the amount of these mutually missing reactions and metabolites, the more different are the two networks. This distance can formally be realized as edit distance or as maximum common subgraph-based distance. However, this measure does not take into account how many reactions and metabolites are common to both networks. These can be considered by dividing this distance measure by the total number of reactions and metabolites in the supernetwork constructed by joining both networks to be compared. This procedure, at the same time, normalizes the distance measure. In this case the resulting distance corresponds to a Soergel type distance. A normalization can also be achieved by dividing the distance measure by twice the number of reactions and metabolites of the largest of all networks involved in the comparison. Note that normalizing by a value that changes for different pairs of networks, like reactions and metabolites in their supernetwork, influences the properties of the distance measure, which is not the case when normalizing by a constant value, like twice the number of reactions and metabolites of the largest of all networks involved in the comparison.

Instead of focusing on the differences, distance measures can also be based on the number of common or shared reactions and metabolites. The higher this amount, the more similar are the compared organisms. For creating a distance measure this value needs to be subtracted from a maximum value like the amount of reactions and metabolites in the larger of the two networks. Naturally, these distances are closely related to the ones based on mutually missing reactions, and they can also be realized as edit distance

or as maximum common subgraph-based distance. In particular, if the amount of reactions and metabolites in the supernetwork is used as maximum value, the resulting distance is identical to the one described above that counts mutually missing reactions and metabolites. These distances can be normalized by dividing them by the amount of reactions and metabolites in the larger of both networks or the number of reactions and metabolites in the supernetwork.

All distance measures considered so far take both reactions and metabolites into account. However, they can also be constructed to take into account either only reactions or only metabolites. A distance measure based only on reactions is very similar to comparing organisms based on their gene content, i.e. counting the number of shared and mutually missing orthologous genes. One difference is that, when comparing metabolic networks, only metabolic genes are taken into account, i.e. genes coding for enzymes that catalyze metabolic reactions. Another difference is that the relation between genes and enzymes is not always a one-to-one relation, since several different genes may code for the same enzyme or a particular enzyme may actually be a complex of several proteins, each of which might be encoded by a different gene.

In this thesis, two distance measures that focus on the differences between two metabolic networks will be implemented, firstly, number of reactions and metabolites that are not common to both networks divided by number of reactions and metabolites in the supernetwork, and secondly, number of reactions and metabolites that are not common to both networks divided by twice the number of reactions and metabolites in the largest of all networks involved in the comparative analysis. The former distance will be referred to as **edit distance** and the latter as **Soergel type edit distance**. Furthermore, a distance measure that focuses on the common parts of two metabolic networks will be constructed relying on the number of reactions and metabolites that are common to both networks divided by the number of reactions and metabolites in the larger of the two networks. This measure will be referred to as **mcs type edit distance**. Each distance measure will be implemented in three versions: firstly, based on reactions and metabolites, secondly, based only on reactions and thirdly, based only on metabolites.

### 3.1.2. Reaction Neighborhood

For constructing distance measures, the structure or topology of the metabolic network can be exploited as well. The structure provides information on which reactions produce the metabolites that serve as substrates for another reaction. In this sense, the structure also carries information on the network function.

The idea for developing a neighborhood sensitive distance measure is to incorporate information on the reaction neighborhood into the edit cost for deleting and inserting a certain reaction node. Heymans and Singh (2003) already considered incoming and outgoing edges for comparing two reaction nodes, but they exploited only structural similarity. In contrast to this, the goal here is to deduce the importance of a particular reaction node in the metabolic network by analyzing the relationship between this reaction and its neighboring reactions. One might say that reactions that are connected to many other reactions (via some metabolite) are more important than reactions with less connections, because they are hubs (branching points) in the metabolic network. So the higher the number of incident edges (coming from different reactions via some metabolite) the higher the cost should be for deleting this node. On the other hand, in

an unbranched chain of reactions, any reaction is of profound importance for the functionality of the whole chain, since the removal of any single reaction would already lead to malfunction. However, if several reactions exist converting the same substrate(s) into the same product(s) (e.g. differing only in the use of a co-factor), it does not matter for the functionality of the network if one of them is missing.

For the development of this distance measure, reactions that are adjacent to a particular reaction are categorized into one of two classes: synonymous reactions and adjacent reactions. Synonymous reactions are reactions that catalyze the same reaction, i.e. the same substrates are transformed into the same products and the reaction directionality is the same. Adjacent reactions comprise reactions that produce metabolites that are consumed by the reaction under investigation, and reactions that consume metabolites that are produced by the reaction under investigation. Based on these considerations a distance that assigns a cost only to reaction nodes and weights each reaction according to the number of synonymous and adjacent reactions can be constructed. The higher the number of synonymous reactions, the smaller the weight, and the higher the amount of adjacent reactions, the larger the weight.

In this thesis, this type of distance measure will be implemented in three versions, each relying on its own function for combining synonymous and adjacent reactions. These distances will be referred to as **neighborhood sensitive reaction edit distances**.

### 3.1.3. Network Function

Automatically assessing whether a metabolic network is functional based only on the network data available in public databases is not yet possible. To answer this question, wet-lab experiments would have to be performed for curating the data and input as well as output metabolites of the networks would have to be defined. Ye *et al.* (2005) implemented a semi-automated approach, in which input and output metabolites of the network have to be specified manually by the user. Other approaches are EFM, EP or Petri Net analysis. However, applying any of these approaches implies a huge amount of manual effort for curating the data. Thus, they cannot be applied in an automated approach.

Therefore, in the approach developed in this thesis, no automated classification into functional or non-functional is attempted. Instead, the decision as to whether a pathway is functional or not is left to the user to make. However, network function can, to some extent, be taken into account by subdividing the overall network into smaller pathways representing functional units. This can be done by relying on an existing set of pathways, like the ones defined in the KEGG database, or by defining one's own pathways. Subsequently, any of the above discussed distance measures can be applied to assess the similarity between the organism specific implementations of these pathways. This approach is realized in this thesis by performing the comparative analysis on the overall reaction network, the KEGG pathways, as well as on user defined pathways.

## 3.2. Data Source – Decision

Databases that come into consideration as data source for metabolic reaction data as well as functional annotation data of organisms were reviewed in Section 2.2. In this

section the most suitable will be chosen. A suitable database needs to contain information on metabolic reactions including reaction equation, direction and stoichiometry as well as relevant metabolites. Moreover, annotation information about which organisms implement the given reactions is needed. It would be beneficial if the database provided definitions of metabolic pathways grouping reactions together that are involved in the same functional process. The more comprehensive the database is, i.e. the more organisms and pathways are represented in the database, the more substantial will the analysis results be. In order to enable automatic processing, data access should be possible as flat file download. Furthermore, access to the data should be free of charge at least for academic users. For data consistency reasons, a solution relying on a single database is preferred over a solution that involves combining data from different data sources.

Most of the databases do not meet all criteria and therefore cannot be used as data source. At the time the data was needed, the BRENDA database was not available for download. The BioCyc family of databases does not yet provide pathway annotation data for a comprehensive set of organisms. The EMP database does not include any data about which organism implements the EMP pathways. ExpASY, like BRENDA, offers comprehensive data on metabolic reactions, but these reactions are not organized in pathways. KEGG combines genome annotation data and pathways in computerized format. Also, this data is available for most of the currently sequenced organisms. However, annotation and pathway reconstruction is mainly undertaken automatically. Therefore, pathway annotation data might not be without errors. PUMA2 is as comprehensive as KEGG. However, data is available only as genome annotation data, and pathway reconstruction would have to be done by mapping EC numbers to pathways. Reactome and UM-BBD are very specialized databases, which do not provide the required comprehensiveness.

Considering the above summarized benefits and drawbacks, the decision is made to use **KEGG** as data source for the comparative metabolic network analysis in this thesis. In addition to the data mandatory for the analyses, the KEGG database provides with the KEGG pathways a segmentation of the overall metabolic network into pathways of functionally related reactions, which can be readily used as a segmentation of the overall metabolic network for taking functionality into account as discussed in Section 3.1.3.

### 3.3. Data Model – Decision

As has been described in Section 3.1, a number of different distance measures are implemented and compared against each other, namely reaction-based, metabolite-based, reaction and metabolite-based, as well as reaction neighborhood sensitive distance measures. Since for different distance measures different types of data are needed, two strategies can be followed when it comes to modeling the data: one can either choose the most suitable model for each single distance measure, or rely on models that are suitable for more than one distance measure. The latter option has the advantage that the same theory can be relied upon for constructing different distance measures.

For a distance measure that assesses presence or absence only of reactions, the set model is appropriate (see Section 2.3 for a review on different models). The same is true for a distance measure assessing only metabolites. Even if presence or absence of reactions and metabolites is considered, the set model suffices. In this case the set model

would have to comprise both reactions and metabolites. Neighborhood information is not represented in the set model, which is why this model cannot be used for developing neighborhood sensitive distance measures. Also the metabolite graph model cannot be used, since it does not capture explicit information on reactions. The reaction graph model would be suitable, because this model captures reactions as well as reaction adjacency information. However, this model does not hold information on metabolites, so it cannot be used to develop reaction and metabolite-based distance measures. On the other hand, all envisaged distance measures can be developed based on a graph model.

Therefore, the necessary theory will be developed on graphs, namely **directed node-labeled graphs**. Reaction-based distances as well as neighborhood sensitive distances will be developed on reaction graphs, metabolite-based distances on metabolite graphs, and reaction and metabolite-based distances on bipartite graphs comprising both reactions and metabolites. The graphs are directed in order to capture information on reaction directionality and have labels assigned to their nodes in order to distinguish different reactions and metabolites.

### 3.4. Concept for Metabolic Network Comparison – Decision

Different concepts for developing distance measures were introduced in Section 2.4. The concept of graph isomorphisms has the disadvantage that network structure, like reaction neighborhood, cannot be explicitly taken into account. Moreover, maximum common subgraph type distance measures can be defined based on the concept of edit distances.

The benefit of the feature-based concept is that it can be used to construct a distance measure based on all sorts of abstract graph properties like number of nodes, number of edges, shortest path, etc. However, special features of individual nodes in the graph, like the number of outgoing and incoming edges, cannot be considered and thus this approach does not allow the assessment of whether the neighborhood of two identical nodes in two graphs is similar or not, which is important for the neighborhood sensitive distance measures.

The concept of edit distances allows for the definition of distance measures that take into account presence or absence of reactions and metabolites. Furthermore, distance measures that additionally consider network structure can be developed by defining edit costs based on the reaction neighborhood. Therefore, in this thesis the concept of **edit distances** is employed as the theoretical framework for developing distance measures on metabolic networks.

### 3.5. Clustering Methods – Decision

In this thesis cluster analysis is used to classify organisms based on the distances calculated between the organisms' metabolic networks. Organisms are considered similar if the distance between their networks is small. Therefore, clustering methods detecting compact clusters seem to be a good choice. Since distances will be calculated using the distance measures described above (see Section 3.1), methods that require only the

distance data as input rather than the original items (i.e. the metabolic networks) are needed.

For application in this thesis, hierarchical clustering methods are chosen. They generally require only the distance data on the objects (Jain and Dubes, 1988; Day and Edelsbrunner, 1984). In this thesis **average linkage agglomerative** and **complete linkage agglomerative** as well as the **Ward clustering technique** will be used. These established and widely-accepted approaches provide methods following different criteria like compactness and connectedness, which presumably makes them capable of detecting clusters of similar pathway variants.

Each of these clustering methods results in a clustering dendrogram which requires further processing in order to obtain a classification of the analyzed items. However, since the number of clusters is not known in advance, methods are needed to determine the most appropriate number of clusters. Preliminary studies have shown that the **cophenetic correlation coefficient** *cpcc* (see Section 2.5) performs well for this task and therefore it is relied upon in this thesis.

Since it is not clear from the type of data that is investigated which of the different clustering techniques will perform best, the decision is not yet made for a single clustering technique. As an alternative to relying on one single clustering method, classifications can be derived from several clustering dendrograms constructed by different clustering techniques and the *cpcc* between the cophenetic matrix of the partitioning and the original distance matrix can be used to determine the best partitioning. Whether one best clustering technique can be found or the *cpcc* will be used to determine the best classification will be decided in Section 6.1, where all clustering techniques will be evaluated on two test scenarios.





---

## Distance Measures

---

This chapter first provides the theoretical background on graphs. Then the distance measures that were informally introduced in Section 3.1 are formally defined.

### 4.1. Graph Theory

In this section a formal introduction to graphs, graph isomorphisms and edit distances on graphs is given. It mainly relies on Valiente (2002), Bunke (1997, 1999) as well as Fernández and Valiente (2001).

#### 4.1.1. Graphs and Subgraphs

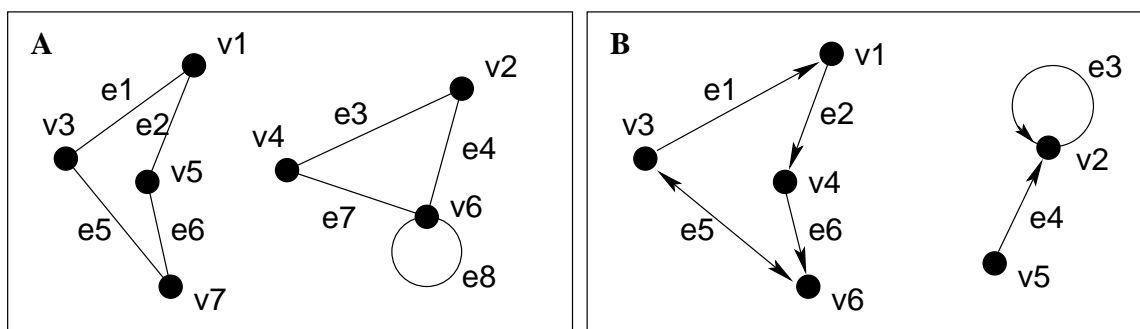
A graph is a structure consisting of vertices (or nodes) and arcs (or edges). An arc always connects two vertices.

**Definition 1.** *graph*

A graph  $G = (V, E)$  consists of a finite set of vertices  $V$  and a finite set of edges  $E \subseteq V \times V$ . If  $V = \emptyset$  then the graph  $G$  is called the empty graph. The graph is directed if the edge  $e_1 = (v_1, v_2)$  is to be distinguished from the edge  $e_2 = (v_2, v_1)$ ,  $e_1, e_2 \in E$ ,  $v_1, v_2 \in V$ , and undirected otherwise. The order of a graph  $G = (V, E)$ , denoted by  $n$ , is the number of vertices,  $n = |V|$ , the size, denoted by  $m$ , is the number of edges,  $m = |E|$ . In this thesis  $|G|$  stands for the order of the graph,  $|G| := n = |V|$ . An edge  $e = (v_1, v_2)$  is said to be incident with vertices  $v_1$  and  $v_2$ , where  $v_1$  is the source and  $v_2$  the target of edge  $e$ , and vertices  $v_1$  and  $v_2$  are said to be adjacent. Edges  $(v_1, v_2)$  and  $(v_2, v_3)$  are said to be adjacent, as are edges  $(v_1, v_2)$  and  $(v_3, v_2)$ , and  $(v_1, v_2)$  and  $(v_1, v_3)$ .

Graphs are often visualized as sets of points in the plane. Edges are drawn as lines connecting these points. Two examples are given in Figure 4.1.

In a *bipartite graph* the vertex set is partitioned into two subsets in a way such that every edge of the graph joins a vertex of one subset with a vertex of the other subset.



**Figure 4.1.:** A: Graphical depiction of an undirected graph with size 8 and order 7. B: For directed graphs the lines are substituted by arrows indicating the direction of the edge. A bidirectional edge, such as  $e5$ , is indicated by two arrowheads, one at each end of the line representing the edge.

**Definition 2.** *bipartite graph*

A graph  $G = (V, E)$  is said to be a bipartite graph if  $V$  can be partitioned into two subsets  $U, W \subseteq V$ ,  $U \cap W = \emptyset$  such that for all  $(v_1, v_2) \in E$ , either  $v_1 \in U$  and  $v_2 \in W$ , or  $v_1 \in W$  and  $v_2 \in U$ .

Labeled graphs have attributes or labels assigned to nodes and edges.

**Definition 3.** *labeled graph*

A labeled graph  $G$  is a quintuple  $G = (V, E, L, \alpha, \beta)$ , where  $V$  and  $E$  are the sets of nodes and edges, respectively,  $L$  is a set of labels,  $\alpha : V \rightarrow L$  is the node labeling function, and  $\beta : E \rightarrow L$ , the edge labeling function.  $\alpha(v)$  is called the label of vertex  $v \in V$ ,  $\beta(e)$  is called the label of edge  $e \in E$ .

**Definition 4.** *node labeled graph*

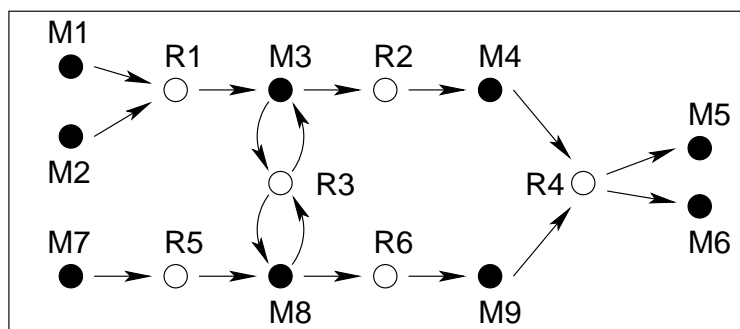
A node labeled graph  $G$  is a quadruple  $G = (V, E, L, \alpha)$ , where  $V$  and  $E$  are the sets of nodes and edges, respectively,  $L$  is a set of labels, and  $\alpha : V \rightarrow L$  is the node labeling function.  $\alpha(v)$  is called the label of vertex  $v \in V$ .

For modeling metabolic networks, two different types of graphs are employed depending on the distance measure that is to be calculated. Firstly, the concept of *bipartite directed node labeled graphs* is used for distance measures that take into account both reaction and metabolite nodes. Secondly, *directed node labeled graphs* are used to model a metabolic network if distance measures based on metabolites only or based on reactions only, or neighborhood sensitive distance measures are to be calculated. In all cases, nodes are assigned descriptive labels, namely reaction or metabolite identifiers, while edges are not assigned any labels.

A metabolic network can be modeled as metabolic network graph.

**Definition 5.** *metabolic network graph*

A metabolic network graph is a bipartite directed node-labeled graph and thus a quadruple  $G = (V, E, L, \alpha)$ , where  $V$  is a finite set of vertices and  $E$  a finite set of edges. The set of vertices  $V$  is partitioned into two subsets  $V_R$  and  $V_M$  constituting reactions and metabolites, respectively. Edges are directed to indicate reaction directionality and always connect either reactions to metabolites or vice versa. The label set  $L$  contains all reaction and metabolite identifiers.  $\alpha : V \rightarrow L$  is the node labeling function assigning a reaction identifier to the reaction nodes and a metabolite identifier to the metabolite nodes.



**Figure 4.2.:** Metabolic network modeled as bipartite directed node-labeled graph. Reaction identifiers start with an R followed by a unique number, whereas metabolite identifiers start with an M followed by a unique number.

An example is given in Figure 4.2.

If the focus is on reactions, a metabolic network can be modeled as reaction graph.

**Definition 6.** *reaction graph*

A reaction graph is a directed node-labeled graph and thus a quadruple  $G = (V, E, L, \alpha)$ , where  $V$  is a finite set of vertices representing metabolic reactions and  $E$  a finite set of edges. Edges are directed to indicate reaction directionality and connect reactions sharing an intermediate metabolite. The label set  $L$  contains all reaction identifiers, and  $\alpha : V \rightarrow L$  is the node labeling function assigning a unique reaction identifier to each node.

If the focus is on metabolites, a metabolic network can be modeled as metabolite graph.

**Definition 7.** *metabolite graph*

A metabolite graph is a directed node-labeled graph and thus a quadruple  $G = (V, E, L, \alpha)$ , where  $V$  is a finite set of vertices representing metabolites and  $E$  a finite set of edges. Edges represent the conversion of one metabolite into another metabolite by some reaction, and they are directed to indicate reaction directionality. The label set  $L$  contains all metabolite identifiers, and  $\alpha : V \rightarrow L$  is the node labeling function assigning a unique metabolite identifier to each node.

**Remark 1.**

1. Metabolic network graphs are special, because for each reaction node the stoichiometry of the reaction defines to which metabolite nodes it is connected via edges. These edges are not allowed to be altered individually, since this would correspond to altering the reaction stoichiometry, which is not possible. The same applies to reaction graphs.
2. Node labels are unique in metabolic network graphs, reaction graphs and metabolite graphs (the node labeling function is injective), since labels are used to distinguish individual reactions and metabolites.
3. Due to the above described stoichiometric constraints, edges in metabolic network graphs only describe which metabolites are connected to which reactions and whether

they act as substrates or products of the respective reactions or both. Therefore no edge labels are needed for further distinguishing the edges.

4. In reaction graphs, edges could be assigned the names of the intermediate metabolites. However, this is not done in this thesis, because this information is not used for comparing reaction graphs.
5. For the metabolite graphs it has to be decided whether they are to be used in the form they are defined originally or in a modified form, i.e. with all edges removed. The decision to be made here is whether two metabolites in different networks should be treated as identical if their labels are identical or only if additionally their connections to neighboring metabolites are identical. In the latter case a metabolite that is synthesized via a chain of certain intermediate metabolites in one metabolite graph is not mapped to a metabolite with identical label in another metabolite graph if that metabolite is synthesized via a chain of different metabolites. Since the aspect of similar chains of reactions is considered already in reaction neighborhood sensitive distance measures as well as in distance measures based on both reactions and metabolites, the decision is made not to take this information into account in distance measures based on metabolite alone. Therefore, from all metabolite graphs used in this thesis the edges will be removed. In particular, due to this decision the desired metabolite-based distance measures can be defined in the same way as those based on reactions and those based on reactions and metabolites, and the same proofs can be applied for showing the metric property.

### 4.1.2. Isomorphisms on Graphs

If there exists a *graph isomorphism* between two graphs then there exists a correspondence between these two graphs in terms of their structure and the labels in the case the graphs are labeled graphs.

**Definition 8.** *graph isomorphism*

Let  $G_1 = (V_1, E_1, L, \alpha_1)$  and  $G_2 = (V_2, E_2, L, \alpha_2)$  be two graphs.

A graph isomorphism between  $G_1$  and  $G_2$  is a bijective function  $f : V_1 \rightarrow V_2$  such that  $\alpha_1(v) = \alpha_2(f(v))$  for all  $v \in V_1$ , and  $(v_1, v_2) \in E_1 \Leftrightarrow (f(v_1), f(v_2)) \in E_2$  for all  $(v_1, v_2) \in E_1$ .

A *subgraph* of a graph is a graph whose node and edge sets are contained in the respective sets of the larger graph. The subgraph of a graph *induced* by a subset of its nodes has as edges the set of edges in the larger graph whose source and target belong to the subset of nodes.

**Definition 9.** *subgraph, supergraph*

Let  $G = (V, E)$  be a graph, and let  $V' \subseteq V$ . A graph  $G' = (V', E')$  is a subgraph of  $G$ ,  $G' \subseteq G$ , if  $E' \subseteq E$ . In this case  $G$  is called a supergraph of  $G'$ . The subgraph of  $G$  induced by  $V'$  is the graph  $(V', E \cap (V' \times V'))$ .

**Remark 2.** No matter, whether the metabolic network graph, the reaction graph or the metabolite graph without edges is used for modeling a metabolic network, metabolic

subnetworks of a given metabolic network always are induced subgraphs of the original graphs in the model. This is due to the special nature of metabolic networks: edges are bound to represent reaction stoichiometry and therefore are not allowed to be altered. If there are no edges at all, as in the metabolite graphs, the statement obviously is true as well.

An isomorphism might not exist between two given graphs, but between a subgraph of the one and a subgraph of the other graph. Then each subgraph is called a *common subgraph*.

**Definition 10.** *common subgraph*

Let  $G_1$  and  $G_2$  be two graphs and  $G'_1 \subseteq G_1$ ,  $G'_2 \subseteq G_2$  induced subgraphs of  $G_1$  and  $G_2$ , respectively.

If there exists a graph isomorphism between  $G'_1$  and  $G'_2$  then  $G'_1$  and  $G'_2$  are called common subgraphs of  $G_1$  and  $G_2$ .

The largest such common subgraph is called *maximum common subgraph*.

**Definition 11.** *maximum common subgraph*

Let  $G_1$  and  $G_2$  be two graphs.

A graph  $G$  is called a maximum common subgraph (mcs) of  $G_1$  and  $G_2$  if  $G$  is a common subgraph of  $G_1$  and  $G_2$  and there exists no other common subgraph  $G'$  of  $G_1$  and  $G_2$  such that  $|G'| > |G|$ .

The *common supergraph* of two graphs is a graph that contains two subgraphs such that the first subgraph is isomorphic to the first graph and the second subgraph is isomorphic to the second graph. In a sense, the common supergraph comprises both graphs. Of special interest is the smallest such supergraph, which is called *minimum common supergraph*.

**Definition 12.** *minimum common supergraph*

A graph  $G$  is a common supergraph of two graphs  $G_1$  and  $G_2$  if there exist subgraphs  $G'_1 \subseteq G$  and  $G'_2 \subseteq G$  such that  $G'_1$  is isomorphic to  $G_1$  and  $G'_2$  is isomorphic to  $G_2$ .

$G$  is a minimum common supergraph if there exists no other common supergraph  $G'$  of  $G_1$  and  $G_2$  such that  $|G'| < |G|$ .

Figure 4.3 shows two example graphs and their maximum common subgraph as well as their minimum common supergraph.

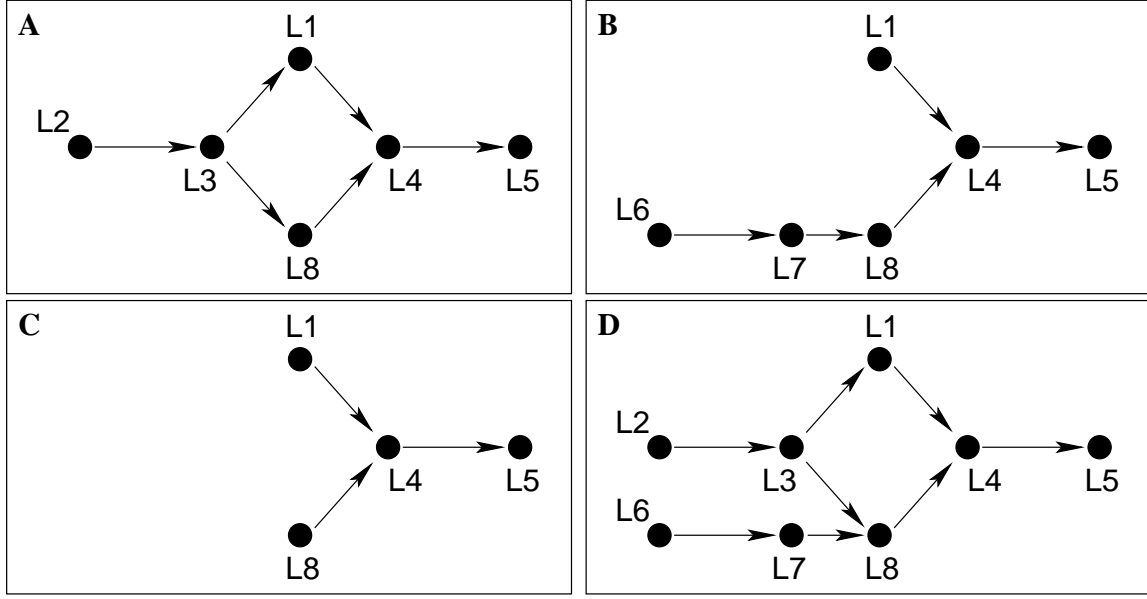
### 4.1.3. Graph Edit Distance

In this section the theory on graph edit distance is presented as far as needed. This section mainly relies on the theory as published by Bunke (1997, 1999).

**Definition 13.** *error-tolerant graph matching*

Let  $G_1 = (V_1, E_1, L, \alpha_1)$  and  $G_2 = (V_2, E_2, L, \alpha_2)$  be two graphs.

An error-tolerant graph matching (etgm) from  $G_1$  to  $G_2$  is a bijective function  $f: \hat{V}_1 \rightarrow \hat{V}_2$ , for some  $\hat{V}_1 \subseteq V_1$  and  $\hat{V}_2 \subseteq V_2$ . The graphs induced by  $\hat{V}_1$  and  $\hat{V}_2$  are named  $\hat{G}_1$  and  $\hat{G}_2$ , respectively.



**Figure 4.3.:** Two example graphs in A and B. C: maximum common subgraph of the graphs in A and B. D: minimum common supergraph of the graphs in A and B.

One says that node  $v \in \hat{V}_1$  is *substituted* by node  $v' \in \hat{V}_2$  if  $f(v) = v'$ . If  $\alpha_1(v) = \alpha_2(f(v))$  then the substitution is called *identical*, otherwise *non-identical*. Furthermore, one says that every node in  $V_1 \setminus \hat{V}_1$  is *deleted* from  $G_1$  under  $f$ , and every node in  $V_2 \setminus \hat{V}_2$  is *inserted* in  $G_2$  under  $f$ . Therewith every mapping  $f$  directly defines an edit operation for each node in  $G_1$  and  $G_2$ . Also,  $f$  indirectly defines an edit operation for each edge in  $G_1$  and  $G_2$ : If  $f(v) = v'$  and  $f(w) = w'$  and there exist edges  $e = (v, w) \in E_1$  and  $e' = (v', w') \in E_2$  then  $e$  is *substituted* by  $e'$  under  $f$ . If there exists only an edge  $e \in E_1$ , but no corresponding edge  $e' \in E_2$  then  $e$  is *deleted* under  $f$ . If there exists only an edge  $e' \in E_2$ , but no corresponding edge  $e \in E_1$  then  $e'$  is *inserted* under  $f$ . If a node  $v$  is deleted under  $f$  then all edges incident with  $v$  are deleted as well. Similarly, for every node  $v'$  inserted under  $f$  all edges incident with  $v'$  are inserted as well. Obviously, any etgm  $f$  can be understood as a set of edit operations (substitutions, deletions, and insertions of both nodes and edges) transforming a given graph  $G_1$  into another graph  $G_2$ .

The cost of an etgm is defined as follows:

**Definition 14.** *cost of an etgm*

The cost of an etgm  $f : \hat{V}_1 \rightarrow \hat{V}_2$  from a graph  $G_1 = (V_1, E_1, L, \alpha_1)$  to a graph  $G_2 = (V_2, E_2, L, \alpha_2)$  under a cost function  $c := (c_{ns}(v), c_{nd}(v), c_{ni}(v), c_{es}(e), c_{ed}(e), c_{ei}(e))$  is given by:

$$\begin{aligned} \gamma_c(f) = & \sum_{v \in \hat{V}_1} c_{ns}(v) + \sum_{v \in V_1 \setminus \hat{V}_1} c_{nd}(v) + \sum_{v \in V_2 \setminus \hat{V}_2} c_{ni}(v) \\ & + \sum_{e \in E_s} c_{es}(e) + \sum_{e \in E_d} c_{ed}(e) + \sum_{e \in E_i} c_{ei}(e), \end{aligned} \quad (4.1)$$

where  $c_{nd}(v)$ ,  $c_{ni}(v)$ ,  $c_{ns}(v)$  are the costs for deleting node  $v$ , inserting node  $v$  and substituting node  $v$  by node  $f(v)$ , respectively, and  $c_{ed}(e)$ ,  $c_{ei}(e)$ ,  $c_{es}(e)$  the costs for deleting,

inserting and substituting an edge  $e$ . All costs are nonnegative real numbers or infinite. The sets  $E_s$ ,  $E_d$ , and  $E_i$  are implicitly defined by the etgm  $f$  as follows: Let  $e_1 = (v_1, v_2)$ ,  $f(v_1) = v'_1$ ,  $f(v_2) = v'_2$ , and  $e'_1 = (v'_1, v'_2)$ . If edge  $e_1 \in E_1$  and  $e'_1 \notin E_2$  then  $e_1 \in E_d$ , if  $e_1 \notin E_1$ , and  $e'_1 \in E_2$  then  $e'_1 \in E_i$ , and if  $e_1 \in E_1$  and  $e'_1 \in E_2$  then  $e_1 \in E_s$ .

**Remark 3.**

1. Note that in this definition the cost of each edit operation depends on the particular node or edge it is applied to. In most distance measures in this thesis, the costs are the same for all edit operations. However, this feature in particular allows the definition of edit costs that depend on the neighborhood of the respective node in the graph, which is essential for defining the neighborhood sensitive reaction edit distance.
2. Bunke (1997, 1999) define graphs with labels on nodes as well as on edges and give their proofs for this type of graphs. Since in this thesis the graphs have labels assigned only to nodes, edge substitutions cannot be discriminated into identical and non-identical. Therefore, there exists only one type of edge substitutions. However, the proofs given in Bunke (1997, 1999) can still be applied, because one can introduce an artificial edge labeling function that assigns the same label to every edge. Also in this case there exists only one type of edge substitutions, namely identical edge substitutions.

**Definition 15.** *optimal etgm*

Let  $f$  be an etgm from a graph  $G_1$  to a graph  $G_2$  under a particular cost function  $c$ .

We call  $f$  an optimal etgm if there exists no other etgm  $f'$  from  $G_1$  to  $G_2$  with  $\gamma_c(f') < \gamma_c(f)$ .

**Definition 16.** *edit distance*

The cost of an optimal etgm from a graph  $G_1$  to a graph  $G_2$  under a given cost function  $c$  is called the edit distance of  $G_1$  and  $G_2$ , and is denoted by  $ed(G_1, G_2)$ :

$$ed(G_1, G_2) = \min\{\gamma_c(f) \mid f \text{ is an etgm from } G_1 \text{ to } G_2\}. \quad (4.2)$$

**Remark 4.**

1. For a given cost function there can be several optimal etgms from  $G_1$  to  $G_2$ .
2. The set of allowed edit operations is not fixed, but rather has to be adjusted to the particular problem domain. This can be achieved by appropriately defining the cost of edit operations. These costs considerably influence the minimal cost mapping between the two graphs. Setting the cost for a particular edit operation to infinity is an elegant way to prevent this operation from occurring in the minimal cost set of edit operations if there exists at least one such set with finite cost.

In order to prove the metric properties for the distance measures to be defined and for showing that certain edit distances correspond to mcs type distances, the following lemma is needed. It proposes that if the cost function meets certain conditions, there exists a correspondence between the matched (identically substituted) nodes of two graphs  $G_1$  and  $G_2$  and a maximum common subgraph of both.

**Lemma 1.**

Let  $G_1 = (V_1, E_1, L, \alpha_1)$  and  $G_2 = (V_2, E_2, L, \alpha_2)$  be two graphs. For a cost function  $c$  meeting the following conditions, the subgraph  $\hat{G}_1 \subseteq G_1$  induced by the set of identically substituted nodes  $\hat{V}_1$ , defined by the optimal etgm, is a maximum common subgraph of  $G_1$  and  $G_2$ , and  $\hat{G}_1$  is isomorphic to the graph  $\hat{G}_2 \subseteq G_2$ , induced by the set of identically substituted nodes  $\hat{V}_2$ . Furthermore, the resulting edit distance can be written as:

$$ed(G_1, G_2) = |V_1| + |V_2| - 2|V| \quad (4.3)$$

for any maximum common subgraph  $G = (V, E, L, \alpha)$  of  $G_1$  and  $G_2$ .

$$\begin{aligned} c_{ns}(v) &= \left\{ \begin{array}{ll} 0 & \text{for all identical substitutions} \\ \infty & \text{for all non-identical substitutions} \end{array} \right\} \text{ for any } v \in \hat{V}_1, \\ c_{nd}(v) &= 1 \text{ for any } v \in V_1 \setminus \hat{V}_1, \\ c_{ni}(v) &= 1 \text{ for any } v \in V_2 \setminus \hat{V}_2, \\ c_{\hat{e}s}(e) &= 0 \text{ for any } e \in \hat{E}_1, \\ c_{\hat{e}d}(e) &= \infty \text{ for any } e \in \hat{E}_1, \\ c_{\hat{e}i}(e) &= \infty \text{ for any } e \in \hat{E}_2, \\ c_{ed}(e) &= 0 \text{ for any } e \in E_1 \setminus \hat{E}_1, \\ c_{ei}(e) &= 0 \text{ for any } e \in E_2 \setminus \hat{E}_2, \end{aligned} \quad (4.4)$$

where  $\hat{E}_1 = E_1 \cap (\hat{V}_1 \times \hat{V}_1)$ ,  $\hat{E}_2 = E_2 \cap (\hat{V}_2 \times \hat{V}_2)$ , and  $c_{\hat{e}s}(e)$ ,  $c_{\hat{e}d}(e)$ ,  $c_{\hat{e}i}(e)$  are the costs for edit operations on edges  $e \in \hat{E}_1$  or  $e \in \hat{E}_2$ , respectively.

*Proof.* Introducing an artificial edge labeling function to both graphs  $G_1$  and  $G_2$  that assigns the same label to each edge and additionally demanding  $c_{es}(e) = \infty$  for all non-identical edge substitutions of any edge  $e \in \hat{E}_1$ , the conditions on the cost function (Eqs. 4.4) resemble those demanded in Bunke (1997). Therewith the proposition follows from Lemma 1, Lemma 2, and Theorem 1 in the same publication.  $\square$

As has been motivated above, distance measures in this thesis are constructed in a way that they satisfy the metric properties.

**Definition 17. metric**

A distance measure  $d$  on graphs is a metric if:

$$\begin{aligned} (1) \quad & d(G_1, G_2) \geq 0 \\ (2) \quad & d(G_1, G_2) = 0 \Leftrightarrow G_1 \text{ isomorphic to } G_2 \\ (3) \quad & d(G_1, G_2) = d(G_2, G_1) \\ (4) \quad & d(G_1, G_2) \leq d(G_1, G_3) + d(G_3, G_2), \end{aligned} \quad (4.5)$$

where  $G_1, G_2, G_3$  are graphs.



## 4.2. Distance Measures on Metabolic Networks

In this section, the distance measures that were discussed and informally introduced in Section 3.1 are defined formally and their metric properties are proven. Furthermore, whenever a correspondence exists between an edit distance-based measure and another distance measure from the literature, like an mcs type distance or Soergel type distance, this correspondence is shown as well. The section starts with summarizing general requirements on costs for edit operations that apply for all distance measures to be defined. Then the reaction and metabolite-based edit distances are defined followed by the reaction-based and the metabolite-based edit distances. Then the Soergel type edit distances and the mcs type edit distances are defined, each based on reactions only, metabolites only, as well as based on reactions and metabolites. Finally, the neighborhood sensitive reaction edit distances are defined.

### 4.2.1. Cost Function Requirements

Each edit distance is defined via its underlying cost function, which assigns weights to each edit operation. Thus, for developing an edit distance an appropriate cost function needs to be defined. Some specific aspects of metabolic networks need to be considered when defining this cost function. Also the metric property depends on the definition of the cost function. These considerations are equally important for all distance measures and are therefore discussed in the following, before the distance measures are defined.

In a metabolic network, identical node substitutions refer to the case that a node in one network is mapped to a node in the other network and that both nodes have the same label. Thus, identical node substitutions occur between the identical parts of the networks and are therefore assigned zero cost. Non-identical node substitutions match a reaction or metabolite in one of the networks with a reaction or metabolite that has a different label in the other network. This operation is undesirable, because different reactions and metabolites have different functions. However, the goal is to match as many reactions and metabolites with identical function as possible. Therefore the cost of non-identical node substitutions is set to infinity.

Edge substitutions do not alter the network (in analogy to node substitutions), thus they are assigned zero cost. Edge deletions and insertions of any edge  $e \in E_1 \cap (\hat{V}_1 \times \hat{V}_1)$  or  $e \in E_2 \cap (\hat{V}_2 \times \hat{V}_2)$  respectively, do not occur, because nodes in  $\hat{V}_1$  match nodes in  $\hat{V}_2$  and there are no non-identical node substitutions, and edges are bound to represent the reaction stoichiometry, which cannot be altered. The respective costs are set to infinity.

Edge deletions and insertions of any edge  $e \in E_1 \setminus (\hat{V}_1 \times \hat{V}_1)$  or  $e \in E_2 \setminus (\hat{V}_2 \times \hat{V}_2)$ , respectively, are not counted in this thesis, since an edge connects two nodes due to the respective reaction stoichiometry. Edit operations on these edges cannot be performed without editing the connected nodes. Therefore, if at all, they are supposed to be counted together with incident nodes, when these are deleted or inserted. Consequently, the cost for edge deletions and insertions is set to zero. Note that deleting or inserting a reaction node or metabolite node automatically triggers the deletion or insertion of all incident edges. Edit operations on edges obviously do not occur in metabolite graphs without edges. Therefore, costs for these edit operations may be assigned any value in the case these graphs are employed.

Furthermore, the cost for deleting a particular node has to be the same as for inserting

this node (symmetry). If this is not guaranteed, the resulting distance measure might not be symmetric and thus cannot be a metric.

These settings ensure that nodes (and thus also the edges) will be mapped to nodes having the same label if these exist, or otherwise be deleted or inserted, respectively. This already determines the etgm with minimal cost so that there is no need to search for the best mapping, which simplifies computation considerably.

### 4.2.2. Edit Distances on Reactions and Metabolites

For realizing the distance measures that count presence and absence of reactions and metabolites, the metabolic network graph model is employed and all node deletions and insertions are equally weighted by cost 1.

**Definition 18.** *reaction and metabolite-based cost function and edit distance*

The cost function  $c^{rm}$  for counting reactions and metabolites is defined as:

$$c^{rm} = (c_{ns_i}, c_{ns_n}, c_{nd}, c_{ni}, c_{\hat{e}s}, c_{\hat{e}d}, c_{\hat{e}i}, c_{ed}, c_{ei}) := (0, \infty, 1, 1, 0, \infty, \infty, 0, 0), \quad (4.6)$$

where  $c_{ns_i}$ ,  $c_{ns_n}$  are the costs for identical and non-identical node substitutions,  $c_{nd}$ ,  $c_{ni}$  the costs for node deletions and node insertions,  $c_{\hat{e}s}$ ,  $c_{\hat{e}d}$ , and  $c_{\hat{e}i}$  the costs for substituting, deleting, or inserting an edge  $e \in \hat{E}_1$  or  $e \in \hat{E}_2$ , and  $c_{ed}$ , and  $c_{ei}$  the corresponding values for edit operations on edges  $e \in E_1 \setminus \hat{E}_1$  or  $e \in E_2 \setminus \hat{E}_2$ .

Let  $G_1 = (V_1, E_1, L, \alpha_1)$  and  $G_2 = (V_2, E_2, L, \alpha_2)$  be two metabolic network graphs. According to Definitions 14 and 16, the edit distance based on  $c^{rm}$  is:

$$ed^{rm}(G_1, G_2) = \sum_{v \in V_1 \setminus \hat{V}_1} c_{nd}(v) + \sum_{v \in V_2 \setminus \hat{V}_2} c_{ni}(v) = |V_1 \setminus \hat{V}_1| + |V_2 \setminus \hat{V}_2|, \quad (4.7)$$

where  $\hat{V}_1$  and  $\hat{V}_2$  are the sets of identically substituted nodes.

This edit distance corresponds to a particular mcs type distance, because the sets of identically substituted nodes,  $\hat{V}_1$  and  $\hat{V}_2$ , are isomorphic to each other and are maximum common subgraphs of  $G_1$  and  $G_2$  (see Lemma 1).

**Corollary 1.**

For the cost function Equation 4.6 defined in Definition 18 the edit distance (Equation 4.7) can be written as:

$$ed^{rm}(G_1, G_2) = |V_1| + |V_2| - 2|\hat{V}_{12}|, \quad (4.8)$$

where  $\hat{V}_{12}$  stands for the node set of a maximum common subgraph  $\hat{G}_{12}$  of  $G_1$  and  $G_2$  that is isomorphic to both  $\hat{G}_1$  and  $\hat{G}_2$ , the graphs induced by the identically substituted nodes  $\hat{V}_1$  and  $\hat{V}_2$ , respectively.

*Proof.* The proposition follows from Equation 4.6 and Lemma 1. □

**Lemma 2.**

The reaction and metabolite-based edit distance  $ed^{rm}(G_1, G_2)$  introduced above (see Definition 18, Equation 4.7) is a metric.

*Proof.* Fernández and Valiente (2001) introduced a distance measure based on minimum common supergraph  $\hat{G}_{12}$  and maximum common subgraph  $\check{G}_{12}$  of two graphs  $G_1$  and  $G_2$ :  $d_{FV}(G_1, G_2) := |\hat{G}_{12}| - |\check{G}_{12}|$ , and proved that it is a metric (Theorem 17 in Fernández and Valiente (2001)). Although they define  $|G|$  as  $|G| := |V| + |E|$  for a graph  $G = (V, E)$ , their result can be shown to be valid for the alternative definition  $|G| = |V|$  using the very same proofs. Furthermore, it holds that  $d_{FV}(G_1, G_2) = |G_1| + |G_2| - 2|\hat{G}_{12}|$  (proof of Theorem 17 in Fernández and Valiente (2001)) and thus it follows  $ed^{rm}(G_1, G_2) = d_{FV}(G_1, G_2) = |\hat{G}_{12}| - |\check{G}_{12}|$ .  $\square$

For normalizing the edit distance  $ed^{rm}(G_1, G_2)$  (see Definition 18) an appropriate factor has to be found such that  $ed^{rm}(G_1, G_2) \in [0, 1]$  for any two metabolic network graphs  $G_1$  and  $G_2$ . Different choices are possible, like twice the number of nodes in the larger graph, sum of number of nodes of both graphs or number of nodes of the supergraph of both graphs, etc. The factor has to be chosen in a way that the metric property of the distance is maintained.

The maximum value that  $ed^{rm}(G_1, G_2)$  can take for any  $G_1 = (V_1, E_1)$  and  $G_2 = (V_2, E_2)$ , is  $|V_1| + |V_2|$ , which occurs if the maximum common subgraph of  $G_1$  and  $G_2$  is the empty graph. Dividing the distance by its maximum value ensures that it is always in the interval  $[0, 1]$ . The first idea for defining a normalized distance might then be:  $ed_{norm1}^{rm}(G_1, G_2) := (|V_1| + |V_2| - 2|\hat{V}_{12}|) / (|V_1| + |V_2|)$ , and another idea might be to define:  $ed_{norm2}^{rm}(G_1, G_2) := (|V_1| + |V_2| - 2|\hat{V}_{12}|) / 2 \max(|V_1|, |V_2|)$ . However, both do not satisfy the metric properties.

**Proposition 1.**

*The two distance measures*

$$ed_{norm1}^{rm}(G_1, G_2) := \frac{|V_1| + |V_2| - 2|\hat{V}_{12}|}{|V_1| + |V_2|}, \text{ and}$$

$$ed_{norm2}^{rm}(G_1, G_2) := \frac{|V_1| + |V_2| - 2|\hat{V}_{12}|}{2 \max(|V_1|, |V_2|)}$$

*are not metric.*

*Proof.* Both distance measures fail to satisfy the triangle inequality as can be seen with the following counter example:

Let  $G_1 = (\{v_1\}, \emptyset)$ ,  $G_2 = (\{v_2\}, \emptyset)$ , and  $G_3 = (\{v_1, v_2\}, \emptyset)$ . If the triangle inequality was valid, it would follow:

$$1 = \frac{|V_1| + |V_2| - 2|\hat{V}_{12}|}{|V_1| + |V_2|} \leq \frac{|V_1| + |V_3| - 2|\hat{V}_{13}|}{|V_1| + |V_3|} + \frac{|V_2| + |V_3| - 2|\hat{V}_{23}|}{|V_2| + |V_3|} = \frac{2}{3},$$

$$1 = \frac{|V_1| + |V_2| - 2|\hat{V}_{12}|}{2 \max(|V_1|, |V_2|)} \leq \frac{|V_1| + |V_3| - 2|\hat{V}_{13}|}{2 \max(|V_1|, |V_3|)} + \frac{|V_2| + |V_3| - 2|\hat{V}_{23}|}{2 \max(|V_2|, |V_3|)} = \frac{1}{2},$$

which both are contradictions.  $\square$

A normalized distance measure satisfying the properties of a metric can be defined as follows:

**Definition 19.** *normalized reaction and metabolite-based edit distance*

Let  $f$  be an etgm between two metabolic network graphs  $G_1 = (V_1, E_1, L, \alpha_1)$  and  $G_2 = (V_2, E_2, L, \alpha_2)$  under the cost function  $c^{rm}$  (see Definition 18), and  $\mathcal{G}$  be the set of all graphs  $G_i$  to be compared against each other in one analysis:  $\mathcal{G} := \{G_i, 1 \leq i \leq n\}$ . Then the normalized reaction and metabolite-based edit distance is defined as:

$$ed_{norm}^{rm}(G_1, G_2) := \frac{|V_1| + |V_2| - 2|\hat{V}_{12}|}{2 \max_{G \in \mathcal{G}}(|G|)}. \quad (4.9)$$

**Lemma 3.**

The normalized reaction and metabolite-based edit distance  $ed_{norm}^{rm}$  (see Definition 19) is a metric.

*Proof.* Criteria Equation 4.5 (1) to (3) are easily verified. For criterion (4) let  $G_1, G_2$ , and  $G_3$  be metabolic network graphs. It has to be shown that the following inequality holds:

$$ed_{norm}^{rm}(G_1, G_2) \leq ed_{norm}^{rm}(G_1, G_3) + ed_{norm}^{rm}(G_3, G_2).$$

This equation can be equivalently transformed into:

$$\frac{|V_1| + |V_2| - 2|\hat{V}_{12}|}{2 \max_{G \in \mathcal{G}}(|G|)} \leq \frac{|V_1| + |V_3| - 2|\hat{V}_{13}|}{2 \max_{G \in \mathcal{G}}(|G|)} + \frac{|V_3| + |V_2| - 2|\hat{V}_{23}|}{2 \max_{G \in \mathcal{G}}(|G|)},$$

which is the same as

$$|\hat{V}_{13}| + |\hat{V}_{23}| \leq |V_3| + |\hat{V}_{12}|.$$

The last inequality is true, because  $\hat{V}_{13}$  is the node set of an mcs of  $G_1$  and  $G_3$ ,  $\hat{V}_{23}$  the node set of an mcs of  $G_2$  and  $G_3$ , and  $\hat{V}_{12}$  the node set of an mcs of  $G_1$  and  $G_2$ . Therefore, it holds that  $\hat{V}_{13} \subseteq V_3$ , and  $\hat{V}_{23} \subseteq V_3$ . Moreover, either the intersection between  $\hat{V}_{13}$  and  $\hat{V}_{23}$  is empty, which is equivalent to  $|\hat{V}_{12}|$  being zero, or the intersection is not empty and it holds that  $\hat{V}_{13} \cap \hat{V}_{23} = \hat{V}_{12}$ . □

The distance  $ed_{norm}^{rm}(G_1, G_2)$  (see Definition 19) is based on both reaction and metabolite nodes of the metabolic network graphs  $G_1$  and  $G_2$ . For evaluating whether this distance measure performs better than a distance measure relying on either reactions alone or metabolites alone, the same type of distance measure is implemented in two further versions, which are based on reaction nodes alone and on metabolite nodes alone, respectively.

For defining an edit distance that is based only on the reaction content the reaction graph model (see Definition 6) is used. In principle, also here the metabolic network model could be used, however, it would not be possible to prove the metric properties, since Equation 4.5 (2) would not be satisfied.

**Definition 20.** *reaction-based cost function and edit distance*

The cost function  $c^r$  for counting reactions is defined as:

$$c^r = (c_{ns_i}, c_{ns_n}, c_{nd}, c_{ni}, c_{\hat{e}s}, c_{\hat{e}d}, c_{\hat{e}i}, c_{ed}, c_{ei}, ) := (0, \infty, 1, 1, 0, \infty, \infty, 0, 0). \quad (4.10)$$

Let  $G_1^r = (V_1^r, E_1^r, L, \alpha_1^r)$  and  $G_2^r = (V_2^r, E_2^r, L, \alpha_2^r)$  be two reaction graphs. The edit distance based on  $c^r$  is:

$$ed^r(G_1^r, G_2^r) = \sum_{v \in V_1^r \setminus \hat{V}_1^r} c_{nd}^r(v) + \sum_{v \in V_2^r \setminus \hat{V}_2^r} c_{ni}^r(v) = |V_1^r \setminus \hat{V}_1^r| + |V_2^r \setminus \hat{V}_2^r|, \quad (4.11)$$

where  $\hat{V}_1^r$  and  $\hat{V}_2^r$  are the sets of identically substituted reaction nodes.

**Corollary 2.**

The reaction edit distance in Definition 20 can be written as:

$$ed^r(G_1^r, G_2^r) = |V_1^r| + |V_2^r| - 2|\hat{V}_{12}^r|, \quad (4.12)$$

where  $\hat{V}_{12}^r$  stands for the node set of a maximum common subgraph  $\hat{G}_{12}^r$  between  $G_1^r$  and  $G_2^r$  that is isomorphic to both  $\hat{G}_1^r$  and  $\hat{G}_2^r$ , the graphs induced by the identically substituted nodes  $\hat{V}_1^r$  and  $\hat{V}_2^r$ , respectively.

*Proof.* The proposition follows from Equation 4.10 and Lemma 1. □

The normalized reaction edit distance is defined as follows:

**Definition 21.** *normalized reaction edit distance*

$$ed_{norm}^r(G_1^r, G_2^r) := \frac{|V_1^r| + |V_2^r| - 2|\hat{V}_{12}^r|}{2 \max_{G \in \mathcal{G}} (|G^r|)} \quad (4.13)$$

**Lemma 4.**

The normalized reaction edit distance (see Definition 21) is a metric.

*Proof.* The proof is identical to the one for Lemma 3, except that the metabolic network graphs are exchanged by reaction graphs. □

The metabolite-based cost function and distance measure is defined in analogy to the cost function and distance measure for the reaction-based case. However, this distance measure is defined on the model of metabolite graphs (see Definition 7) without edges.

**Definition 22.** *metabolite-based cost function and edit distance*

The cost function  $c^m$  for counting metabolites is defined as:

$$c^m = (c_{ns_i}, c_{ns_n}, c_{nd}, c_{ni}, c_{\hat{e}s}, c_{\hat{e}d}, c_{\hat{e}i}, c_{ed}, c_{ei}, ) := (0, \infty, 1, 1, 0, \infty, \infty, 0, 0). \quad (4.14)$$

Let  $G_1^m = (V_1^m, \emptyset, L, \alpha_1^m)$  and  $G_2^m = (V_2^m, \emptyset, L, \alpha_2^m)$  be two metabolite graphs without edges.

The edit distance based on  $c^m$  is:

$$ed^m(G_1^m, G_2^m) = \sum_{v \in V_1^m \setminus \hat{V}_1^m} c_{nd}^m(v) + \sum_{v \in V_2^m \setminus \hat{V}_2^m} c_{ni}^m(v) = |V_1^m \setminus \hat{V}_1^m| + |V_2^m \setminus \hat{V}_2^m|, \quad (4.15)$$

where  $\hat{V}_1^m$  and  $\hat{V}_2^m$  are the sets of identically substituted metabolite nodes.

**Corollary 3.**

The metabolite edit distance in Definition 22 can be written as:

$$ed^m(G_1^m, G_2^m) = |V_1^m| + |V_2^m| - 2|\hat{V}_{12}^m|, \quad (4.16)$$

where  $\hat{V}_{12}^m$  stands for the node set of a maximum common subgraph  $\hat{G}_{12}^m$  between  $G_1^m$  and  $G_2^m$  that is isomorphic to both  $\hat{G}_1^m$  and  $\hat{G}_2^m$ , the graphs induced by the identically substituted nodes  $\hat{V}_1^m$  and  $\hat{V}_2^m$ , respectively.

*Proof.* The proposition follows from Equation 4.14 and Lemma 1. □

The normalized edit distance based on metabolites is defined as follows:

**Definition 23.** *normalized metabolite edit distance*

$$ed_{norm}^m(G_1^m, G_2^m) := \frac{|V_1^m| + |V_2^m| - 2|\hat{V}_{12}^m|}{2 \max_{G \in \mathcal{G}}(|G^m|)} \quad (4.17)$$

**Lemma 5.**

The normalized metabolite edit distance (see Definition 23) is a metric.

*Proof.* The proof is identical to the one for Lemma 3, except that the metabolic network graphs are exchanged by metabolite graphs without edges. □

A distance measure that only differs in normalization from the ones defined above can be defined in a way that the resulting distance measure corresponds to the Soergel distance (Willett *et al.*, 1998).

**Definition 24.** *Soergel type reaction and metabolite-based edit distance*

Let  $c^{rm}$  be the cost function defined in Definition 18, and let  $G_1 = (V_1, E_1, L, \alpha_1)$  and  $G_2 = (V_2, E_2, L, \alpha_2)$  be two metabolic network graphs.

The Soergel type reaction and metabolite-based edit distance is defined as:

$$ed_S^{rm}(G_1, G_2) := \frac{|V_1| + |V_2| - 2|\hat{V}_{12}|}{|V_1| + |V_2| - |\hat{V}_{12}|}, \quad (4.18)$$

where  $\hat{V}_{12}$  is the node set of a maximum common subgraph  $\hat{G}_{12}$  of  $G_1$  and  $G_2$  that is isomorphic to both  $\hat{G}_1$  and  $\hat{G}_2$ , the graphs induced by the set of identically substituted nodes  $\hat{V}_1$  and  $\hat{V}_2$ , respectively.

**Lemma 6.**

The Soergel type reaction and metabolite-based edit distance (see Definition 24) is a metric.

*Proof.* It is shown that  $ed_S^{rm}$  corresponds to a Soergel type distance, which is a metric. Consider a metabolic network being a vector of reactions and metabolites  $X$ . The organism specific implementation of this pathway in organism  $i$  is denoted by the vector  $X_i = (x_{i1}, x_{i2}, \dots, x_{iN})$ , where  $x_{ij} = 1$  if reaction or metabolite  $j$  is present in this organism and  $x_{ij} = 0$  otherwise. The *Soergel distance* for dichotomous variables (which is

equivalent to the *Tanimoto (or Jaccard) coefficient* transformed into a distance (Willett *et al.*, 1998)) is defined as

$$d_T(X_i, X_j) := 1 - \frac{X_{ij}}{X_{ii} + X_{jj} - X_{ij}} = \frac{X_{ii} + X_{jj} - 2X_{ij}}{X_{ii} + X_{jj} - X_{ij}}, \quad (4.19)$$

where  $X_{ij} = X_i X_j$  is the scalar product between the two pathway vectors. Therewith  $X_{ii}$  is the number of reactions and metabolites present in the first organism,  $X_{jj}$  the number of reactions and metabolites present in the second one, and  $X_{ij}$  the number of reactions and metabolites both have in common. Späth (1980) gives a proof that the Tanimoto distance is a metric. Lipkus (1999) provides an alternative proof for the triangle inequality. Since the Soergel type edit distance  $ed_S^m$  is based on the cost function defined in Definition 18, Lemma 1 holds and thus  $\hat{G}_1$  is an mcs of  $G_1$  and  $G_2$ , and  $\hat{G}_1$  is isomorphic to  $\hat{G}_2$ , and for any mcs  $\hat{G}_{12}$  of  $G_1$  and  $G_2$ :  $|\hat{V}_1| = |\hat{V}_2| = |\hat{V}_{12}|$ . Furthermore,  $|\hat{V}_{12}|$  is the number of reactions and metabolites both graphs have in common,  $|V_1|$  and  $|V_2|$  the number of reactions and metabolites of the first and second graph, respectively, and therefore  $ed_S^m$  is equivalent to  $d_T$ . Thus, (1), (3), and (4) of the metric properties (Equation 4.5) are proven.

The metric property Equation 4.5 (2) has to be shown explicitly for the graph model: Let  $G_1$  be isomorphic to  $G_2$ . Then  $|V_1| = |V_2| = |\hat{V}_{12}|$  and thus  $ed_S^m(G_1, G_2) = 0$ . On the other hand,  $ed_S^m(G_1, G_2) = 0$  only if  $|V_1| + |V_2| - 2|\hat{V}_{12}| = 0$ , which can be achieved only by  $G_1$  being isomorphic to  $G_2$ , since always  $|\hat{V}_{12}| \leq \min(|V_1|, |V_2|)$ .

Thus,  $ed_S^m$  is a metric. □

This type of distance measure can also be defined to take into account reactions only or metabolites only:

**Definition 25.** *Soergel type reaction edit distance*

Let  $c^r$  be the cost function defined in Definition 20, and let  $G_1^r = (V_1^r, E_1^r, L, \alpha_1^r)$  and  $G_2^r = (V_2^r, E_2^r, L, \alpha_2^r)$  be two reaction graphs.

The Soergel type reaction edit distance is defined as:

$$ed_S^r(G_1^r, G_2^r) := \frac{|V_1^r| + |V_2^r| - 2|\hat{V}_{12}^r|}{|V_1^r| + |V_2^r| - |\hat{V}_{12}^r|}, \quad (4.20)$$

where  $\hat{V}_{12}^r$  is the node set of a maximum common subgraph  $\hat{G}_{12}^r$  of  $G_1^r$  and  $G_2^r$  that is isomorphic to both  $\hat{G}_1^r$  and  $\hat{G}_2^r$ , the graphs induced by the set of identically substituted nodes  $\hat{V}_1^r$  and  $\hat{V}_2^r$ , respectively.

**Definition 26.** *Soergel type metabolite edit distance*

Let  $c^m$  be the cost function defined in Definition 22, and let  $G_1^m = (V_1^m, E_1^m, L, \alpha_1^m)$  and  $G_2^m = (V_2^m, E_2^m, L, \alpha_2^m)$  be two metabolite graphs.

The Soergel type metabolite edit distance is defined as:

$$ed_S^m(G_1^m, G_2^m) := \frac{|V_1^m| + |V_2^m| - 2|\hat{V}_{12}^m|}{|V_1^m| + |V_2^m| - |\hat{V}_{12}^m|}, \quad (4.21)$$

where  $\hat{V}_{12}^m$  is the node set of a maximum common subgraph  $\hat{G}_{12}^m$  of  $G_1^m$  and  $G_2^m$  that is isomorphic to both  $\hat{G}_1^m$  and  $\hat{G}_2^m$ , the graphs induced by the set of identically substituted nodes  $\hat{V}_1^m$  and  $\hat{V}_2^m$ , respectively.

**Lemma 7.**

The Soergel type reaction edit distance as well as the Soergel type metabolite edit distance (see Definitions 25 and 26) are metrics.

*Proof.* The proof follows the same line as the one for the reaction and metabolite-based version of this distance (see Lemma 6). The only difference is that here the graphs are either reaction graphs or metabolite graphs without edges and therefore the vector  $X$  describing the metabolic network denotes a vector of reactions or metabolites, respectively. □

**Remark 5.** The difference between the above defined (standard) edit distances and the Soergel type edit distances is the normalization factor. Whereas for the latter only the order of the two graphs to be compared is taken into account, for the former the normalization factor is based on the order of all graphs that are to be compared in a given analysis. Therefore, the Soergel type edit distances weight deletions and insertions of nodes relative to the order of the two graphs being compared, whereas the (standard) edit distances weight all operations equally.

Whereas the above defined distance measures are based on the differences between the two networks to be compared, the following distances are based on what is common to both metabolic networks.

**Definition 27.** *mcs type reaction and metabolite-based edit distance*

Let  $c^m$  be the cost function defined in Definition 18, and let  $G_1$  and  $G_2$  be two metabolic network graphs.

The mcs type reaction and metabolite-based edit distance is defined as:

$$ed_{mcs}^m(G_1, G_2) := 1 - \frac{|\hat{V}_1|}{\max(|G_1|, |G_2|)}, \quad (4.22)$$

where  $\hat{V}_1 \subseteq V_1$  is the set of identically substituted nodes.

**Remark 6.** Although this distance is based on the same cost function and thus the same optimal etgm as, for example, the reaction and metabolite-based edit distance, there are no costs involved in the definition of this distance, since only those reactions and metabolites are considered that are identically substituted and the cost for any identical substitution is zero.

**Lemma 8.**

The mcs type reaction and metabolite-based edit distance (see Definition 27) is a metric and can be written as:

$$ed_{mcs}^m(G_1, G_2) := 1 - \frac{|\hat{G}_{12}|}{\max(|G_1|, |G_2|)}, \quad (4.23)$$

where  $\hat{G}_{12}$  is a maximum common subgraph of  $G_1$  and  $G_2$ .



*Proof.* From Lemma 1 it follows that the subgraphs  $\hat{G}_1$  and  $\hat{G}_2$  induced by the sets of identically substituted nodes  $\hat{V}_1$  and  $\hat{V}_2$ , respectively, are isomorphic to each other and that each is an mcs of  $G_1$  and  $G_2$ . Therefore, for any mcs  $\hat{G}_{12}$  of  $G_1$  and  $G_2$ :  $|\hat{G}_{12}| = |\hat{G}_1| = |\hat{G}_2|$ . Thus, this distance corresponds to the mcs type distance measure discussed in Bunke and Shearer (1998):

$$d_{BS}(G_1, G_2) := 1 - \frac{|mcs(G_1, G_2)|}{\max(|G_1|, |G_2|)} \quad (4.24)$$

Bunke and Shearer (1998) give a proof that this distance is a metric. In particular they show that  $0 \leq d_{BS}(G_1, G_2) \leq 1$ , for any graphs  $G_1, G_2$ , and thus this distance is already normalized.  $\square$

This distance measure can also be defined to take into account either reactions or metabolites only:

**Definition 28.** *mcs type reaction edit distance*

Let  $c^r$  be the cost function defined in Definition 20, and let  $G_1^r$  and  $G_2^r$  be two reaction graphs.

The mcs type reaction edit distance is defined as:

$$ed_{mcs}^r(G_1^r, G_2^r) := 1 - \frac{|\hat{V}_1^r|}{\max(|G_1^r|, |G_2^r|)}, \quad (4.25)$$

where  $\hat{V}_1^r \subseteq V_1^r$  is the set of identically substituted reaction nodes.

**Definition 29.** *mcs type metabolite edit distance*

Let  $c^m$  be the cost function defined in Definition 22, and let  $G_1^m$  and  $G_2^m$  be two metabolite graphs.

The mcs type metabolite edit distance is defined as:

$$ed_{mcs}^m(G_1^m, G_2^m) := 1 - \frac{|\hat{V}_1^m|}{\max(|G_1^m|, |G_2^m|)}, \quad (4.26)$$

where  $\hat{V}_1^m \subseteq V_1^m$  is the set of identically substituted metabolite nodes.

**Lemma 9.**

The mcs type reaction edit distance and the mcs type metabolite edit distance (see Definitions 28 and 29) are metrics and can be written as:

$$ed_{mcs}^r(G_1^r, G_2^r) := 1 - \frac{|\hat{G}_{12}^r|}{\max(|G_1^r|, |G_2^r|)}, \quad (4.27)$$

where  $\hat{G}_{12}^r$  is an mcs of  $G_1^r$  and  $G_2^r$ , and

$$ed_{mcs}^m(G_1^m, G_2^m) := 1 - \frac{|\hat{G}_{12}^m|}{\max(|G_1^m|, |G_2^m|)}, \quad (4.28)$$

respectively, where  $\hat{G}_{12}^m$  is an mcs of  $G_1^m$  and  $G_2^m$ .

*Proof.* The proofs are identical to the one given above for Lemma 8 with the only difference that the metabolic network graphs are exchanged by reaction graphs and metabolite graphs without edges, respectively.  $\square$

### 4.2.3. Neighborhood Sensitive Reaction Edit Distances

Based on the considerations in Section 3.1, the neighborhood sensitive reaction edit distance is defined to assign a non-zero, finite cost to edit operations only on reaction nodes, and the costs for each node are calculated based on its neighborhood in the network. The graph model employed for this distance is the reaction graph.

**Definition 30.** *reaction-based neighborhood sensitive cost function and edit distance*

$$c^{rns} = (c_{ns_i}, c_{ns_n}, c_{nd}, c_{ni}, c_{\hat{e}s}, c_{\hat{e}d}, c_{\hat{e}i}, c_{ed}, c_{ei}, ) := (0, \infty, score, score, 0, \infty, \infty, 0, 0), \quad (4.29)$$

where  $score : V \rightarrow \mathbb{R}$  denotes a function that assigns the individual edit cost to a reaction based on the amount of its synonymous and adjacent reactions. Three different scoring functions are evaluated against each other in this thesis:

$$score_1(v) = \exp(-syn(v) + \frac{1}{2} * adj(v)), \quad (4.30)$$

$$score_2(v) = \frac{1}{2^{syn(v)}} + \frac{1}{2} * adj(v), \quad (4.31)$$

$$score_3(v) = \max(1 - syn(v) + \frac{1}{2} * adj(v), \epsilon), \quad (4.32)$$

where  $syn(v)$  denotes the number of synonymous reactions,  $adj(v)$  the number of adjacent reactions for a reaction node  $v$ , and  $\epsilon > 0$  is a suitable small value, e.g.  $10^{-6}$ .

The resulting edit distance based on  $c^{rns}$  is defined as:

$$ed^{rns}(G_1^r, G_2^r) = \sum_{v \in V_1^r \setminus \hat{V}_1^r} score(v) + \sum_{v \in V_2^r \setminus \hat{V}_2^r} score(v), \quad (4.33)$$

where  $G_1^r$  and  $G_2^r$  are the two graphs to be compared,  $V_1^r$  and  $V_2^r$  are the node sets of the two graphs,  $\hat{V}_1^r$ ,  $\hat{V}_2^r$  are the sets of identical nodes, and  $score$  is one of  $score_1$ ,  $score_2$ ,  $score_3$ .

**Remark 7.**

- The scoring functions are defined in a way that the higher the number of synonymous reactions, the smaller is the resulting weight, and the higher the amount of adjacent reactions, the larger is the weight. The scoring functions differ in the way they weight synonymous and adjacent reactions relative to each other as well as in their gradient.
- The number of synonymous and adjacent reactions for a particular reaction is determined by analyzing the supergraph of all reaction graphs to be compared in a particular analysis.

The neighborhood sensitive reaction edit distance is a metric.

**Lemma 10.**

The neighborhood sensitive reaction edit distance  $ed^{rns}$  (see Definition 30) is a metric.

*Proof.* The four metric properties (see Definition 17, Equation 4.5) need to be shown for  $ed^{rns}$ . Equation 4.5 (1) and (3) are obvious. (2) is true because an edit distance of zero implies that all nodes are identically substituted. Since edges are bound to represent reaction stoichiometry, both graphs are isomorphic. Clearly, the edit distance equals zero if  $G_1$  is isomorphic to  $G_2$ .

Now the triangle inequality is proven (Equation 4.5 (4)). Let  $G_1 = (V_1, E_1, L, \alpha_1)$ ,  $G_2 = (V_2, E_2, L, \alpha_2)$ , and  $G_3 = (V_3, E_3, L, \alpha_3)$  be reaction graphs.  $ed^{rns}(G_1, G_2)$  is the cost of an optimal etgm from  $G_1$  to  $G_2$ . Due to the cost function (Equation 4.29) a node  $v_1 \in V_1$  or  $v_2 \in V_2$ , respectively will be either identically substituted, deleted or inserted under the optimal etgm. Since edges are bound to represent reaction stoichiometry,  $v_1 \in V_1$  is identically substituted by  $v_2 \in V_2$  if  $\alpha_1(v_1) = \alpha_2(v_2)$ . If either such a node  $v_1 \in V_1$  or  $v_2 \in V_2$  does not exist,  $v_1$  has to be deleted or  $v_2$  inserted, respectively.

Now it is shown that the edit cost between  $G_1$  and  $G_2$  is always less than the sum of edit costs between  $G_1$  and  $G_3$  and between  $G_3$  and  $G_2$ . Consider the following three cases:

(1)  $v_1 \in \hat{V}_1$  and therefore  $v_2 \in \hat{V}_2$  with  $\alpha_1(v_1) = \alpha_2(v_2)$  and thus zero edit cost. (1.1) If there exists  $v_3 \in V_3$  with  $\alpha_1(v_1) = \alpha_3(v_3)$ , it follows that  $\alpha_2(v_2) = \alpha_3(v_3)$  and thus there is zero edit cost. (1.2) If there exists no such  $v_3 \in V_3$ ,  $v_1$  is deleted and  $v_2$  inserted with costs greater than zero.

(2)  $v_1 \in V_1 \setminus \hat{V}_1$  and therefore no  $v_2 \in \hat{V}_2$  with  $\alpha_1(v_1) = \alpha_2(v_2)$ , so that  $v_1$  is deleted with some cost greater than zero. (2.1) If there exists  $v_3 \in V_3$  with  $\alpha_1(v_1) = \alpha_3(v_3)$ , there are no costs involved so far, but then  $v_3$  needs to be deleted, which involves the same costs as deleting  $v_1$ . (2.2) If there exists no such  $v_3 \in V_3$  then  $v_1$  is deleted which involves the same costs as above.

(3)  $v_2 \in V_2 \setminus \hat{V}_2$  and therefore no  $v_1 \in \hat{V}_1$  with  $\alpha_1(v_1) = \alpha_2(v_2)$ . (3.1) If there exists  $v_3 \in V_3$  with  $\alpha_2(v_2) = \alpha_3(v_3)$ , the edit cost here is zero, but then  $v_1$  needs to be inserted which involves an edit cost greater than zero. (3.2) If there exists no such  $v_3 \in V_3$  then  $v_2$  is inserted involving the same cost as above.

□

A normalized version of this distance measure can be defined as follows:

**Definition 31.** *normalized neighborhood sensitive reaction edit distance*

$$ed_{norm}^{rns}(G_1^r, G_2^r) := \frac{ed^{rns}(G_1^r, G_2^r)}{2 \max_{G \in \mathcal{G}}(|G|) \max_{v \in G^{super}} score(v)} = \frac{\sum_{v \in V_1^r \setminus \hat{V}_1^r} score(v) + \sum_{v \in V_2^r \setminus \hat{V}_2^r} score(v)}{2 \max_{G \in \mathcal{G}}(|G|) \max_{v \in G^{super}} score(v)}, \quad (4.34)$$

where  $G^{super}$  is the supergraph of all graphs  $G \in \mathcal{G}$  that are involved in the analysis.

**Lemma 11.**

The normalized neighborhood sensitive reaction edit distance  $ed_{norm}^{rns}$  (see Definition 31) is a metric.

*Proof.* Multiplying a metric by a scalar preserves the metric properties. Since  $\max_{G \in \mathcal{G}}(|G|) \max_{v \in G^{super}} score(v)$  does not change once the set of graphs  $\mathcal{G}$  to be compared is chosen, the proposition holds.

□



---

## Implementation: the CPA Web Server

---


The developed approach for comparative metabolic network analysis is implemented as perl scripts and accessible via a web frontend called CPA: Comparative Pathway Analyzer (Oehm *et al.*, 2008). In this chapter the functionality of the web server is demonstrated by means of a usecase.

### 5.1. Clustering Metabolic Pathway Data

From CPA's homepage the user can choose to follow three links. One leads to the clustering start page shown in Figure 5.1. The remaining two links can be followed to directly access the visualization facilities provided by CPA without prior comparative analysis of pathway variants in a set of organisms.

On the clustering start page the user can choose organisms and pathways to be analyzed. In addition to organism's annotation data from the KEGG database, users can upload their own reaction annotation data and submit it to the analysis. Currently, two file formats are supported: text files containing one KEGG reaction identifier (e.g. R00001) per line, and files in EMBL format (Kulikova *et al.*, 2007) containing EC number annotations. Users need to specify the format of the data they want to upload. EC numbers are translated into KEGG reactions numbers. Note that this transformation is ambiguous. In order to not miss any reaction possibly catalyzed by the respective enzyme specified by an EC number, all possible translations into KEGG reaction numbers are kept. The drawback is that the organism might be assigned more reactions than it is actually able to catalyze.

The comparative analysis can be performed for any pathway defined in the KEGG database and for the overall metabolic network constructed by merging reactions from all KEGG pathways into one single network. Moreover, users can define their own pathway by uploading a file containing a list of KEGG reaction identifiers. A pathway name as well as a pathway number have to be provided for internal reference. A generic




[CeBiTec SOFTWARE HOME](#) [CPA HOME](#) [HELP FOR THIS PAGE](#)

[CPA REACTION CONTENT VISUALIZER](#)

[CPA DIFFERENTIAL REACTION CONTENT VISUALIZER](#)

[CPA CLUSTERING](#)



## CPA - Clustering

If you want to include your own annotation data in your analysis upload this data first (below). Select organisms and pathways to cluster from the lists, using the 'Add' and 'Remove' buttons. (In case you uploaded your own data, you need to explicitly include it as well.) Then click 'Start clustering' to invoke calculations.

Currently there are 150 pathways and 890 organisms to choose from.


### Select organisms


Select from the list of organisms from the KEGG database and your uploads which ones to include in your clustering analysis. A maximum number of 30 organisms per analysis is allowed.

**Organisms**

**All organisms** | **Prokaryotes** | **Eukaryotes**

- Acaryochloris marina [amr]
- Acholeplasma laidlawii [acl]
- Acidiphilium cryptum JF-5 [acr]
- Acidithiobacillus ferrooxidans [afe]
- Acidobacteria bacterium [aba]
- Acidothermus cellulolyticus [ace]
- Acidovorax avenae [aav]
- Acidovorax sp. JS42 [ajs]
- Acinetobacter baumannii ACICU [abc]
- Acinetobacter baumannii ATCC 17978 [acb]
- Acinetobacter baumannii AYE [aby]
- Acinetobacter baumannii SDF [abm]

  
Add

  
Remove


**Selected organisms**


### Select pathways

Select from the list of pathways from the KEGG database which ones to include in your analysis. Pathways for which the KEGG pathway map does not contain any reaction numbers are not included here.

**Pathways**

- 1,1,1-Trichloro-2,2-bis(4-chlorophenyl)ethane (DDT) degradation [00351]
- 1,2-Dichloroethane degradation [00631]
- 1,4-Dichlorobenzene degradation [00627]
- 1- and 2-Methylnaphthalene degradation [00624]
- 2,4-Dichlorobenzoate degradation [00623]
- 3-Chloroacrylic acid degradation [00641]
- Acridone alkaloid biosynthesis [01058]
- Alanine and aspartate metabolism [00252]
- Alkaloid biosynthesis I [00950]
- Alkaloid biosynthesis II [00960]
- Aminoacyl-tRNA biosynthesis [00970]
- Aminophosphonate metabolism [00440]

  
Add

  
Remove

**Selected pathways**

[Start clustering](#)

### Upload your own data

Upload genome annotation data (EMBL format or list of KEGG reactions)  
You need to specify the file format!

filename	shortcut	file format	
<input type="text" value=""/> <a href="#">Browse...</a>	<input type="text" value="AAA"/>	<input type="radio"/> EMBL <input type="radio"/> KEGG reactions	<a href="#">Upload</a>

### Define your own pathway

Upload a data file defining your own pathway (list of KEGG reactions)

filename	pathway name	pathway number	
<input type="text" value=""/> <a href="#">Browse...</a>	<input type="text" value="User Pathway No. 1"/>	<input type="text" value="99980"/>	<a href="#">Upload</a>

**Figure 5.1.:** CPA clustering start page. On top of this web page organisms and pathways can be chosen for analysis. On the bottom of the page users can upload their own organism annotation data and define their own pathways to include them in the analysis.

pathway name and a generic pathway number are automatically generated and provided as default values.

As an example usecase for describing CPA's functionality five *Corynebacteria* are chosen to be analyzed, namely *C. diphtheriae* (KEGG abbreviation: *cdi*), *C. efficiens* (*cef*), *C. glutamicum* (*cgl*), *C. jeikeium* (*cjk*), and *C. urealyticum* (*cur*). These organisms have to be added to the list of *selected organisms* by selecting them from the list of *organisms* and confirming this choice with the *add* button. Then pathways have to be added that are to be analyzed. In the example usecase, the analyses are to be performed for all pathways except the overall reaction network, so these pathways are added to the list of *selected pathways* in the same way as explained for the organisms. If users want to analyze their own organisms or want to use their own pathway definition they have to upload the respective data before selecting organisms and pathways for analysis, since prior selections will not be kept. Uploaded organisms and pathways appear in the respective lists above the KEGG organisms and pathways and can be selected for analysis as explained above.

Once organisms and pathways are chosen, the user starts the analyses by invoking the *start clustering* button. A new page appears that informs the user about the progress. This page is automatically reloaded every 10 seconds until the analyses are finished.


## 5.2. Results Overview

When all calculations are done the user is directed to a new page that gives an overview of the results for all analyzed pathways (see Figure 5.2). For each pathway the results for each clustering method are summarized in a table. If results for several clustering methods are identical, only one table is displayed. Each cluster analysis results in a subdivision of the set of all analyzed organisms into subsets of organisms with similar pathway variants. For all possible pairs of these subgroups the differential reaction content (*drc*) is calculated and displayed. The *drc* can be subdivided into several classes:


- reactions occurring in all organisms of the first group and none of the second
- reactions occurring in some organisms of the first group and none of the second
- reactions occurring in all organisms of the first group and some of the second
- reactions occurring in all organisms of the second group and none of the first
- reactions occurring in some organisms of the second group and none of the first
- reactions occurring in all organisms of the second group and some of the first

Each row in the table contains from left to right the organisms in the first group of the pair, those in the second group, the total number of reactions in the analyzed pathway, the number of reactions all organisms in both groups have in common, as well as the values for the different classes of the *drc*. In the last column two links are provided that lead to a more detailed presentation of the *drc* in tabular format and a graphical visualization of the *drc* on pathway maps, respectively.

By default, the list of pathways is sorted according to the amount of mutually missing reactions (*ammr*) resulting from the analysis of each pathway, which is based on the *drc* as explained in the following. For each clustering technique and each pair of groups of



[CeBiTec SOFTWARE HOME](#)   [CPA HOME](#)   [HELP FOR THIS PAGE](#)  
[CPA REACTION CONTENT VISUALIZER](#)  
[CPA DIFFERENTIAL REACTION CONTENT VISUALIZER](#)  
[CPA CLUSTERING](#)



---

## Comparative Pathway Analyzer - Clustering Results Overview

For each pathway the clustering methods and corresponding clustering results are displayed. For each clustering method all automatically suggested clusters of organisms are paired and the differential reaction content is listed. Choose from the different visualization options provided in the columns to the right.

---

### 00071: Fatty acid metabolism

Average, Complete and Ward clustering results ([Show clustering dendrogram](#))

Automatically suggested groupings:

group 1	group 2	all reactions	all 1 all 2	all 1 no 2	some 1 no 2	all 2 no 1	some 2 no 1	
5 <input checked="" type="checkbox"/> cdi <input checked="" type="checkbox"/> cgl	<input checked="" type="checkbox"/> cef	47	10	0	0	12	0	<a href="#">View differential reaction content</a> <a href="#">CPA KEGG Visualizer</a>
5 <input checked="" type="checkbox"/> cdi <input checked="" type="checkbox"/> cgl	<input checked="" type="checkbox"/> cjk	47	9	1	0	23	0	<a href="#">View differential reaction content</a> <a href="#">CPA KEGG Visualizer</a>
5 <input checked="" type="checkbox"/> cdi <input checked="" type="checkbox"/> cgl	<input checked="" type="checkbox"/> cur	47	10	0	1	13	0	<a href="#">View differential reaction content</a> <a href="#">CPA KEGG Visualizer</a>
5 <input checked="" type="checkbox"/> cef	<input checked="" type="checkbox"/> cjk	47	24	1	0	11	0	<a href="#">View differential reaction content</a> <a href="#">CPA KEGG Visualizer</a>
5 <input checked="" type="checkbox"/> cef	<input checked="" type="checkbox"/> cur	47	17	8	0	8	0	<a href="#">View differential reaction content</a> <a href="#">CPA KEGG Visualizer</a>
5 <input checked="" type="checkbox"/> cjk	<input checked="" type="checkbox"/> cur	47	24	11	0	1	0	<a href="#">View differential reaction content</a> <a href="#">CPA KEGG Visualizer</a>

---

### 00860: Porphyrin and chlorophyll metabolism

Average, Complete and Ward clustering results ([Show clustering dendrogram](#))

Automatically suggested groupings:

group 1	group 2	all reactions	all 1 all 2	all 1 no 2	some 1 no 2	all 2 no 1	some 2 no 1	
5 <input checked="" type="checkbox"/> cdi	<input checked="" type="checkbox"/> cef <input checked="" type="checkbox"/> cgl	96	20	15	0	1	1	<a href="#">View differential reaction content</a> <a href="#">CPA KEGG Visualizer</a>

**Figure 5.2.:** The clustering results overview page lists all pathways that were analyzed and for each of them provides a summary of the clustering results for each clustering method. The groups resulting from the automatic classification of organisms are compared pairwise. Each line in the presented table displays the drc of the organisms in group1 versus those in group2. Columns: all reactions: total number of reactions in the pathway, all 1 all 2: number of reactions all organisms in both groups have in common, all 1 no 2: number of reactions annotated for all organisms in group1 and for none in group2, some 1 no 2: number of reactions annotated for some, but not all, organisms in group1 and for none in group2, all 2 no 1 and some 2 no 1 are defined analogously. The last column provides two links leading to a tabular view of the drc and a graphical visualization of the drc on KEGG pathway maps, respectively.



organisms for a particular pathway the number of reactions occurring in all organisms of the first group while missing in all organisms of the second group and the number of reactions occurring in all organisms of the second group while missing in all organisms of the first group are summed up. If there exists only one group, the value for ammr is set to zero. The maximum of these values over all pairs and all clustering techniques is the maximum ammr for this pathway. By sorting according to the maximum ammr pathways with a large number of differences in reaction content appear on top of the list.

Pathways in Figure 5.2 are sorted according to this strategy. The pathway with highest maximum ammr for the organisms in the usecase is fatty acid metabolism (KEGG pathway number 00071). The maximum ammr for this pathway and the analyzed *Corynebacteria* is realized for the first group consisting of organisms *cdi* and *cgl* and the second group consisting of *cjk* (see Figure 5.2, second row). The list also depicts the composition of the automatically detected groups: only *cdi* and *cgl* are grouped together, *cef*, *cjk*, and *cur* are put into singleton groups. The next pathway in the list is porphyrin and chlorophyll metabolism (KEGG pathway number 00860).

The list of pathways can also be sorted according to the relative maximum ammr, which is the maximum ammr divided by the size of the respective pathway, where the size of a pathway is the number of constituting reactions. This means that it does not matter for the position of a pathway in the sorted result list whether half of the reactions of a huge pathway are missing or half of the reactions of a small pathway, both will appear close to each other in the list.

The list of pathways can also be filtered. The web page allows the user to specify a set of organisms and to update the list, so that only pathways are displayed for which these organisms were classified into the same group. This simplifies, for example, searching for pathways for which all pathogens or all organisms with the same habitat are put together. Additionally, the user can specify whether other than the specified organisms are allowed to be in the group or not. The filtered list can be sorted either according to the absolute ammr or according to the relative ammr as explained above.

## 5.3. Single Pathway Clustering Results

For each pathway in the results overview list a web link called *show clustering dendrograms* leads to another web page providing more detailed information on the clustering results (see Figure 5.3). On this page, the clustering dendrograms are displayed for each clustering method. A red line in the dendrogram indicates where the automated method suggested to cut the dendrogram for obtaining the classification.

Below each dendrogram a group of checkboxes allows the user to manually select organisms and thus define two groups to be compared. Buttons are provided to invoke displaying the drc of these two groups either on KEGG pathway maps (*visualize on pathway map*) or in tabular format (*view differential reaction content*). Additionally, all possible pairings of the automatically derived groups are displayed, and buttons are provided for invoking visualizations.

### Select pathway

Choose one pathway from the list of analyzed pathways and press 'Display results'.

Fatty acid metabolism [00071]

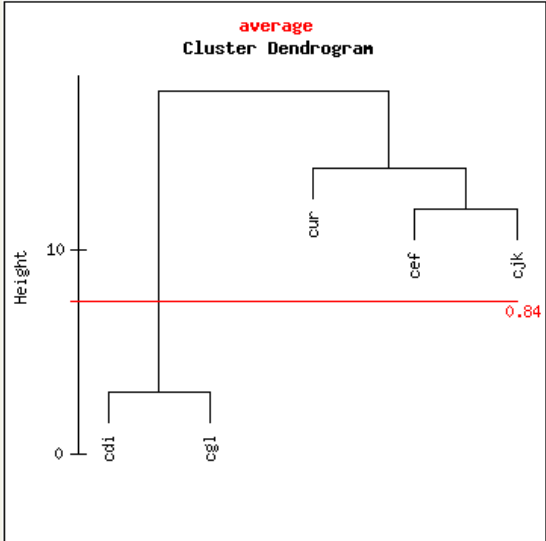
---

Go to: [top](#), [average](#), [complete](#), [ward](#)

---

### Average clustering results

Move the mouse pointer over an organism's abbreviation to display its proper name.



Make your own choice based on the dendrogram:

<b>group 1</b>	<input type="checkbox"/> cdi <input type="checkbox"/> cef <input type="checkbox"/> cgl <input type="checkbox"/> cjk <input type="checkbox"/> cur	<input type="button" value="Visualize on pathway map"/>
<b>group 2</b>	<input type="checkbox"/> cdi <input type="checkbox"/> cef <input type="checkbox"/> cgl <input type="checkbox"/> cjk <input type="checkbox"/> cur	<input type="button" value="check complement"/> <input type="button" value="View differential reaction content"/>

Automatically suggested groupings:

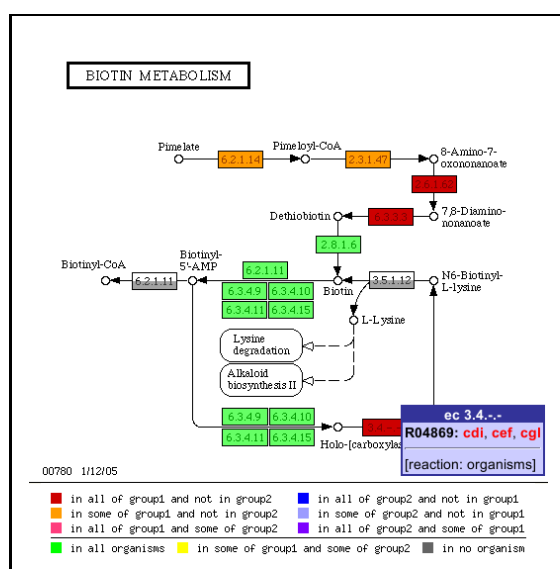
group 1	group 2	
<input checked="" type="checkbox"/> cdi <input checked="" type="checkbox"/> cgl	<input checked="" type="checkbox"/> cef	<input type="button" value="Visualize on pathway map"/>
<input checked="" type="checkbox"/> cdi <input checked="" type="checkbox"/> cgl	<input checked="" type="checkbox"/> cjk	<input type="button" value="Visualize on pathway map"/>
<input checked="" type="checkbox"/> cdi <input checked="" type="checkbox"/> cgl	<input checked="" type="checkbox"/> cur	<input type="button" value="Visualize on pathway map"/>
<input checked="" type="checkbox"/> cef	<input checked="" type="checkbox"/> cjk	<input type="button" value="Visualize on pathway map"/>
<input checked="" type="checkbox"/> cef	<input checked="" type="checkbox"/> cur	<input type="button" value="Visualize on pathway map"/>
<input checked="" type="checkbox"/> cjk	<input checked="" type="checkbox"/> cur	<input type="button" value="Visualize on pathway map"/>

**Figure 5.3.:** For one pathway at a time, detailed results can be displayed on a separate web page. For each clustering method the clustering dendrogram is provided. Groups of checkboxes for manual selection as well as predefined groups are provided for invoking visualizations of the respective drc. The red line in the dendrogram indicates where it was cut for deducing the classification. The value in red next to this line depicts the respective value of the *cpcc*.

## 5.4. Displaying Differential Reaction Content

Two visualization options exist for displaying the drc. The first is a tabular view, which provides a table for each class of the drc that for all organisms lists the presence or absence of all reactions. For completeness sake two additional tables are provided: the first lists all reactions occurring in all analyzed organisms, while the second lists all reactions that are not annotated for any of the analyzed organisms.

The second option is a visualization of the drc for a particular pathway and two sets of organisms on the respective KEGG pathway map. Each class of the drc is assigned a unique color, and each reaction on the map is colored according to which class it belongs to (see Figure 5.4).



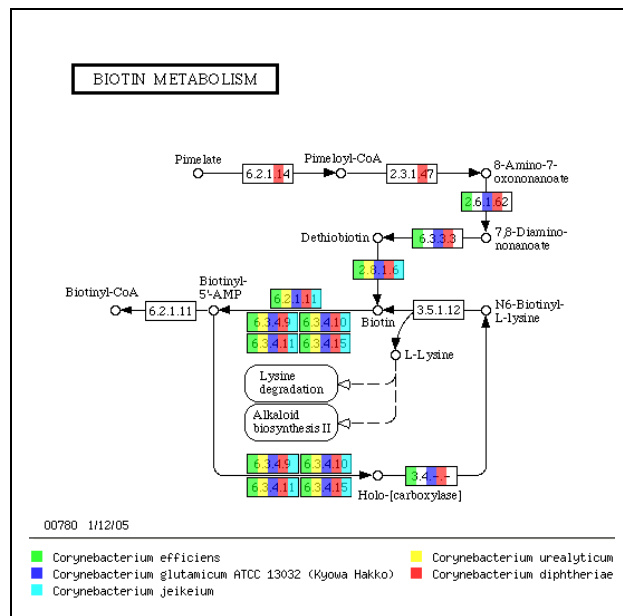
**Figure 5.4.:** Visualization of the drc on the KEGG biotin metabolism for the five *Corynebacteria* *C. diphtheriae* (cdi), *C. efficiens* (cef), and *C. glutamicum* (cgl) in the first set compared against *C. jeikeium* (cjk) and *C. urealyticum* (cur) in the second set. For each box containing an EC number, a tooltip lists all KEGG reactions associated with the respective EC number, and all organisms this reaction is annotated for. Red organisms belong to set one, blue organisms to set two (tooltip). See the legend attached to the image for the color code of the enzyme boxes. If more than one reaction corresponds to a particular EC number the respective box is subdivided into several parts, each representing one reaction and colored accordingly. Boxes whose lower half is colored in grey indicate that none of the corresponding reactions are annotated in any of the analyzed organisms.

If a user-defined pathway was subjected to analysis there does not exist a KEGG pathway on which the drc could be displayed. In this case, a pathway layout is automatically generated using the software graphviz (Gansner and North, 2000), and the drc is visualized on this pathway layout.

In the case this page is invoked from one of the clustering results pages (results overview or single pathway clustering results), a pathway and a set of organisms are already selected and the respective pathway map is displayed automatically. In the case a user came to this page via the link on the CPA start page, lists of organisms and pathways are provided for selection.

## 5.5. Simultaneously Displaying Reaction Content of Several Organisms

In addition to the visualizations that focus on displaying the drc, CPA provides yet another visualization for the reaction content of several organisms. The respective web page can be reached via the third link on the CPA start page. Very much like the visualizations of the KEGG web interface, it permits the display of the reaction content of organisms on KEGG's metabolic pathway maps by coloring the respective enzyme boxes. However, in difference to the KEGG website, here the user can choose up to six organisms, for which the CPA web server simultaneously displays their reaction content using a user-specified color for each organism (see Figure 5.5). This visualization also allows to easily assess the drc. However, in contrast to the visualization of the drc on pathway maps, the number of organisms that can be displayed simultaneously is limited.



**Figure 5.5.:** Reaction content visualization on the KEGG biotin metabolism for five *Corynebacteria* simultaneously. The *Corynebacteria* are *C. diphtheriae* (*cdi*), *C. efficiens* (*cef*), *C. glutamicum* (*cgl*), *C. jeikeium* (*cjk*), and *C. urealyticum*. EC boxes are subdivided vertically into several parts, each representing one organism. These parts are colored if the respective reaction is annotated for the respective organism or left blank otherwise.

This chapter is subdivided into two parts. In the first part, the suitability of the developed methodology for comparing metabolic networks is evaluated. Different combinations of a distance measure on the one hand and a clustering technique on the other hand are compared to each other on two test scenarios in order to find the best suited combination. In the second part, the chosen distance measure and clustering techniques are applied for the comparative analysis of five members of the genus *Corynebacterium*.

### 6.1. Comparison of Different Distance Measures and Clustering Techniques

Several different distance measures were defined in Section 4.2 (see Table 6.1 for a complete list). Here, clustering dendrograms and automatic classifications of organisms based on their metabolic pathway variants are compared to each other. Each dendrogram results from a combination of a distance measure and a clustering technique. The clustering techniques are average and complete linkage agglomerative clustering as well as Ward clustering, which in the following will be referred to as *average*, *complete*, and *ward*, respectively. The two goals are on the one hand to validate the developed metabolic pathway comparison approach, and on the other hand to find those distance measures and clustering techniques that are best suited for comparing these networks.

Validation is usually done by comparing the results with a standard of truth. For the approach introduced in this thesis, the standard of truth would be selected pathways and sets of organisms and a verified grouping of these organisms according to their pathway variants. Since such data does not exist, it has to be created artificially. In the following, two test scenarios, comprising a pathway and a set of organisms each, are manually analyzed in order to derive a classification of the organisms.

The first test case consists of a set of manually designed pseudo-organisms and an artificial pathway, whereas the second test case consists of a set of existing organisms

**Table 6.1.:** Distance measures being evaluated. The table lists for each distance measure the abbreviation used as reference in the text, a brief description of the distance measure, and a reference to the mathematical definition in Section 4.2.  $s$ : number of synonymous reactions,  $a$ : number of adjacent reactions, mcs: maximum common subgraph, nsred: neighborhood sensitive reaction edit distance.

abbreviation	distance measure description	reference
<i>m1</i>	normalized reaction edit distance	Eq. 4.13
<i>m2</i>	mcs type edit distance based on reactions	Eq. 4.27
<i>m3</i>	Soergel type edit distance based on reactions	Eq. 4.20
<i>m4</i>	normalized reaction and metabolite edit distance	Eq. 4.9
<i>m5</i>	mcs type edit distance based on reactions and metabolites	Eq. 4.23
<i>m6</i>	Soergel type edit distance based on reactions and metabolites	Eq. 4.18
<i>m7</i>	normalized metabolite edit distance	Eq. 4.17
<i>m8</i>	mcs type edit distance based on metabolites	Eq. 4.28
<i>m9</i>	Soergel type edit distance based on metabolites	Eq. 4.21
<i>m10</i>	nsred, <i>reaction edit cost</i> = $e^{-s+\frac{a}{2}}$	Eqs. 4.34 and 4.30
<i>m11</i>	nsred, <i>reaction edit cost</i> = $\frac{1}{2^s} + \frac{a}{2}$	Eqs. 4.34 and 4.31
<i>m12</i>	nsred, <i>reaction edit cost</i> = $\max(1 - s + \frac{a}{2}, 10^{-6})$	Eqs. 4.34 and 4.32

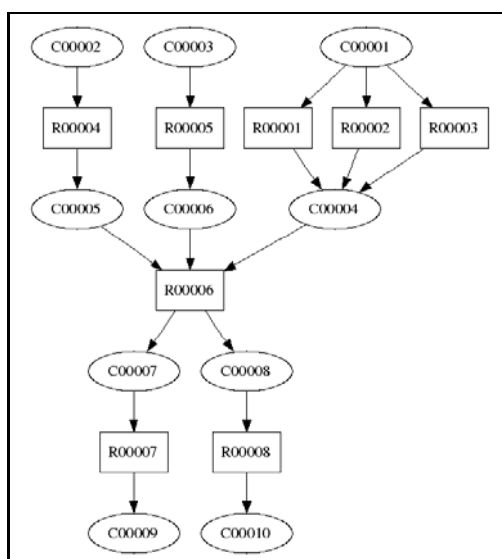
and a subpathway of the lysine biosynthesis pathway defined in the KEGG database. The first test case is constructed in order to test whether distance measures and clustering techniques in principle work as expected on metabolic networks, whereas in the second test case the performance on real world data is assessed. Each test case will be analyzed and discussed on its own, before, based on both test cases, the choice is made which combination of distance measure and clustering technique to use for subsequent applications.

### 6.1.1. Artificial Test Scenario

This test scenario is constructed in order to test whether distance measures and clustering techniques work in a way that would theoretically be expected from their design. The artificial pathway is constructed to contain on the one hand a set of synonymous reactions (reactions converting the same substrates into the same products) and on the other hand a reaction with many adjacent reactions, i.e. a reaction that is connected (via metabolites) to many other reactions. Organisms are designed to implement various parts of the pathway. Some organisms only implement one of the synonymous reactions and two organisms are missing the highly connected reaction. The principles for designing the pathway and the organisms are as follows: Firstly, it should be possible to unambiguously group the organisms manually for the pathway with respect to the following criteria: reaction content, reaction and metabolite content, metabolite content, and functionality. Secondly, the resulting groupings should be distinct for the different distance measures. For example,  $E$  is close to  $C$ , and  $D$  is close to  $G$  if the reaction content is the criterion, but  $E$  is close to  $D$  if the focus is on functionality.

A diagram of the constructed test pathway is shown in Figure 6.1. The reaction content of the pseudo-organisms is shown in Table 6.2 and the metabolite content in Table 6.3. Note that the metabolite content of an organism depends on the reaction content as follows: a particular metabolite is said to be present in an organism, if a reaction is present that either consumes or produces this metabolite. Thus, it is possible that, like

in this example, the metabolite content is almost identical for all organisms, although the reaction content is not. Moreover, if one reaction is missing in an organism, reactions that directly or indirectly (via other intermediate reactions) depend on the metabolites produced by the missing reaction and thus theoretically cannot be active, because their substrate metabolites are missing, are not removed from the network of this organism. This is handled in this way in order to be more error-tolerant: if the lack of the reaction in some organism is due to a missing or erroneous annotation, the effect of this error would be increased if other reactions would be removed from the network, which is not desired.



**Figure 6.1.:** Graphical representation of the artificial test pathway comprising reactions R00001 to R00008 and metabolites C00001 to C00010. Reactions R00001, R00002, and R00003 are synonymous reactions. R00006 is a highly connected reaction, i.e. it has many adjacent reactions connected via substrate and product metabolites.

**Table 6.2.:** Presence (x) and absence (-) of reactions in analyzed pseudo-organisms for the artificial test pathway. Pseudo-organisms are named A, B, C, D, E, F, G.

reaction	pseudo-organisms						
	A	B	C	D	E	F	G
R00001	x	-	-	x	-	x	x
R00002	-	x	-	x	-	x	x
R00003	-	-	x	x	x	x	x
R00004	x	x	x	x	x	-	x
R00005	x	x	x	x	x	x	x
R00006	x	x	x	-	-	x	x
R00007	x	x	x	x	x	x	x
R00008	x	x	x	x	x	x	x

**Table 6.3.:** Presence (x) and absence (-) of metabolites in analyzed pseudo-organisms for the artificial test pathway. Pseudo-organisms are named A, B, C, D, E, F, G.

metabolite	pseudo-organisms						
	A	B	C	D	E	F	G
C00001	x	x	x	x	x	x	x
C00002	x	x	x	x	x	-	x
C00003	x	x	x	x	x	x	x
C00004	x	x	x	x	x	x	x
C00005	x	x	x	x	x	x	x
C00006	x	x	x	x	x	x	x
C00007	x	x	x	x	x	x	x
C00008	x	x	x	x	x	x	x
C00009	x	x	x	x	x	x	x
C00010	x	x	x	x	x	x	x

**Table 6.4.:** Manually derived classifications of the artificial organisms for the artificial test pathway according to different criteria.

classification criteria	manual classification			
reaction content-based	CE	DFG	A	B
reaction and metabolite content-based	CE	DG	F	A B
metabolite content-based	ABCDEG	F		
functionality-based	ABCG	DE	F	

### 6.1.1.1. Manual Classification of Organisms

As indicated above, the manual classification of organisms is done according to four different criteria: based on the reaction content of the organisms, based on both reaction and metabolite content, based on the metabolite content, as well as according to functional aspects.

If the reaction content is taken as a basis, the pseudo-organisms can be classified into three groups. *E* and *C* differ by only a single reaction and each of them by two or more reactions from all other pseudo-organisms. Therefore, these two make up one group. *D* differs from *G* by a single reaction and by two or more from all others. *F* also differs from *G* by only one reaction and by more than one from all others. *D* differs from *F* by two reactions. *G* differs from all pseudo-organisms except *D* and *F* by at least two reactions. Therefore *D*, *G*, and *F* make up the second group. The closest neighbor for both *A* and *B* is *G* with a difference of two reactions. However, they differ from *D* and *F* by three reactions and *D* and *F* are in a group together with *G*. *A* and *B* differ from each other by three reactions. *A* and *B* both differ from *C* and *E* by three and four reactions, respectively. Thus *A* and *B* can be put into each of the existing groups or left as singletons.

Although the reaction content differs across the pseudo-organisms, the metabolite content is identical, except for pseudo-organism *F*, which is lacking metabolite C00002. The reason is that almost all metabolites adjacent to any missing reaction are also involved in other reactions which are still present. Hence, the absence of one or two of the three reactions R00001, R00002, and R00003 does not change the metabolite content in the pseudo-organisms and neither does the absence of reaction R00006, since the neighboring reactions are still present. Only the lack of reaction R00004 in *F* makes a difference, since in this case metabolite C00002 is not part of any reaction any longer and therefore removed from the network. Thus, *A*, *B*, *C*, *D*, *E*, and *G* make up one group and *F* another.

When taking into account reactions and metabolites, there is almost no change to the reaction-based grouping, since only the distance from *F* to all other pseudo-organisms is increased by one. Thus, *F* is still closest to *D* and *G*. However, due to a difference of three reactions and metabolites to *D*, *F* is put into a singleton group.

Functionally, the pseudo-organisms can be grouped as follows. *A*, *B*, *C*, and *G* have the same functionality, because reactions R00001, R00002, and R00003 all convert C00001 into C00004. They make up one group. Organisms *D* and *E* make up another group, because they both lack reaction R00006, and the difference in R00001 and R00002 does not matter functionally. *F* is put into a singleton group, since it is the only pseudo-organism lacking reaction R00004.



**Table 6.5.:** Automatically derived classifications of the artificial organisms for the artificial test pathway and different distance measures (abbreviations as defined in Table 6.1) and clustering techniques (*average*, *complete*: average, respectively complete linkage agglomerative clustering method; *ward*: Ward clustering method). Classification results are identical for all three clustering techniques, if no clustering method is specified. For each classification the respective *cpcc* is listed. Organisms are colored according to the group they were manually classified into (see Table 6.4).

distance measure type	distance measure and clustering technique	<i>cpcc</i>	automatic classification			
reaction-based	<i>m1 average, ward</i>	0.83	C,E	D,G,F	A	B
	<i>m1 complete</i>	0.82	C,E	D,G	F	A B
	<i>m2</i>	0.81	A,B,C	D,G,F	E	
	<i>m3</i>	0.83	C,E	D,G,F	A	B
reaction and metabolite-based	<i>m4</i>	0.80	C,E	D,G	F	A B
	<i>m5</i>	0.82	C	E	D,G	F A B
	<i>m6</i>	0.80	C,E	D,G	F	A B
metabolite-based	<i>m7, m8, m9</i>	0.81	A,B,C,D,E,G	F		
neighborhood sensitive	<i>m10</i>	0.92	A,B,C,F,G	D,E		
	<i>m11</i>	0.86	A,B,C,F,G	D,E		
	<i>m12</i>	0.72	A,B,C,F,G	D,E		

### 6.1.1.2. Automatic Classification of Organisms

In this section, for each distance measure the automatically derived classifications are described. The respective dendrograms can be found in the Appendix (see Figures A.1, A.2, A.3, A.4, A.5, and A.6 on pages 96 ff.), whereas Table 6.5 summarizes the automatic classification results. Distance measures are sorted by the type of information they are based upon. Firstly, reaction-based distance measures are described, then reaction and metabolite-based ones, followed by metabolite-based distance measures and the reaction neighborhood sensitive distance measures.

**Reaction-based Distance Measures** (*m1*, *m2*, *m3*). For distance measure *m1*, the automatic grouping yields the same four groups for *average* and *ward*: *D*, *G* and *F* are put together into one group, *C* and *E* into another group, while *A* and *B* are singletons. This grouping resembles the reaction-based manual classification. For the *complete* approach, the first group is split into two subgroups, one containing *D* and *G* and the other containing only *F*. This grouping resembles the reaction and metabolite-based manual classification. The *cpcc* for the classification based on *average* and *ward* is higher than the one for the *complete* grouping indicating that the former classification better resembles the underlying distance data. For *m2*, the automatically derived grouping is identical for all three clustering techniques: *A*, *B*, and *C* are grouped together, as are *D*, *G*, and *F*, while *E* is put into a singleton group. This grouping does not resemble any of the manually derived classifications. For *m3*, the automatically derived grouping is the same for all three clustering techniques. It groups *C* together with *E*, as well as *D* together with *H* and *F*, and leaves *A* as well as *B* as singletons. This grouping resembles the reaction-based manual classification. With the exception of *m1 complete*, the automatically derived groupings for *m1* and *m3* are the same for all clustering techniques.

**Reaction and Metabolite-based Distance Measures** ( $m_4, m_5, m_6$ ). For distance measures  $m_4$  and  $m_6$  and all clustering techniques, the automatic grouping procedure yields the following:  $D$  and  $G$  are put together as are  $C$  and  $E$ , while  $A$ ,  $B$ , and  $F$  are put into singleton groups each. This classification is identical to the one for  $m_1$  *complete* and resembles the reaction and metabolite-based manual classification. For  $m_5$ , the automatically derived grouping is the same for all three clustering techniques:  $D$  and  $G$  are grouped together, while all other pseudo-organisms are put into singleton groups. There is no similarity to any of the manually derived classifications.

**Metabolite-based Distance Measures** ( $m_7, m_8, m_9$ ). The three metabolite-based distance measures  $m_7$ ,  $m_8$ , and  $m_9$  show identical groupings for all clustering techniques:  $F$  is put into a singleton group, while the other pseudo-organisms are put into another group. This grouping exactly resembles the manually derived grouping based on the metabolite content.

**Neighborhood Sensitive Distance Measures** ( $m_{10}, m_{11}, m_{12}$ ). For  $m_{10}$ ,  $m_{11}$ , and  $m_{12}$  *average*, *complete*, and *ward*, the automatic grouping procedure puts  $D$  and  $E$  together into a first group, and all other pseudo-organisms into a second group. This classification is similar to the functionality-based manual classification, even though  $F$  is not put into a singleton group here.

### 6.1.1.3. Discussion & Conclusion

For a good combination of distance measure and clustering technique one would expect the automatically derived grouping (see Table 6.5) to resemble the manually derived one (see Table 6.4). However, this is not the case for all such combinations. These differences are discussed in this section.

Of the reaction-based distance measures, the manual and automatic classifications for  $m_1$  and *complete* differ in  $F$  not joining a group with  $D$  and  $G$ , but being put into a singleton group. This difference results from the way the clusters are joined in the complete linkage agglomerative clustering method: the height at which two clusters are joined resembles the maximum distance between any two items, one item from one cluster and the second item from the other cluster. For the analyzed pathway and organisms this leads to  $F$  joining the cluster containing  $D$  and  $G$  at the same height as  $A$  joins  $B$  (see Figure A.1 B on page 96). Thus, it is not possible to deduce a classification of organisms into four groups from the resulting dendrogram. However, it is possible to deduce such a classification from both *average* and *ward* dendrograms (see Figure A.1 A and C on page 96), which for  $m_1$  is the best classification. This grouping also better resembles the underlying distance data as can be deduced from the *cpcc* for both *average* and *ward*, which is greater than the one for *complete*. This suggests to always apply different clustering techniques in order to generate more dendrograms and thus having a higher chance to generate the best classification. In the case all classifications are identical this can be rated as hint towards a good quality of the result. In the case that classifications differ, the one yielding the highest value for the *cpcc* can be selected.

For distance measure  $m_2$ , the manual and automatic groupings differ. The group consisting of  $A$ ,  $B$ , and  $C$  as well as the singleton group containing  $E$  do not occur in the manual grouping. The different grouping for  $m_2$  as compared to  $m_1$  and  $m_3$  is

partly due to the construction of the distance measure and the implementation of the clustering algorithm. The distance measure only counts what both metabolic networks being compared have in common and does not take into account the differences. This results in  $C$  having the same distance from  $A$  and  $B$  as from  $E$ . When differences between reaction networks are taken into account,  $E$  is closer to  $C$  than to each of  $A$  and  $B$ . In general, if distances are equal, the joins in the dendrogram only depend on the order the clustering algorithm processes the items to be clustered. Exchanging names of pseudo-organisms  $A$  and  $E$  and rerunning the analysis yields dendrograms that indeed resemble the ones from  $m1$  and  $m3$  insofar as  $A$  and  $B$  now form a subgroup as do  $C$  and  $E$  (see Figure A.7 A, B, and C on page 102). However, the automatically derived classification is still different from the one for  $m1$  and  $m3$  and thus from the manual classification, since for  $m2$  pseudo-organisms  $A$  and  $B$  are grouped together.

For the reaction and metabolite-based distance measures  $m4$  and  $m6$ , all respective classifications match perfectly. Here,  $F$  is put into a singleton group, which is the effect from additionally taking metabolite content into account. In terms of metabolite content all organisms are identical, despite  $F$ , which is missing one metabolite as compared to all others. The differences for  $m5$  are due to the same reasons as explained above for  $m2$ . Again, this results in  $C$  having the same distance from  $A$  and  $B$  as from  $E$ . Exchanging names for  $A$  and  $E$  and doing the clustering again results in grouping the former  $E$  next to  $C$  in the dendrogram as well as  $B$  next to the former  $A$  (see Figure A.7 D, E, and F on page 102). The classifications automatically derived from all new dendrograms are identical, but not similar to any of the manually derived ones. They are similar to the ones for  $m4$  and  $m6$ , but differ in that  $A$  and  $B$  are grouped together. All automatically derived groupings for  $m4$ ,  $m6$ , and  $m5$  show the effect of the metabolite content, which is that  $F$  is a clear outsider.

All automatically derived classifications for the metabolite-based distance measures  $m7$ ,  $m8$ , and  $m9$  match perfectly. This is not surprising, since with respect to the metabolite content all organisms are identical, despite pseudo-organism  $F$ , which is lacking one metabolite. The reason is that if one reaction is missing, but the neighboring reactions are present, none of the intermediate metabolites is missing in the network. In this case, distance measures based on metabolites only by design cannot measure that the network is altered, because already their basis, the metabolite content, does not properly capture differences in the metabolic networks.

All three neighborhood sensitive distance measures  $m10$ ,  $m11$ , and  $m12$  yield the same classification, which does not match any of the manually derived classifications. However, the classification is close to the functionality-based manual classification with the difference that in the automatic classification  $F$  is not a singleton, but grouped together with  $A$ ,  $B$ ,  $C$ , and  $G$ .

These groupings can be explained with the edit costs listed in Table 6.6. Organisms  $D$  and  $E$  differ by reactions R00001 and R00002, and both lack reaction R00006, while R00006 is present in all other organisms (see Table 6.2). Since the edit costs for deleting or inserting each of the first two reactions are very low and the edit cost for deleting or inserting reaction R00006 is very high, these two organisms are grouped together for all three distance measures. The remaining organisms are put together into another group by the automatic procedure. Analyzing the clustering dendrograms (see Figures A.5 C, D, and E on page 100 and A.6 C, D, and E on page 101) yields that  $F$  is separated from  $A$ ,  $B$ ,  $C$ , and  $G$  for distance measures  $m10$  and  $m12$ , although this is not reflected in the

**Table 6.6.:** Costs of edit operations for reactions of the artificial test pathway for all neighborhood sensitive distance measures.  $s$ : number of synonymous reactions for a particular reaction,  $a$ : number of adjacent reactions for a particular reaction, reaction edit cost formulae for  $m10$ :  $e^{-s+\frac{1}{2}(a)}$ ,  $m11$ :  $\frac{1}{2s} + \frac{a}{2}$ ,  $m12$ :  $\max(1 - s + \frac{a}{2}, 10^{-6})$ .

reaction	s	a	reaction edit cost		
			$m10$	$m11$	$m12$
R00001	3	1	0.082	0.625	0.000
R00002	3	1	0.082	0.625	0.000
R00003	3	1	0.082	0.625	0.000
R00004	1	1	0.607	1.000	0.500
R00005	1	1	0.607	1.000	0.500
R00006	1	7	12.182	4.000	3.500
R00007	1	1	0.607	1.000	0.500
R00008	1	1	0.607	1.000	0.500

automatically derived grouping. For  $m11$ ,  $F$  forms a subgroup with  $G$  in all clustering dendrograms. The reason is that for  $m11$  the edit cost for R00004 is higher than that for any two of R00001, R00002, or R00003, while for  $m10$  and  $m12$  it is vice versa.

Regarding different clustering techniques, one can say that for some distance measures, namely  $m2$ ,  $m3$ ,  $m4$ , and  $m6$ , *average* and *complete* dendrograms are qualitatively identical, while *ward* differs from them (see Figures A.1, A.2, and A.3 on pages 96 ff.), and that for one distance measure, namely  $m5$ , *average* and *ward* are qualitatively identical, while *complete* differs from them (see Figure A.3 on page 98). A specialty of the Ward clustering method is that it has a tendency to avoid outsiders, as has already been reported in the literature, e.g. in Eckes and Roßbach (1980). In the current test scenario this effect can, for example, be found for distance measure  $m2$ : pseudo-organism  $E$  is much closer to the subgroup containing  $A$ ,  $B$ , and  $C$  in the *ward* dendrogram than in the *average* and *complete* dendrograms (see Figure A.1 D, E, F on page 96). A similar effect can be found for pseudo-organism  $F$  in the respective dendrograms of distance measures  $m4$  (see Figure A.2 D, E, F on page 97) and  $m6$  (see Figure A.3 D, E, F on page 98).

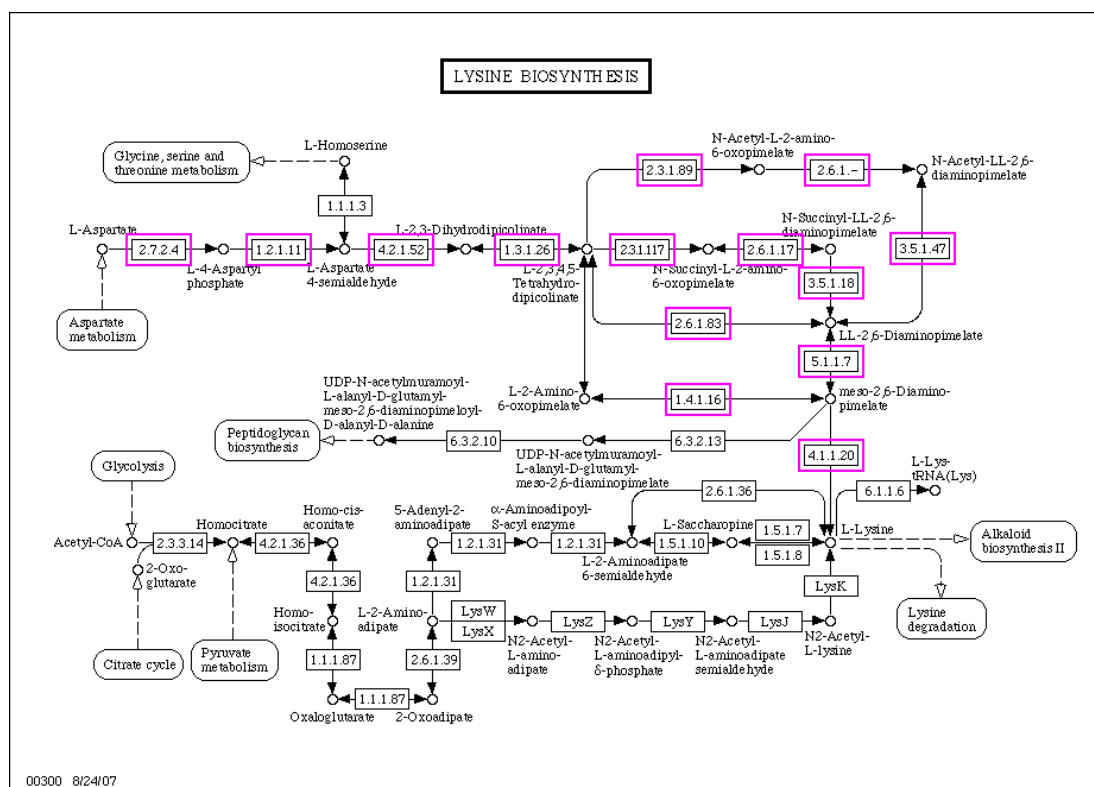
However, there does not seem to be the one best clustering technique, especially since the resulting groupings are the same despite for  $m1$  *complete*. In principle, it is good to have dendrograms originating from different clustering techniques for the case that a particular clustering dendrogram cannot be cut to yield a certain number of groups (as described above for  $m1$  and *complete*). If classifications differ, the respective value of the *cpcc* can be used to choose the classification that best resembles the underlying distance data.

Summarizing the analysis for this test scenario one can state the following. Edit distances and Soergel type edit distances based on reactions as well as based upon both reactions and metabolites perform well for almost all clustering techniques. All neighborhood sensitive reaction edit distances yield the same result, but are not perfect. Maximum common subgraph type edit distances generally perform less well than edit distances and Soergel type edit distances. Moreover, all metabolite-based edit distances perform poorly. There does not seem to exist the one best clustering technique. On the contrary, it seems beneficial to keep all clustering techniques and choose as best classification the one maximizing the *cpcc*. This would also enable a user to manually compare possibly different automatically derived classifications of organisms.

### 6.1.2. Lysine Subpathway Test Scenario

In this test scenario the goal is to evaluate how well the distance measures perform on real world data by assessing how well organisms can be classified according to their individual pathway variants. Classifying real world data may be more difficult, because pathway variants in different organisms may differ significantly in their length (number of associated reactions) or may be affected by misannotations, which both hampers the comparison.

The pathway relied upon in this test case is a subpathway of the KEGG lysine biosynthesis. Figure 6.2 highlights the associated reactions on the KEGG map of lysine biosynthesis. This pathway is chosen, because there exist three different routes from L-aspartate to L-lysine, which moreover have different lengths (in terms of number of associated reactions). This pathway can be used to evaluate whether the automated approach is able to classify organisms according to which metabolic route they implement. Furthermore, it can be assessed how different distance measures cope with the differing lengths of the different routes. A subpathway is used instead of the entire pathway for simplifying manual analysis and thus enabling a more reliable manual classification.



**Figure 6.2.:** KEGG lysine biosynthesis pathway map. Reactions associated with the chosen subpathway are highlighted in magenta. The biosynthesis in this subpathway starts at L-aspartate and ends at L-lysine. It starts with a single reaction chain, splits up into three alternative routes, and merges again into a last reaction that is common to all routes.

For this test scenario organisms that implement different routes of the lysine subpathway are chosen from the KEGG database. Some implement more than one route, and some implement only parts of one or more routes, which might be due to missing or erro-

neous annotations or might reflect an intermediate state of evolution (the organism has lost or aquired, for example by lateral gene transfer, the respective genes). This choice is made in order to assess how the classification results are influenced by imperfect data. The resulting list of organisms comprises (abbreviations in brackets): *Acinetobacter* sp. ADP1 (aci), *Bifidobacterium longum* (blo), *Clostridium acetobutylicum* ATCC 824 (cac), *Clostridium tetani* E88 (ctc), *Corynebacterium glutamicum* (cgl), *Enterococcus faecalis* V583 (efa), *Lactobacillus plantarum* WCFS1 (lpl), *Listeria welshimeri* SLCC5334 (lwe), *Staphylococcus aureus* subsp. aureus COL (sac), and *Staphylococcus epidermidis* ATCC 12228 (sep). For all reactions in the subpathway Table 6.7 shows whether they are implemented in the respective organism, while Table 6.8 shows the metabolite content of these organisms.

**Table 6.7.:** Presence (x) and absence (-) of reactions in analyzed organisms for the lysine biosynthesis subpathway. Bars between columns indicate the manually derived grouping. Organisms' abbreviations are given in the text.

description	reaction	EC number	organisms (KEGG abbreviation)									
			sac	sep	lwe	efa	lpl	cac	aci	blo	cgl	ctc
first steps	R00480	2.7.2.4	x	x	x	x	x	x	x	x	x	x
	R02291	1.2.1.11	x	x	x	x	x	x	x	x	x	x
	R02292	4.2.1.52	x	x	x	x	x	x	x	x	x	x
	R04198	1.3.1.26	x	x	x	x	x	x	x	x	x	x
	R04199	1.3.1.26	x	x	x	x	x	x	x	x	x	x
succinyl	R04365	2.3.1.117	x	-	-	x	x	x	x	x	x	-
	R04475	2.6.1.17	-	-	-	-	-	x	x	-	-	-
	R02734	3.5.1.18	x	x	-	-	x	x	x	x	x	-
	R02735	5.1.1.7	-	-	x	x	x	x	x	x	x	-
acetyl	R04364	2.3.1.89	-	x	x	-	-	-	-	-	-	-
	R04467	2.6.1.-	x	x	x	x	x	x	-	-	-	-
	R02733	3.5.1.47	-	-	x	x	x	-	-	-	-	-
	R02735	5.1.1.7	-	-	x	x	x	x	x	x	x	-
<i>dhh</i>	R02755	1.4.1.16	-	-	-	-	-	-	-	-	x	x
last step	R00451	4.1.1.20	x	x	x	x	x	x	x	x	x	x

### 6.1.2.1. Manual Classification of Organisms

Since the goal is to assess how well the organisms can be grouped according to their pathway variants, the reaction content of the organisms is compared manually in order to group organisms together according to the different routes of the pathway they implement. Organisms with incomplete pathway variants are grouped together with organisms that fully implement them, unless only very few reactions are present. Metabolite content is not explicitly considered, since it depends on the reaction content.

The manual analysis yields the following. There is no difference in the first part of this pathway, consisting of reactions R00480, R02291, R02292, R04198, R04199 (EC 2.7.2.4, 1.2.1.11, 4.2.1.52, and 1.3.1.26). All organisms also implement the last reaction of the pathway, R00451 (EC 4.1.1.20). Inbetween, the pathway is split into three different routes, the succinyl route consisting of R04365, R04475, R02734, and R02735

**Table 6.8.:** Presence (x) and absence (-) of metabolites in analyzed organisms for the lysine biosynthesis subpathway. Metabolites are represented by their KEGG identifier, the abbreviations of the organisms are given in the text.

metabolite	sac	sep	lwe	efa	lpl	cac	aci	blo	cgl	ctc
C00001	x	x	x	x	x	x	x	x	x	x
C00002	x	x	x	x	x	x	x	x	x	x
C00003	x	x	x	x	x	x	x	x	x	x
C00004	x	x	x	x	x	x	x	x	x	x
C00005	x	x	x	x	x	x	x	x	x	x
C00006	x	x	x	x	x	x	x	x	x	x
C00008	x	x	x	x	x	x	x	x	x	x
C00009	x	x	x	x	x	x	x	x	x	x
C00010	x	x	x	x	x	x	x	x	x	-
C00011	x	x	x	x	x	x	x	x	x	x
C00014	-	-	-	-	-	-	-	-	x	x
C00022	x	x	x	x	x	x	x	x	x	x
C00024	-	x	x	-	-	-	-	-	-	-
C00025	x	x	x	x	x	x	x	-	-	-
C00026	x	x	x	x	x	x	x	-	-	-
C00033	-	-	x	x	x	-	-	-	-	-
C00042	x	x	-	-	x	x	x	x	x	-
C00047	x	x	x	x	x	x	x	x	x	x
C00049	x	x	x	x	x	x	x	x	x	x
C00080	x	x	x	x	x	x	x	x	x	x
C00091	x	-	-	x	x	x	x	x	x	-
C00441	x	x	x	x	x	x	x	x	x	x
C00666	x	x	x	x	x	x	x	x	x	-
C00680	x	x	x	x	x	x	x	x	x	x
C03082	x	x	x	x	x	x	x	x	x	x
C03340	x	x	x	x	x	x	x	x	x	x
C03871	-	-	-	-	-	-	-	-	x	x
C03972	x	x	x	x	x	x	x	x	x	x
C04390	x	x	x	x	x	x	-	-	-	-
C04421	x	x	-	-	x	x	x	x	x	-
C04462	x	-	-	x	x	x	x	x	x	-
C05539	x	x	x	x	x	x	-	-	-	-

(EC 2.3.1.117, 2.6.1.17, 3.5.1.18, and 5.1.1.7), the acetyl route consisting of R04364, R04467, R02733, and R02735 (EC 2.3.1.89, 2.6.1.-, 3.5.1.47, and 5.1.1.7), as well as the *ddh* route, consisting of reaction R02755 (EC 1.4.1.16) catalyzed by the enzyme *ddh*. The succinyl and acetyl routes share one reaction, namely R02735 (EC 5.1.1.7). Organisms implementing the complete succinyl route are *Acinetobacter* (*aci*), and *C. acetobutylicum* (*cac*). *C. acetobutylicum*, however, also implements one reaction of the acetyl route, whereas *Acinetobacter* lacks all reactions of this route (except the one reaction common to both routes). *B. longum* (*blo*) implements almost the same reactions as *Acinetobacter*, both differ in the absence of only one reaction in the succinyl route. *Acinetobacter*, *B. longum* and *C. glutamicum* (*cgl*) have in common that all reactions of the acetyl route are missing, except for reaction R02735, which is part of both routes. However, *C. glutamicum* additionally implements reaction R02755, which is the only reaction of the *ddh* route. *C. tetani* (*ctc*) implements no reaction of both the succinyl and acetyl route, but is the only organism, besides *C. glutamicum*, that implements the single reaction of the *ddh* route. *E. faecalis* (*efa*) implements three of four reactions of the acetyl route and only one additional reaction in the succinyl route. *L. plantarum* (*lpl*) differs by one reaction from *E. faecalis*, implementing three out of four reactions of the succinyl route. *L. welshimeri* (*lwe*) implements all reactions of the acetyl route and no additional reactions of the succinyl route. Nevertheless, it is still similar to the two aforementioned organisms. The remaining organisms, *S. aureus* (*sac*) and *S. epidermidis* (*sep*), share one reaction of the succinyl and one reaction of the acetyl route. *S. aureus* additionally implements a reaction in the succinyl route and *S. epidermidis* another reaction in the acetyl route. Both lack reaction R02735, which is common to both routes.

The organisms *C. acetobutylicum*, *Acinetobacter*, *B. longum*, and *C. glutamicum* could be put into a first group of organisms. These organisms either entirely or at least almost completely implement the succinyl route, while lacking the acetyl route. *C. acetobutylicum* and *Acinetobacter* differ by one reaction, *Acinetobacter* and *B. longum* by another reaction, and *B. longum* and *C. glutamicum* by yet another one. This group might be split into *C. acetobutylicum* and *Acinetobacter* on the one hand and *C. glutamicum* and *B. longum* on the other hand due to the larger distance of three reactions between *C. acetobutylicum* and *C. glutamicum*. *E. faecalis*, *L. plantarum*, and *L. welshimeri* can be put into another group. These organisms completely, respectively almost completely, implement the acetyl route, and only some reactions of the first route. *S. aureus* and *S. epidermidis* have either one or two reactions annotated for each route and therefore can be put together into a group. Both routes seem to be unfunctional for these organisms. *C. tetani* is the only organism lacking all reactions from both the succinyl route and the acetyl route. However, *C. tetani* and *C. glutamicum* have the presence of the *ddh* route in common, and these two organisms are the only ones with this trait. But as for the remaining reaction content, *C. glutamicum* is identical to *B. longum*.

### 6.1.2.2. Automatic Classification of Organisms

All combinations of distance measures and clustering techniques were applied to the lysine subpathway and the above defined set of organisms. The resulting dendrograms can be found in the Appendix (see Figures A.8, A.9, A.10, A.11, A.12, and A.13 on pages 103 ff.), whereas the automatically derived groupings are listed in Table 6.9.



**Table 6.9.:** Classification results for the lysine biosynthesis subpathway: for all distance measures and clustering methods the automatically derived classification of organisms is shown. The first column (DM) holds the distance measure identifier (abbreviations as defined in Table 6.1), the second column (CM) the clustering method (A: average linkage agglomerative clustering, C: complete linkage agglomerative clustering, W: Ward clustering). The next column depicts the *cpcc* of the classification. The remaining columns hold the groups of organisms in the automatically derived classification. Organisms are colored according to the manual grouping: dark blue for existence of succinyl route, light blue for succinyl and *ddh* route, green for *ddh* route, red for acetyl route, pink for succinyl and acetyl route, and orange for organisms that are difficult to classify, because only few reactions of any route are present. Abbreviations of organisms are given in the text.

DM	CM	cpcc	automatic classification						
m1	A C W	0.73	aci,blo,cac,cgl	ctc	efa,lpl,lwe	sac,sep			
m2	A C W	0.70	aci,blo,cac,cgl	ctc	efa,lpl,lwe	sac,sep			
m3	A C W	0.74	aci,blo,cac,cgl	ctc	efa,lpl,lwe	sac,sep			
m4	A W	0.68	aci,blo,cgl	cac,efa,lpl,sac	ctc	lwe,sep			
m4	C	0.67	aci,blo,cgl	cac,lpl,sac	ctc	efa,lwe	sep		
m5	A C W	0.64	aci,blo,cgl	cac,lpl,sac	ctc	efa,lwe	sep		
m6	A W	0.69	aci,blo,cgl	cac,efa,lpl,sac	ctc	lwe,sep			
m6	C	0.68	aci,blo,cgl	cac,lpl,sac	ctc	efa,lwe	sep		
m7	A C	0.66	aci,blo,cgl	cac,efa,lpl,sac	ctc	lwe,sep			
m7	W	0.65	aci,blo,cgl	cac,efa,lpl,lwe,sac,sep	ctc				
m8	A C W	0.66	aci,blo,cgl	cac,efa,lpl,lwe,sac,sep	ctc				
m9	A C	0.67	aci,blo,cgl	cac,efa,lpl,sac	ctc	lwe,sep			
m9	W	0.66	aci,blo,cgl	cac,efa,lpl,lwe,sac,sep	ctc				
m10	A C W	0.85	aci,blo,cac,efa,lpl,lwe,sac,sep	cgl,ctc					
m11	A	0.73	aci,blo,cac,sac	cgl,ctc	efa,lpl,lwe	sep			
m11	C W	0.72	aci,blo,cac	cgl,ctc	efa,lpl,lwe	sac,sep			
m12	A	0.71	aci,blo,cac,sac	cgl	ctc	efa	lpl	lwe	sep
m12	C W	0.72	aci,blo,cac,sac	cgl,ctc	efa	lpl	lwe	sep	

**Reaction-based Distance Measures** ( $m1, m2, m3$ ). For all three reaction-based distance measures  $m1$ ,  $m2$ , and  $m3$ , and all three clustering techniques the classification result is identical, namely *C. tetani* in a singleton group, *S. aureus* and *S. epidermidis* in another group, *E. faecalis* and *L. plantarum*, and *L. welshimeri* in yet another group, and *Acinetobacter*, *B. longum*, *C. acetobutylicum*, and *C. glutamicum* in the last group. This grouping is in accordance with the manual grouping: all organisms implementing the succinyl route are grouped together including *C. glutamicum*, which additionally implements the *ddh* route. *C. tetani*, the other organism implementing the *ddh* route is put into a group on its own. The two organisms *S. aureus* and *S. epidermidis*, which are not easy to classify, are grouped together, and *L. plantarum* is grouped together with *E. faecalis* and *L. welshimeri*.

**Reaction and Metabolite-based Distance Measures** ( $m4, m5, m6$ ). For the reaction and metabolite-based distance measures, the classifications of  $m4$  and  $m6$  are identical. The four resulting groups for *average* and *ward* clustering are *C. tetani* in a singleton group, *L. welshimeri* and *S. epidermidis* in the second group, *Acinetobacter*, *B. longum*, and *C. glutamicum* in the third group, and *C. acetobutylicum*, *E. faecalis*, *L. plantarum*, and *S. aureus* in the fourth group. In contrast to this, for both  $m4$  and  $m6$  *complete*, *S. epidermidis* forms a singleton group, and *E. faecalis* is not grouped together with *C. acetobutylicum*, *L. plantarum*, and *S. aureus*, but with *L. welshimeri*. The grouping for  $m5$  and all clustering techniques is identical to that for  $m4$  and  $m6$  *complete*. For both  $m4$  and  $m6$  the *cpcc* is higher for the respective classifications based on *average* and *ward* than for *complete* indicating that the former better resemble the respective distance data.

**Metabolite-based Distance Measures** ( $m7, m8, m9$ ). For the metabolite-based distance measures  $m7$  and  $m9$  *average* and *complete* the grouping is identical to that for  $m4$  and  $m6$  *average* and *ward*. *C. tetani* in a singleton group, *L. welshimeri* and *S. epidermidis* in the second group, *Acinetobacter*, *B. longum*, and *C. glutamicum* in the third group, and *C. acetobutylicum*, *E. faecalis*, *L. plantarum*, and *S. aureus* in the fourth group. Both  $m7$  and  $m9$  *ward* are identical to  $m8$  and all clustering techniques: *C. tetani* is put into a singleton group, *Acinetobacter*, *B. longum*, and *C. glutamicum* make up the second group, and the remaining organisms form the third group. For both  $m7$  and  $m9$  the *cpcc* is higher for the respective classifications based on *average* and *complete* than for *ward* indicating that the former better resemble the respective distance data.

**Neighborhood Sensitive Distance Measures** ( $m10, m11, m12$ ). The neighborhood sensitive distance measure  $m10$  shows the same grouping for all clustering techniques: *C. glutamicum* and *C. tetani* are grouped together, while all other organisms are put into the second group. For  $m11$  and the *average* approach *S. epidermidis* is put into a singleton group, *C. glutamicum* and *C. tetani* form another group, *E. faecalis*, *L. plantarum*, and *L. welshimeri* are put together and so are *Acinetobacter*, *B. longum*, *C. acetobutylicum*, and *S. aureus*, while for *complete* and *ward*, *S. aureus* does not join *Acinetobacter*, *B. longum*, and *C. acetobutylicum*, but *S. epidermidis*. For distance measure  $m12$  *average*, *Acinetobacter*, *B. longum*, *C. acetobutylicum*, and *S. aureus* are put together into one group, while all other organisms form singleton groups. For *complete* and *ward*

the same classification is deduced besides that *C. glutamicum* and *C. tetani* are grouped together. For *m11* the *cpcc* for the classification based on the *average* dendrogram is higher than the *cpcc* for the *complete* and *ward*-based classification indicating that the former better resembles the respective distance data. For *m12*, however, the *average* classification yields a lower *cpcc* than the classification deduced from the *complete* and *ward* dendrograms.

### 6.1.2.3. Discussion & Conclusion

As in the first test scenario, for a good combination of distance measure and clustering technique one would expect the automatically derived grouping (see Table 6.9) to resemble the manually derived one (see Table 6.7). However, this is not the case for some distance measures and clustering techniques.

All reaction-based distance measures combined with any clustering technique show identical results. The deduced classification perfectly resembles the manual classification. *L. plantarum*, which was not easy to classify manually, is put together with *E. faecalis* and *L. welshimeri*. *C. tetani* is always put into a group on its own. Analyzing the respective dendrograms (see Figures A.8 and A.9 A, B, C on pages 103 ff.) yields that *C. tetani* is not even close to *C. glutamicum*, although both are the only organisms implementing the *ddh* route. This results from the higher number of differences in reaction content for both the succinyl and the acetyl route between these two organisms. However, when analyzing the dendrograms, one recognizes a slight difference: whereas for *m3* and *m2* *Acinetobacter* joins *C. acetobutylicum* and *B. longum* joins *C. glutamicum* and then these groups join, for *m1* *Acinetobacter* joins *B. longum*, then these are joined by *C. acetobutylicum*, and then joined by *C. glutamicum*. This is another case for which the order of joinings depends on the names of the organisms to cluster. For distance measure *m1*, *C. acetobutylicum* and *Acinetobacter* as well as *Acinetobacter* and *B. longum* have the same distance, whereas for *m2* and *m3* *C. acetobutylicum* is closer to *Acinetobacter* than *Acinetobacter* to *B. longum*. Exchanging names of *C. acetobutylicum* and *B. longum* and rerunning the analysis yields new dendrograms that now resemble the ones from *m3* and *m2* (results not shown).

For all reaction and metabolite-based distance measures and all clustering techniques, *Acinetobacter*, *B. longum*, and *C. glutamicum* group together as for the manual classification and the reaction-based distance measures. However, *C. acetobutylicum* is not put into this group here, but instead is grouped together with *S. aureus* and *L. plantarum*. For *m4* *average* and *ward* and *m6* *average* and *ward*, *E. faecalis* also joins this group. Analyzing the metabolite content (see Table 6.8) explains this difference to the reaction-based distance measures: *C. acetobutylicum* and *S. aureus* have all metabolites in common, and *L. plantarum* differs only by one metabolite, whereas all other discrepancies are larger. Another difference to the manual classification is the group consisting of *L. welshimeri* and *S. epidermidis*, which occurs whenever *E. faecalis* is grouped together with *C. acetobutylicum*, *L. plantarum*, and *S. aureus*. Altogether, these combinations of distance measures and clustering techniques are not suitable for grouping organisms according to their pathway variants.

All metabolite-based distance measures result in groupings that do not resemble the manually derived grouping. *C. acetobutylicum* is grouped together with *S. aureus*, instead of with the very close organism *Acinetobacter*. Analyzing the metabolite content

yields that *C. acetobutylicum* and *S. aureus* share all metabolites, which explains this grouping. Instead, *C. acetobutylicum* and *L. plantarum* are always grouped together, because they only differ by one metabolite. Also these combinations of distance measures and clustering techniques are not considered well suited.

The neighborhood sensitive distance measures *m10* and *m11* have in common that for all clustering techniques *C. glutamicum* is grouped together with *C. tetani*. This correlates with these two organisms being the only ones implementing the *ddh* route. However, *C. glutamicum* also implements the succinyl route, which *C. tetani* does not. Analyzing the costs for edit operations on the reactions (see Table 6.10) reveals that reaction R02755 (EC 1.4.1.16), which is catalyzed by *ddh*, has a large number of adjacent reactions, which makes deleting this reaction having a huge impact on the distances calculated by any of the neighborhood sensitive distance measures. For *m10* and all clustering techniques the other organisms are put together into a second group, which is not even close to the manually derived classification of the organisms. However, analyzing the dendrograms shows that *E. faecalis*, *L. plantarum*, and *L. welshimeri* are always put closely together as in the manually derived grouping. But even if the dendrogram would be cut at another height, still *S. aureus* and *S. epidermidis* would be grouped together with *Acinetobacter*, *B. longum*, and *C. acetobutylicum*. For *m11 complete* and *ward* besides *C. glutamicum* and *C. tetani*, *Acinetobacter*, *B. longum*, and *C. acetobutylicum* form one group, *S. aureus* and *S. epidermidis* another, and *E. faecalis*, *L. plantarum*, and *L. welshimeri* the last. This grouping is in accordance with the manually derived one. For the *average* approach, *S. epidermidis* is put into a singleton group, while *S. aureus* joins the group that also contains *Acinetobacter*, *B. longum*, and *C. acetobutylicum*. This grouping still is considered as conform with the manual classification, because *S. aureus* and *S. epidermidis* are organisms that are not easy to classify. For the neighborhood sensitive distance measure *m12* and *average*, *Acinetobacter*, *B. longum*, *C. acetobutylicum*, and *S. aureus* are put into one group, while all remaining organisms are classified into singleton groups. For *complete* and *ward* the only difference is that here *C. glutamicum* and *C. tetani* make up another group, as for distance measures *m10* and *m11*. So far the grouping is in agreement with the manual classification. However, *E. faecalis* does not join *L. welshimeri*, and neither *L. plantarum* is grouped together with *E. faecalis* nor *L. welshimeri* with *Acinetobacter*, *B. longum*, and *C. acetobutylicum*. Thus, no classification based on *m12* resembles the manually derived grouping.

Summarizing this test case, one can say that distance measures based on reactions as well as the neighborhood sensitive distance measure *m11* perform well and much better than those based on both reactions and metabolites and those based on metabolites alone. There is no clear preference for any of the clustering techniques.

### 6.1.3. Choice of Distance Measure and Clustering Technique

In this section the analysis results of both test scenarios are summarized and a conclusion is drawn as to which distance measure and clustering technique to use for further applications.

Of all analyzed distance measures, the reaction-based edit distance *m1* as well as the Soergel type reaction edit distance *m3* perform best. The neighborhood sensitive reaction edit distance *m11* also performs well, but not as well as the former two (see the conclusion of the first test scenario, in Section 6.1.1.3 on page 64). The difference

**Table 6.10.:** Costs of edit operations for reactions of the lysine biosynthesis subpathway for all neighborhood sensitive distance measures.  $s$ : number of synonymous reactions for a particular reaction,  $a$ : number of adjacent reactions for a particular reaction, reaction edit cost formulae for  $m10$ :  $e^{-s+\frac{1}{2}(a)}$ ,  $m11$ :  $\frac{1}{2s} + \frac{a}{2}$ ,  $m12$ :  $\max(1 - s + \frac{a}{2}, 10^{-6})$ .

reaction	s	a	reaction edit cost		
			$m10$	$m11$	$m12$
R00451	1	2	1.000	1.500	1.000
R00480	1	1	0.607	1.000	0.500
R02291	1	9	33.115	5.000	4.500
R02292	1	8	20.086	4.500	4.000
R02733	1	8	20.086	4.500	4.000
R02734	1	6	7.389	3.500	3.000
R02735	1	4	2.718	2.500	2.000
R02755	1	14	403.429	7.500	7.000
R04198	1	8	20.086	4.500	4.000
R04199	1	12	148.413	6.500	6.000
R04364	1	6	7.389	3.500	3.000
R04365	1	6	7.389	3.500	3.000
R04467	1	4	2.718	2.500	2.000
R04475	1	4	2.718	2.500	2.000

between the reaction edit distance and the Soergel type reaction edit distance is the normalization factor. Normalizing can be understood as additionally weighting the cost for an edit operation. The first option is to weight all edit operations in each comparison relative to the size of the two organisms being compared, which results in different weights for different pairs of organisms (Soergel type reaction edit distance). The second option is to weight all edit operations equally over all pairs of organisms to be compared (reaction edit distance). In order to avoid distortion, the decision is made in favor of the **reaction edit distance**.

In both test scenarios no single best clustering technique could be determined. On the contrary, it seems beneficial to keep all clustering techniques, namely *average*, *complete*, and *ward* for being able to manually compare the automatically derived classification of organisms. This increases the chance to find the best classification for a given pathway and set of organisms. If all classifications are identical this indicates a good quality of the grouping. In the case classifications differ, the best classification is determined as the one yielding the highest value for the *cpcc*.

## 6.2. Comparative Metabolic Pathway Analysis of Five Corynebacteria

Five species from the *Corynebacterium* genus are analyzed using the newly developed approach for metabolic network comparison. The comparative analysis is performed for all KEGG pathways as well as for the overall metabolic network of these organisms based on the normalized reaction edit distance  $m1$ . The five analyzed organisms comprise all Corynebacteria currently available in the KEGG database, namely *Corynebacterium diphtheriae* NCTC 13129 (KEGG abbreviation: cdi) *Corynebacterium efficiens* YS-314

(cef) *Corynebacterium glutamicum* ATCC 13032 in the KEGG variant from Kyowa Hakko (cgl), *Corynebacterium jeikeium* K411 (cjk), and *Corynebacterium urealyticum* DSM 7109 (cur). These organisms were chosen because they are taxonomically closely related to each other, but nevertheless occur in different environments or habitats and comprise pathogenic as well as non-pathogenic species. This makes them an interesting set of organisms to compare, since one may find hints concerning pathogenicity factors, clues for lifestyle prerequisites, habitat related adaptations, or pathways containing reactions for which the existing annotation of corresponding genes in the analyzed organisms can be improved. The phylogenetic relationship of Corynebacteria has been analyzed by Khamis *et al.* (2005) using the neighbor-joining method based on concatenated 16S rRNA and *rpoB* gene sequences. In the resulting dendrogram, *C. glutamicum* can be found close to *C. efficiens*, and *C. jeikeium* close to *C. urealyticum*, while *C. diphtheriae* is in the same branch as the first two, but at some distance. While *C. glutamicum* and *C. efficiens* are non-pathogenic (Kalinowski *et al.*, 2003; Nishio *et al.*, 2003), *C. diphtheriae*, *C. jeikeium* and *C. urealyticum* are human pathogens (Cerdeño-Tárraga *et al.*, 2003; Tauch *et al.*, 2005, 2008).

### 6.2.1. Classification Results

In case the comparative analysis of a set of organisms is performed for a large number of pathways, the resulting list of pathways with corresponding classifications of organisms is very long and thus very time-consuming to inspect and to interpret. In order to ease this procedure, it is possible to sort the list of pathways using different strategies. One strategy is to sort the list according to the maximum amount of mutually missing reactions (see Section 5.2). This strategy is called **absolute sorting strategy**. If this sorting strategy is applied, for the listed pathways the set of analyzed organisms for at least one clustering technique is split into groups in a way such that for at least one pair of groups there exist many reactions in all organisms of one group that are missing from the other group or vice versa. If the number of these reactions is reasonably high, the two groups can be regarded well-separated. Well-separatedness of clusters can be interpreted as an indicator for a good clustering result (Handl *et al.*, 2005).

The automatically derived classifications of the five Corynebacteria for the top five pathways according to this sorting strategy are shown in Table 6.11, while Table 6.12 shows the drc for each pair of groups of Corynebacteria for the same pathways. A selection of these pathways will be discussed below in more detail.

However, one may argue that this sorting is not appropriate, because it only accounts for the absolute number of mutually missing reactions. This means that a large pathway (comprising many reactions) might be on top of the list, because half of its reactions make up the difference between two groups of organisms, while a small pathway for which also half of the reactions belong to the differential reaction content might not even be close to the top. In order to address this problem, it is also possible to sort according to the maximum amount of mutually missing reactions relative to the total number of reactions in the respective pathway. Note that in this sorting, the top listed pathway is not necessarily the one with best separated clusters. The top five pathways according to this **relative sorting strategy** are listed in Table 6.13, and a selection of these will be discussed below. The automatically derived classifications of the Corynebacteria into groups are not explicitly listed, since they can be deduced from the same table.

**Table 6.11.:** Automatically derived classification of organisms for top five pathways resulting from comparative pathway analysis of the Corynebacteria *C. diphtheriae* (KEGG abbreviation cdi), *C. efficiens* (cef), *C. glutamicum* ATCC 13032 (cgl), *C. jeikeium* (cjk), and *C. urealyticum* (cur). Results are sorted according to the absolute sorting strategy. The columns provide the pathway name, KEGG pathway number (NO), clustering method (CM, A: average linkage agglomerative, C: complete linkage agglomerative, W: Ward method), and the groups of organisms.

pathway name	NO	CM	groups			
Fatty acid metabolism	00071	A C W	cdi cgl	cef	cjk	cur
Porphyrin and chlorophyll metabolism	00860	A C W	cdi	cef cgl	cjk cur	
Purine metabolism	00230	A C W	cdi cur	cef	cgl	cjk
1- and 2-Methylnaphthalene degradation	00624	A C W	cdi cef cur	cgl	cjk	
Fatty acid biosynthesis	00061	A C W	cdi	cef cgl	cjk cur	

**Table 6.12.:** Top five pathways resulting from comparative pathway analysis of the Corynebacteria *C. diphtheriae* (KEGG abbreviation cdi), *C. efficiens* (cef), *C. glutamicum* ATCC 13032 (cgl), *C. jeikeium* (cjk), and *C. urealyticum* (cur). Results are sorted according to the absolute sorting strategy. The columns provide the pathway name, KEGG pathway number (NO), clustering method (CM; A: average linkage agglomerative, C: complete linkage agglomerative, W: Ward method), two groups of organisms, as well as the differential reaction content for these two groups. The latter is subdivided into the number of all reactions in the respective pathway (pw), number of reactions implemented by all organisms (all), number of reactions occurring in all organisms of group 1, but in no organism from group 2 (1a), number of reactions occurring in some, but not all organisms of group 1, and in no organism from group 2 (1s), number of reactions occurring in all organisms of group 2, but in no organism from group 1 (2a), number of reactions occurring in some, but not all organisms of group 2, and in no organism from group 1 (2s).

pathway name	NO	CM	group1	group2	reaction content					
					pw	all	g1a	g1s	g2a	g2s
Fatty acid metabolism	00071	A C W	cdi cgl	cef	47	10	0	0	12	0
			cdi cgl	cjk	47	9	1	0	23	0
			cdi cgl	cur	47	10	0	1	13	0
			cef	cjk	47	24	1	0	11	0
			cef	cur	47	17	8	0	8	0
			cjk	cur	47	24	11	0	1	0
Porphyrin and chlorophyll metabolism	00860	A C W	cdi	cef cgl	96	20	15	0	1	1
			cdi	cjk cur	96	16	23	0	0	0
			cef cgl	cjk cur	96	15	6	4	0	0
Purine metabolism	00230	A C W	cdi cur	cef	145	49	7	5	1	0
			cdi cur	cgl	145	47	9	2	8	0
			cdi cur	cjk	145	54	2	1	6	0
			cef	cgl	145	51	2	0	10	0
			cef	cjk	145	49	4	0	18	0
			cgl	cjk	145	58	3	0	9	0
1- and 2-Methylnaphthalene degradation	00624	A C W	cdi cef cur	cgl	43	4	0	1	13	0
			cdi cef cur	cjk	43	0	4	4	4	0
			cgl	cjk	43	4	17	0	1	0
Fatty acid biosynthesis	00061	A C W	cdi	cef cgl	53	34	7	0	1	0
			cdi	cjk cur	53	31	10	0	1	0
			cef cgl	cjk cur	53	25	10	0	7	0

**Table 6.13.:** Top five pathways resulting from comparative pathway analysis of the Corynebacteria *C. diphtheriae* (KEGG abbreviation cdi), *C. efficiens* (cef), *C. glutamicum* ATCC 13032 (cgl), *C. jeikeium* (cjk), and *C. urealyticum* (cur). Results are sorted according to the relative sorting strategy. Columns and abbreviations are as in Table 6.12.

pathway name	NO	CM	group1	group2	reaction content					
					pw	all	g1a	g1s	g2a	g2s
Inositol metabolism	00031	A C W	cdi cef cjk cur	cgl	8	0	0	0	6	0
Biosynthesis of siderophore group nonribosomal peptides	01053	A C W	cdi cef cgl cur	cjk	5	0	0	0	3	0
Fatty acid metabolism	00071	A C W	cdi cgl	cef	47	10	0	0	12	0
			cdi cgl	cjk	47	9	1	0	23	0
			cdi cgl	cur	47	10	0	1	13	0
			cef	cjk	47	24	1	0	11	0
			cef	cur	47	17	8	0	8	0
			cjk	cur	47	24	11	0	1	0
1- and 2-Methylnaphthalene degradation	00624	A C W	cdi cef cur	cgl	43	4	0	1	13	0
			cdi cef cur	cjk	43	0	4	4	4	0
			cgl	cjk	43	4	17	0	1	0
Fluorobenzoate degradation	00364	A C W	cdi cjk cur	cef cgl	22	0	0	0	8	2

Sorting the lists of results as described above is appropriate if not much is known about the organisms under investigation. However, if there exists additional knowledge about which grouping of organisms might be of special interest, this can be used to **filter** the list of pathways. If, for example, the taxonomic relationship between the analyzed organisms is known, it might be of interest to list only those pathways that exhibit a grouping differing from the taxonomic classification. Similarly, if the pathogenicity of the analyzed organisms is known, groupings of interest might be those that either group pathogenic or non-pathogenic species together. This helps to find pathways and reactions that are unique for pathogens and therefore possibly important for their pathogenic lifestyle. Thus, the respective enzymes and genes might be of interest for drug design. Furthermore, filtering pathways for which some organism is put into a singleton group helps to find metabolic specialties or to reveal missing or erroneous annotations.

It can be specified whether the filtering is strict or whether the groups may contain other organisms in addition to the ones defined in the filter. Filtering and sorting can be applied to the same data set. Filters can be inclusive or exclusive. As an example, Table 6.14 shows the top 8 pathways for which *C. urealyticum* is clustered into a singleton group sorted according to the absolute sorting strategy. The results for one of these will be discussed below. The automatically derived classifications of the Corynebacteria into groups are not explicitly listed, since they can be deduced from the same table.



**Table 6.14.:** Top eight filtered pathways resulting from comparative pathway analysis of the Corynebacteria *C. diphtheriae* (KEGG abbreviation cdi), *C. efficiens* (cef), *C. glutamicum* ATCC 13032 (cgl), *C. jeikeium* (cjk), and *C. urealyticum* (cur). Only pathways for which *C. urealyticum* is grouped into a singleton cluster are displayed. The list is sorted according to the absolute sorting strategy. Columns and abbreviations are as in Table 6.12.

pathway name	NO	CM	group1	group2	reaction content					
					pw	all	1a	1s	2a	2s
Fatty acid metabolism	00071	A C W	cdi cgl	cef	47	10	0	0	12	0
			cdi cgl	cjk	47	9	1	0	23	0
			cdi cgl	cur	47	10	0	1	13	0
			cef	cjk	47	24	1	0	11	0
			cef	cur	47	17	8	0	8	0
			cjk	cur	47	24	11	0	1	0
Biosynthesis of unsaturated fatty acids	01040	A C W	cdi cef cgl cjk	cur	41	0	3	0	10	0
Starch and sucrose metabolism	00500	A C W	cdi cjk	cef	84	17	0	2	2	0
			cdi cjk	cgl	84	17	0	1	3	0
			cdi cjk	cur	84	12	5	2	0	0
			cef	cgl	84	18	1	0	3	0
			cef	cur	84	12	7	0	0	0
			cgl	cur	84	12	9	0	0	0
Nitrogen metabolism	00910	A C W	cdi	cef cgl	71	8	0	0	5	1
			cdi	cjk	71	9	1	0	3	0
			cdi	cur	71	8	2	0	6	0
			cef cgl	cjk	71	9	4	1	1	0
			cef cgl	cur	71	11	2	1	1	0
			cjk	cur	71	11	1	0	3	0
Glycolysis/ Gluconeogenesis	00010	A C W	cdi cef cgl	cjk	47	23	2	0	2	0
			cdi cef, cgl	cur	47	23	2	2	1	0
			cjk	cur	47	22	5	0	2	0
Glutamate metabolism	00251	A C W	cdi	cef	35	13	0	0	6	0
			cdi	cgl cjk	35	13	0	0	4	2
			cdi	cur	35	13	0	0	4	0
			cef	cgl cjk	35	16	2	0	1	1
			cef	cur	35	16	3	0	1	0
			cgl cjk	cur	35	15	2	1	1	0
Galactose metabolism	00052	A C W	cdi	cef cgl cjk	51	11	2	0	0	0
			cdi	cur	51	9	5	0	0	0
			cef cgl cjk	cur	51	9	2	1	0	0
Citrate cycle (TCA cycle)	00020	A C W	cdi cef	cgl	29	16	1	0	0	0
			cdi cef	cjk	29	15	2	1	0	0
			cdi cef	cur	29	15	2	1	0	0
			cgl	cjk	29	15	2	0	0	0
			cgl	cur	29	14	3	0	1	0
			cjk	cur	29	14	1	0	1	0

## 6.2.2. Biological Interpretation

As has already been mentioned, it is not only the amount of differences, but also the distribution of organisms into groups that is of interest for interpreting the results for a particular pathway. Especially if groupings are unexpected, this might indicate previously unknown findings.

If for some pathway organisms are grouped according to their habitat or lifestyle, this might indicate the presence of reactions in this pathway that are associated with the organisms' adaptation to their habitat or their particular lifestyle. If, for example, pathogens are grouped together or pathogens are at least not grouped together with non-pathogenic species, this might be due to reactions in this pathway that are associated with the pathogenic lifestyle of the respective organisms.

If the classification is similar to the taxonomic relationship of the organisms, this may reflect alterations in the set of reactions of the respective organisms that are associated with their evolution. These differences might have implications on the lifestyle of the organisms or might reflect the loss of genes in the course of evolution or both.

If the classification reflects none of the two cases above, this might be due to metabolic peculiarities of the organism or might reflect missing or erroneous annotations. In both cases, the respective pathways are of special interest: in the first case, because some special trait of the organism might be the reason for this grouping, and in the second case, because a further analysis of the pathway might lead to an improvement of the existing annotation. Note that incorrect annotations can always influence clustering results and thus any of the above mentioned classifications might be influenced by misannotations.

In the following sections, the classification results are discussed for a selection of the top-ranked pathways from the sorted lists presented above. It is not possible to discuss the results for all pathways, since the amount of analyzed pathways is too large. Pathways are categorized into the above described classes of possible interpretations. Since this classification is not exclusive, pathways might belong to more than one class. Published research results will be used to explain the automatically generated classification of organisms for particular pathways.

### 6.2.2.1. Pathway Analysis in Light of Habitat and Lifestyle

The classifications of organisms for some pathways in the sorted lists can be explained with the habitat or lifestyle of the respective organisms or with metabolic specialties of some of the organisms. In the application case at hand, this might arise due to differences between pathogenic and non-pathogenic *Corynebacteria* and thus be useful for revealing pathogenicity factors. More generally, any metabolic prerequisite for survival in a particular habitat that expresses itself as a number of reactions in a particular metabolic pathway may lead to classifications in which the organisms that implement the respective reactions hold a prominent position. If such pathways are known, classifying organisms with unknown habitat on these pathways may be used as indicator whether they occur in this habitat: on the one hand, if they implement all or almost all necessary reactions, they might be able to survive in the particular habitat, and on the other hand, if too many reactions are missing, they presumably cannot. The KEGG pathways fatty acid metabolism, porphyrin and chlorophyll metabolism, fatty acid biosynthesis, inositol metabolism, and biosynthesis of siderophore group nonribosomal peptides are cases for

which the classification can be explained with the habitat and lifestyle of the organisms.

**Fatty Acid Metabolism** (KEGG pathway number 00071). The top-ranked pathway in the list sorted according to the absolute sorting strategy (see Table 6.12) is fatty acid metabolism. It is not only the amount of mutually missing reactions that makes this pathway an interesting one, but also *C. jeikeium* and *C. urealyticum* not being clustered together, but into a singleton group each, although both are human pathogens. The remaining groups are *C. efficiens* as another singleton group and *C. glutamicum* and *C. diphtheriae* in the last group, which correlates with *C. glutamicum* and *C. diphtheriae* not being capable of degrading fatty acids. In contrast to this, both *C. jeikeium* and *C. urealyticum* are able to metabolize fatty acids. They belong to the lipophilic *Corynebacteria* (Tauch *et al.*, 2005, 2008). However, on the basis of the KEGG annotation they are not grouped together. Analyzing the differential reaction content (data not shown) in more detail reveals that this is mainly due to the lack of an enzyme with EC number 1.1.1.35 catalyzing KEGG reactions R01975, R04737, R04739, R04741, R04743, R04745, and R04748 in *C. urealyticum*. However, this gene is not truly missing in *C. urealyticum*, as is reported below (see Section 6.2.2.2).

*C. jeikeium* and *C. urealyticum* are known to occur on the human skin, so the ability to metabolize fatty acids seems to be an adaptation to their habitat or, from a different point of view, might once has been an advantage and thus has enabled these organisms to colonize this special habitat.

**Porphyrin and Chlorophyll Metabolism** (KEGG pathway number 00860). For the porphyrin and chlorophyll metabolism, *C. diphtheriae* is put into a singleton group. Furthermore, *C. glutamicum* and *C. efficiens* are put together, as are *C. jeikeium* and *C. urealyticum* (see Table 6.12). This classification is similar to the taxonomic relationship of these organisms, and at the same time separates pathogens from non-pathogens. The reason for listing this pathway at second position in the list sorted by the absolute sorting strategy is that for *C. diphtheriae* many reactions are annotated that none of the other analyzed *Corynebacteria* has. KEGG's porphyrin and chlorophyll metabolism is a huge pathway combining many different chains of metabolic conversions required for different cellular needs in different types of organisms. The connection between the different pathway routes is the structural similarity between the respective end products: they all involve porphyrin ring structures. For better understanding the clustering results, it is necessary to examine the individual parts of this pathway. The chlorophyll biosynthesis part obviously is not relevant for the analyzed *Corynebacteria*. However, porphyrin rings are constituents not only of chlorophyll, but also of heme groups whose syntheses make up the second part of this pathway. Heme groups constitute an important part of the cytochromes involved in the respiratory chain. For this reason this part of the pathway is annotated for all analyzed *Corynebacteria*. The third part of this pathway is the synthesis of cobalamin and vitamin B12, both of which also contain porphyrin ring structures. Most reactions involved in this part are only annotated for *C. diphtheriae*, which is the reason for the observed grouping of organisms for this pathway. A reaction chain leads from precorrin 2 to vitamin B12 coenzyme. Only a few prokaryotes are capable of synthesizing cobalamin and vitamin B12, among those *C. diphtheriae* (Rodionov *et al.*, 2003). It is also known that *C. glutamicum* lacks the necessary *cob* genes. Nothing

has been reported so far about *C. efficiens*, *C. jeikeium*, and *C. urealyticum*. Also, a sequence-based homology search using the annotated *cob* genes in those Corynebacteria that lack the respective genes did not return any significant results. Thus, according to current knowledge it seems valid to say that only *C. diphtheriae* is able to de-novo synthesize cobalamin and vitamin B12, and the automatically derived grouping perfectly resembles this distinctive feature.

**Fatty Acid Biosynthesis** (KEGG pathway number 00061). The automated classification groups *C. glutamicum* together with *C. efficiens*, as well as *C. jeikeium* together with *C. urealyticum*, while *C. diphtheriae* is put into a singleton group (see Table 6.12). This perfectly resembles the pathway annotation: *C. glutamicum* and *C. efficiens* have the same annotated reactions, as have *C. jeikeium* and *C. urealyticum*, while both groups differ by some reactions. Every reaction present in any other organism is also annotated for *C. diphtheriae*, despite reaction R07763, which is annotated for all organisms except *C. diphtheriae*. Both *C. glutamicum* and *C. efficiens* have a gene coding for a fatty acid synthase (Radmacher *et al.*, 2005). In contrast to this, *C. jeikeium* is auxotroph for fatty acids (Tauch *et al.*, 2005), and *C. urealyticum* as well (Tauch *et al.*, 2008) due to the absence of a fatty acid synthase gene. Thus, the classification reflects current knowledge about fatty acid biosynthesis capabilities of the organisms, since *C. jeikeium* and *C. urealyticum* are grouped together. Both are known to occur on the human skin, and are able to metabolize fatty acids. Since therefore the ability to de-novo synthesize fatty acids might not be essential, this ability might have been lost in the course of evolution. Interpreting the results for the fatty acid metabolism together with the results for this pathway completes the story: both *C. jeikeium* and *C. urealyticum* are auxotroph for fatty acids in contrast to the other Corynebacteria, while both are able to metabolize fatty acids, which the other Corynebacteria are not.

**Inositol Metabolism** (KEGG pathway number 00031). In the list sorted according to the relative sorting strategy (see Table 6.13), inositol metabolism holds the top position. The automatic classification puts *C. glutamicum* into a singleton group and all other organisms into another group. This grouping is in perfect agreement with the annotation data for this pathway: only for *C. glutamicum* reactions are annotated, and only one reaction is missing. According to Krings *et al.* (2006), *myo*-inositol can be utilized by *C. glutamicum* as a carbon and energy source, but not by the other Corynebacteria being investigated. Nothing has been reported for *C. urealyticum* yet, but a sequence-based homology search for corresponding genes did not return any significant results. Thus, current knowledge is perfectly represented in the organisms' annotations for this pathway and in the results of the automatic classification. This example in particular proves the usefulness of the relative sorting strategy. Using the absolute sorting strategy, this pathway appears at position 23 of the list (results not shown).

**Biosynthesis of Siderophore Group Nonribosomal Peptides** (KEGG pathway number 01053). The second pathway in the relatively sorted list (see Table 6.13) is the biosynthesis of siderophore group nonribosomal peptides, for which only *C. jeikeium* is classified into a group on its own. This perfectly resembles the current pathway annotation in KEGG: for *C. jeikeium* three reactions are annotated, while none are annotated

for any of the other analyzed *Corynebacteria*. According to KEGG, the genes annotated for *C. jeikeium* in this pathway are *jk1285* (EC:5.4.4.2), *jk1821* (EC 3.3.2.1), *jk1819* (EC 1.3.1.28), *jk1820* (EC 2.7.7.58), and *jk1814* (EC 1.14.13.59). Tauch *et al.* (2005) reported that despite *jk1285*, all these genes are organized in a single gene cluster (*jk1805* to *jk1821*), which is involved in siderophore synthesis and iron acquisition. Siderophores are used by many bacteria for iron uptake. The uptake of iron is essential for almost all organisms and growth-limiting in many ecological niches. In particular this is the case for pathogens, because the host specifically limits iron availability as part of its innate defense against invading cellular microorganisms. By using siderophores, pathogens are able to counter the iron restriction imposed by their hosts (Andrews *et al.*, 2003). Therefore, this pathway is of interest also for drug target identification.

A sequence-based homology search for corresponding genes in the remaining *Corynebacteria* did not return any relevant results. Nevertheless, a similar type of iron uptake system can be found in *C. urealyticum* (*pstX* genes, personal communication with PD Dr. Andreas Tauch) as well as in *C. diphtheriae* (Qian *et al.*, 2002). That it does not appear on the KEGG pathway map might either be due to missing annotations in KEGG or due to the lack of an appropriate pathway map.

This example shows how the developed comparative approach can help finding pathways representing metabolic functions that are essential for an organism's survival in a particular habitat.

#### 6.2.2.2. Assisting in Annotation

If the classifications of organisms for some pathway cannot be explained otherwise, they might be caused by missing or erroneous annotations. In this case, searching for genes coding for reactions that are missing in some organism using a sequence-based homology search might help to improve the existing annotation. Two pathways for which candidate genes could be found using this method are the citrate cycle as well as fatty acid metabolism.

**Citrate Cycle** (KEGG pathway number 00020). The citrate cycle appears at position eight of the absolutely sorted filtered list (see Table 6.14). This is an unexpected result, since this pathway is part of the central metabolism, which is not expected to vary much between taxonomically closely related organisms, such as the five analyzed *Corynebacteria*. An expected result would have been that all organisms are grouped together or are grouped in a way similar to their taxonomic relationship.

Analyzing the *drc* for this pathway reveals that the five *Corynebacteria* differ in the annotation of the following four reactions: R00344 (EC 6.4.1.1), R00405, (EC 6.2.1.5), R02570 (EC 2.3.1.61), and R00362 (EC 4.1.3.6). Both *C. jeikeium* and *C. urealyticum* are lacking genes for an enzyme accomplishing the function of a pyruvate carboxylase (Tauch *et al.*, 2005, 2008), which corresponds to reaction R00344 (EC 6.4.1.1). This gene loss may be related to the utilization of exogenous fatty acids as sources for carbon and energy (Tauch *et al.*, 2008). *C. jeikeium*, *C. urealyticum*, and *C. diphtheriae* are moreover missing the *sucCD* genes encoding the alpha and beta subunits of a succinyl-CoA synthetase (R00405, EC 6.2.1.5), which indicates that there exists a variant of the citrate cycle in cutaneous *Corynebacteria* (Tauch *et al.*, 2005, 2008; Cerdeño-Tárraga *et al.*, 2003). The third difference is the apparent lack of a gene coding for a dihydrolipoamide

succinyltransferase (R02570, EC 2.3.1.61) in *C. glutamicum* and *C. jeikeium*. However, results of a sequence-based homology search against these two genomes for genes similar to the ones annotated for the other three analyzed Corynebacteria leads to the assumption that genes coding for this function are present in these species. At last, according to the KEGG data, *C. urealyticum* is missing a gene for the citrate lyase beta subunit (R00362, EC 4.1.3.6). For this case, neither literature could be found confirming this finding, nor did the sequence-based homology search yield any significant results.

This example underlines on the one hand how much the developed pathway comparison approach depends on the quality of the existing annotation, and on the other hand that due to this feature it can lead the track to improve the existing annotation. In particular, the developed approach can be very useful to detect unexpected differences across a set of organisms that otherwise could easily be missed.

**Fatty Acid Metabolism** (KEGG pathway number 00071). As has already been reported above, both *C. jeikeium* and *C. urealyticum* are able to metabolize fatty acids as they belong to the lipophilic Corynebacteria (Tauch *et al.*, 2005, 2008). That they are nevertheless not grouped together for this pathway is mainly due to the lack of a gene in *C. urealyticum* coding for a 3-hydroxyacyl-CoA dehydrogenase catalyzing reactions R01975, R04737, R04739, R04741, R04743, R04745, and R04748 (all EC 1.1.1.35). It has not been reported in the literature that *C. urealyticum* lacks this function, so a sequence-based homology search was conducted using the corresponding gene from the close relative *C. jeikeium*, named *jk0159*. The best hit is a CDS called *cu0178* which is already annotated with the very same function, suggesting that this function is present in *C. urealyticum*. Searching the KEGG database for this gene reveals that it actually is present in the KEGG annotation, but no EC number has been assigned. This might be the reason that this enzyme is not mapped to the respective reaction in the KEGG pathway. Meanwhile the KEGG database has been updated, and as has been predicted here, in the current version this gene is annotated with the respective function. This example again underlines the capability of the developed comparative approach to assist in improving existing annotation.

### 6.2.2.3. Finding Targets for Drug Design

Finding new targets for drug design is an important goal in medicine. The developed approach for metabolic network comparison can be useful in this process in the following way. When comparing pathogenic versus non-pathogenic organisms, in every analyzed pathway for which the grouping separates pathogens from non-pathogens, some reaction might be present in the pathogens only that is obligatory for their pathogenic lifestyle. The analyzed set of Corynebacteria comprises the pathogenic species *C. diphtheriae*, *C. jeikeium*, and *C. urealyticum*. For these species the existence of multidrug resistant strains has been reported (Pereira *et al.*, 2008; Tauch *et al.*, 2005, 2008). Due to their prevalence in clinical settings it is an urgent goal to find new methods to fight these bacteria.

**Fatty Acid Biosynthesis and Fatty Acid Metabolism** (KEGG pathway numbers 00061 and 00071). Pathways of interest here are the fatty acid biosynthesis, for which the two pathogenic species *C. jeikeium* and *C. urealyticum* are clustered together (see

Table 6.12), and the fatty acid metabolism, for which *C. jeikeium* and *C. urealyticum* are clustered into groups on their own. *C. jeikeium* is auxotroph for fatty acids (Tauch *et al.*, 2005), and *C. urealyticum* possibly as well, while both are able to metabolize fatty acids. Thus, inhibiting fatty acid catabolism might be a potent approach to fight these *Corynebacteria*. Cox *et al.* (2004) reported that fatty acid catabolism of *Corynebacteria* that are otherwise able to catabolize fatty acids can actually be inhibited by 4-hydroxy-3-methoxybenzyl alcohol at sub-lethal concentrations. The authors suggest to use this chemical compound as ingredient of deodorants for fighting these bacteria on the surface of the skin for reducing malodor formation. This example shows that metabolic differences relevant for drug design can be detected using the developed comparative approach.

#### 6.2.2.4. Overall metabolic network analysis

Besides analyzing the *Corynebacteria* on each of the KEGG pathways separately, the same analysis was performed on the overall metabolic network of these organisms. The overall metabolic network was constructed by merging all KEGG pathways into a single metabolic network. The dendrograms resulting from this analysis are depicted in the Appendix (see Figure A.14 on page 109). The automatically derived classification of organisms is the same for all three clustering techniques and groups *C. jeikeium* and *C. urealyticum* together, while *C. glutamicum*, *C. efficiens*, and *C. diphtheriae* are put into singleton groups. Also, all three dendrograms are qualitatively identical: in each of them *C. jeikeium* is grouped together with *C. urealyticum*. At some distance, this group is joined by *C. diphtheriae*. *C. glutamicum* is grouped together with *C. efficiens*. These dendrograms closely resemble the phylogenetic relationship as deduced by Khamis *et al.* (2005). Thus, there is not much to be learned from the results of the overall metabolic network comparison. In particular, this analysis fails to detect pathway specific differences, which was possible by comparing smaller pathways. This underlines the appropriateness of a comparative approach based on smaller pathways, like the one developed in this thesis, for revealing interesting details on similarities and differences in metabolism across a set of organisms.

Altogether, this application case shows that the developed approach for metabolic network comparison can assist in finding metabolic peculiarities of the analyzed organisms, detecting metabolic functions necessary for survival in a particular habitat, finding new candidate genes for drug design, and revealing missing or erroneous annotations. In particular, it is beneficial to base the comparisons on smaller pathways instead of on the overall metabolic network. Automatically grouping the analyzed organisms saves a lot of time and effort, since significant metabolic differences for particular organisms, e.g. pathogens, can be detected easily by sorting and filtering the pathways based on the automatic classification for individual pathways and the respective differential reaction content.





---

# Conclusion

---

### 7.1. Summary

In this thesis a fully automated approach for comparative analysis of organisms on the functional level of metabolism yielding a classification of the analyzed organisms according to their individual metabolic pathway variants was developed. In contrast to gene sequence-based comparison techniques, the approach developed herein is based on the functional annotation of genes, namely metabolic reactions. Moreover, instead of comparing individual reactions one at a time, sets of reactions that are jointly involved in the same cellular process, also known as metabolic pathways, are compared.

Data on metabolic pathways were taken from the KEGG database. This includes definitions of metabolic reactions, reaction annotation data for individual organisms as well as data on organization of reactions into metabolic pathways. Metabolic pathways were modeled as directed node labeled graphs. Distance measures were developed based on the theory of edit distances on graphs. It was proven that the distance measures are metrics, and, where appropriate, correspondences between the implemented edit distance-based distance measures and already published distance measures were shown.

The developed comparative analysis approach comprises the following steps. Firstly, pairwise distances are calculated between the pathway variants of a set of organisms to be analyzed. Then, organisms are clustered based on these distances using various clustering approaches which results in a dendrogram for each clustering method. Subsequently, these dendrograms are cut at a certain height and thus a classification (partitioning) of the analyzed organisms into groups is achieved. The number of groups is determined as the value for which the cophenetic correlation coefficient between the cophenetic matrix of the partitioning and the distance matrix is maximized. Finally, the differential reaction content is calculated for each pair of groups and can either be presented in a table or visualized on KEGG's metabolic pathway maps. The entire functionality is implemented as a web-based application called Comparative Pathway Analyzer, which is publicly accessible.

Several distance measures were implemented, namely reaction-based distance measures, metabolite-based distance measures, reaction and metabolite-based distance measures, as well as distance measures that, when calculating the edit cost for the deletion or insertion of a reaction, take into account the neighboring reactions. All distance measures were evaluated against each other in order to find the one that is most adequate for the given data. The evaluation was performed on two manually designed test scenarios, since a standard of truth did not exist. Three different clustering techniques, namely average and complete linkage agglomerative clustering as well as Ward clustering, were evaluated for their suitability to group organisms based on distance data on the organisms' pathway variants.

Furthermore, as an application example, five *Corynebacteria* were compared against each other using the newly developed approach and the results were discussed in light of their biological relevance.

## 7.2. Discussion

### Evaluating distance measures and clustering techniques

For evaluating the performance of distance measures and clustering techniques two test scenarios were constructed. Each scenario consists of a pathway, a set of organisms, and a manually derived classification of the organisms according to their particular pathway variants. The first test scenario was artificially created for testing the correct functioning of the distance measures, whereas the second one is derived from the KEGG lysine biosynthesis pathway for evaluating the performance on a real world example. A distance measure and a clustering technique are considered well suited if the automatically derived classification of organisms resembles the manually derived one.

In the first test scenario most distance measures performed well. Exceptions are the neighborhood sensitive reaction edit distances and the maximum common subgraph (mcs) type edit distances. This presumably is due to their design: since mcs type edit distances are based on the maximum common subgraph, they count what is common to the two reaction networks being compared and do not take into account the differences. Even the normalization relative to the larger of the two networks cannot compensate for that, although in a sense this takes into account the differences at least for the larger of both networks. For the neighborhood sensitive reaction edit distances the automated classification almost perfectly matches the manual one. The difference is that one organism could not be separated from its closest neighbors by the automated procedure, although it functionally differs from them.

In the second test scenario the performance of the distance measures is quite different. The best resemblance between automatic and manual classification was achieved by all reaction-based distance measures (reaction edit distance *m1*, mcs type reaction edit distance *m2*, and Soergel type reaction edit distance *m3*). Their classification perfectly resembles the manual classification.

Metabolite-based distance measures did not perform well. They suffer from the fact that the metabolite content does not properly reflect differences in the metabolic networks, which is due to the derivation of the metabolite content: reactions are set to be present if some gene in the organism's genome has the respective function. Via reaction stoichiometry metabolites are associated with one or more reactions as substrates

or products and a metabolite is set to be present in the network if at least one reaction it is associated with is present. If, for example, one reaction is removed from a network, while all neighboring reactions remain present, none of the intermediate metabolites is removed from the network. In this case metabolite-based distance measures cannot capture that the network is altered. Once there exist good measurements of the metabolite content of an organism, this information could be used to deduce the metabolite content independently from the reaction content. Then, distance measures based on metabolite content alone, or based on both reaction and metabolite content, might become more powerful.

The distance measures based on both reactions and metabolites performed better than those based on metabolites alone. However, they suffer from similar effects. This is probably due to the influence of the metabolite content and the associated problems discussed above.

For the neighborhood sensitive reaction edit distances  $m10$  and  $m12$ , the grouping of organisms is not even close to the manually derived classification of the organisms. This indicates that either the concept for weighting edit operations on reactions based on their network neighborhood (synonymous and adjacent reactions) or the choice of the scoring function is not appropriate. Only the neighborhood sensitive reaction edit distance  $m11$  shows an acceptable result as the automatic classification is close to the manually derived one. The difference between both is the classification of one organism. An analysis of the edit costs suggested that the relation between edit costs of different reactions is not well-balanced. Like for the first test scenario, this underlines that for this type of distance measure the choice of a proper scoring function is essential. The scoring functions tested so far are not fully satisfying and thus leave room for improvements, which could be the subject of further research.

Summarizing both test scenarios one can say the following: of all analyzed distance measures, the reaction edit distance  $m1$  as well as the Soergel type reaction edit distance  $m3$  perform well. Third best is the neighborhood sensitive reaction edit distance  $m11$ . However, the scoring function of the latter for calculating the reaction edit cost needs to be improved.

The difference between the reaction edit distance and the Soergel type reaction edit distance is the normalization factor. Normalizing can be understood as weighting the cost for an edit operation. In the Soergel type reaction edit distance all edit operations in a single pairwise comparison are weighted relative to the size of the two organisms being compared, which results in different weights for a particular reaction when different pairs of organisms are compared. In the reaction edit distance all edit operations are weighted equally over all pairs of organisms to be compared. In order to avoid a possible distortion, the decision was made to weight all edit operations equally and therefore to rely on the reaction edit distance for further analyses.

In both test scenarios, and for most distance measures, all clustering techniques showed qualitatively identical clustering dendrograms. For some distance measures *average* and *complete* dendrograms are qualitatively identical, while *ward* differs from them, and for some distance measures *average* and *ward* are qualitatively identical, while *complete* differs. The automatically derived classification in some cases differs across different clustering techniques. Altogether, no best clustering technique could be identified. On the contrary, it seems beneficial to keep all clustering techniques for being able to manually compare the automatically derived classifications of organisms. This increases the

chance to find the best classification for a given pathway and set of organisms. This is especially the case since it can occur that a particular clustering dendrogram cannot be cut to yield a certain number of groups due to several group joins at the same distance level. If classifications derived from different clustering techniques differ, the best classification is selected as the one for that the *cpcc* between the cophenetic matrix of the partitioning and the original distance matrix is maximized. The higher the value of the *cpcc* the better is the correspondence of the classification with the original distance data. In the case that all three clustering methods yield the same classification of organisms this can be rated as supporting the adequateness of the classification results.

### Application to Five *Corynebacteria* on all KEGG Pathways

The developed approach was applied to five members of the *Corynebacterium* genus on all KEGG pathways as well as the organisms' overall reaction network. For selected pathways the resulting classifications were discussed in more detail in light of their biological relevance. In the following paragraphs, different types of classifications and their biological implication are summarized. Results from the comparative analysis of five *Corynebacteria* are assigned to these types, where appropriate. Note that a particular pathway may be interpreted in more than one way and thus might belong to more than one of these types. Furthermore, incorrect annotations can always influence the calculated distances and thus any of the automatically derived classifications might be influenced by these misannotations.

**Habitat and Lifestyle.** If, for some pathway, the classification is in accordance with the habitat of the organisms, e.g. organisms living in the same environment are grouped together or at least are not grouped together with organisms occurring in other habitats, the difference in reaction content might lead to findings about special adaptations associated with occupation of, and survival in, this habitat. Adaptations include acquisition of new functions as well as loss of functions. A particular habitat is often associated with a certain lifestyle. If, for example, pathogenic species form a group on their own, this might indicate the presence of reactions in this pathway that are associated with the pathogenic lifestyle of the respective organisms. These pathways could be of special interest in medical applications because every reaction, or the corresponding gene, that is unique to any or all of the pathogens in comparison to all non-pathogens is a potential candidate for drug design. Pathways in the application example belonging to this class are fatty acid biosynthesis (KEGG pathway number 00061), and fatty acid metabolism (00071).

**Metabolic Specialty.** The classification of organisms might be due to some metabolic specialty of one or more of the organisms. These pathways are of special interest because some rather specific, and possibly rare, metabolic traits might be discovered. Examples from the analyzed data set are porphyrin and chlorophyll metabolism (00860), as well as inositol metabolism (00031).

**Taxonomic Relationship.** If the classification is similar to the taxonomic relationship of the organisms, this may reflect alterations in the set of reactions that are associated with speciation. However, these differences are often also related to lifestyle and habitat of the organisms. The classifications of the five *Corynebacteria* for fatty acid metabolism (00071) and porphyrin and chlorophyll metabolism (00860), which have already been described above, also resemble the taxonomic relationship of the organisms.

Another example resembling the taxonomic relationship is fluorobenzoate degradation (00364), which has not been further discussed.

**Missing Annotations.** If the classification does not belong to any of the types above, this might be due to missing or erroneous annotations. In this case, the respective pathways are of interest because a further analysis might lead to an improvement of the existing annotation. Examples for this case are citrate cycle (00020) and fatty acid metabolism (00071). For both pathways candidate genes could be found for not yet annotated genes using a sequence-based homology search approach.

The application of the developed approach for comparative metabolic network analysis to five *Corynebacteria* showed impressively that the classifications for most of the top-ranked pathways can already be explained using accepted knowledge from literature. Given the outcome of the current research, it appears probable that for the classifications of the remaining pathways suitable explanations can also be found. This could be the topic of interesting and promising further research projects.

Altogether, this research project has shown that comparative metabolic network analysis, and in particular the developed approach, is valuable for finding differences in metabolism across a set of organisms that may help to broaden the knowledge about metabolic specialties of the analyzed organisms, to detect metabolic functions necessary for survival in a particular habitat or for following a certain lifestyle, to find new candidate genes for drug design, and to reveal missing or erroneous annotations.

The developed approach in particular allows to perform comparative metabolic network analysis in a fully automated fashion. Furthermore, the analysis can be performed on the overall metabolic network, the KEGG pathways as well as on user-defined pathways. Results can be sorted according to the amount of differences between the detected groups and filtered for quickly finding relevant results. What is more, results can be visualized on KEGG pathway maps or on automatically generated pathway diagrams.

## 7.3. Outlook

This section comprises ideas for both further technical developments of the proposed metabolic network comparison approach and for further application areas.

### Technical Improvements

The evaluation of the neighborhood sensitive edit distances showed that the scoring function for calculating the particular cost for each reaction edit operation needs to be improved. It seemed that reactions which are highly connected to other reactions are weighted too strongly. Therefore, an improvement could be achieved by utilizing a scoring function that grows less fast for increasing values of the number of adjacent reactions.

Presumably, the clustering procedure can also be improved. In the current implementation, standard hierarchical techniques are used for clustering the distance data. However, a comprehensive evaluation of other existing clustering approaches might lead to the discovery of a technique that is better suited for clustering distance data on metabolic networks. Alternatively, developing a new clustering technique that is specifically tailored to clustering distance data on metabolic networks could also refine classification results.

Another feature to be improved is the set of available pathways on which the comparative analysis is based. Analyzing the *Corynebacteria* on KEGG pathways as well as on the overall metabolic network has shown that the classification results strongly depend on the particular pathway, i.e. on the specific assembly of reactions. The available KEGG pathways might not provide the combination of reactions best suited for the analysis of special metabolic processes. A software called CARMEN that enables the generation of new pathways has recently been developed in the Computational Genomics group at Bielefeld University. Pathways are generated automatically for individual organisms based on KEGG reaction data and the functional annotation of the organism. Pathways can be edited manually in order to satisfy individual research needs. Thus, reactions can be regrouped into new pathways providing a different view on metabolic processes, which might be more appropriate for comparative analysis.

### Comparative Analysis of Metagenomics Data

In metagenomics, environmental samples of natural microbial communities are sequenced and subsequently analyzed. The term metagenome subsumes the genomes of all individual members of the community. Metagenome approaches allow the analysis of microbes which have so far eluded genomic studies because they cannot be cultivated. Moreover, analyzing the gene content of an entire community has the potential to reveal comprehensive information on the evolution (Hansen *et al.*, 2007), lifestyle (Tyson *et al.*, 2004), diversity (Venter *et al.*, 2004; Krause *et al.*, 2008) as well as metabolism (Gill *et al.*, 2006) of coexisting free-living microbes. Existing metagenome studies on metabolism mostly focus on reconstructing the metabolic network and mapping for visual inspection as, for example, in Tyson *et al.* (2004) or Kalyuzhnaya *et al.* (2008). However, comparing the metabolic capabilities of entire microbial communities from different habitats is a valuable approach for elucidating metabolic adaptations for particular environments.

The developed approach for metabolic network comparison across a set of organisms can be readily applied for comparing metabolic pathways across different metagenome data sets. Firstly, this facilitates visual functional analysis, since the metabolic reactions are mapped onto pathway diagrams and thus arranged into their functional context in cellular metabolism. Secondly, the automated clustering procedure groups microbial communities from different sample sites together if their reconstructed pathway variants are similar to each other. This allows for easy detection of pathways for which the individual metabolic capabilities of the communities differ as well as which of the communities are different. Finally, sorting the list of analyzed pathways according to the amount of differences in reaction content and filtering for particular groups enables the quick detection of the most profound metabolic differences across the analyzed communities, as well as the discovery of metabolic peculiarities of particular communities.

One challenging question in such an approach is how to perform the gene calling and annotation based on genome fragments, because these cannot always be assembled into larger contigs or complete genomes, and in this case could be rather short (e.g. contigs of length 826 bp – 2.1 Mbp for an environmental sample after assembly in Venter *et al.* (2004)). Many gene prediction strategies rely on a training step that is based on a longer sequence of the respective genome prior to the gene calling step. Moreover, it might be that a genome fragment does not contain the entire gene, which makes it difficult to assign a function via sequence homology.

Another challenge is to deal with the different abundances of genome sequences of different species. Due to the sequencing strategy, which produces random shotgun reads, genome sequences of abundant species can be expected to be well represented in the data set in contrast to those of rare species, which may be represented by a small number of sequences only (Gill *et al.*, 2006). From the abundance of the genome sequences one can infer the abundance of genes and encoded metabolic reactions in the microbial community. This information can be used as a measure for the importance of a metabolic reaction or a metabolic pathway in the community. By applying suitable thresholds for distinguishing more relevant reactions and pathways from less relevant ones, and only mapping reactions of either class onto the metabolic pathways to be compared, one might be able to distinguish between more and less relevant pathways for the adaptation of the community to its environment.

Clearly, newly detected proteins cannot be included in the comparative metabolic network analysis, as long as their function remains unknown. Even once their function is elucidated, they might not belong to any existing pathway, so new pathways properly representing their functional context might have to be designed.

### **Comparative Analysis of Gene Expression Data**

Another possible field of application is a comparison of data sets from gene expression analyses. In gene expression analysis the goal is to measure the amount as to which genes are transcribed (expressed) under different conditions, in different tissues or at different points in time. Here, the idea is to map all genes onto pathway maps that correspond to metabolic reactions and are active according to the expression analysis. An individual gene would be said to be active if its expression exceeds a specified significance threshold. The resulting metabolic networks represent the active metabolic network variants under certain conditions, in different tissues or at different points in time. The developed approach has so far only been used to compare the theoretically active (or annotated) metabolic networks. However, it can be readily applied to the new type of metabolic networks. The result is a classification of conditions, tissues, or points in time according to similar active pathway variants.

Questions that can be answered with such an analysis depend on the type of input data. Expression data could, for example, be measured for an uninfected host cell and for the same cell type at different states of infection or on states of infection by different pathogens. The automated metabolic network comparison and sorting can be used for detecting those metabolic pathways that differ in the active reaction content across different conditions. Results from this type of analysis could be helpful in diagnostics. Another example is to compare the metabolism of a pathogen living in the blood versus that of the same pathogen living in a host cell. The goal here is to detect metabolic mechanisms that are active or need to be activated in order to enable the pathogen to invade and survive in the host cell, or, more generally, to discover habitat specific metabolic adaptations. Even if only two data sets are compared against each other and thus the clustering procedure is of no use, the sorting strategy might still be helpful for quickly finding pathways that differ significantly. When time series expression data are analyzed in this way, pathways for which the active pathway variant changes over time can easily be detected.

### **Developing a Library of Habitat Specific Pathway Implementations for Classifying Organisms According to their Habitats**

During the analysis of five species of the *Corynebacterium* genus it became apparent that the developed approach for comparative metabolite network analysis has the potential to detect pathways that are important for the survival of the respective organisms in their particular habitat. These are pathways for which the automatically derived classification groups organisms according to their habitat. By systematically comparing organisms living in the same or similar habitats, pathway variants or sets of pathway variants that are essential for the survival of the organisms could be determined. These can be used to build a library of pathway variants with special relevance for particular habitats. Using pathway variants from this library as indicators, organisms can be tested for their ability to survive in particular environments. A related approach has been published recently by Borenstein *et al.* (2008). The authors analyzed metabolic networks of organisms for deducing the seed set of metabolites which they define as the set of metabolites that, based on the network topology, are exogenously acquired. They showed that the composition of the seed sets significantly correlates with several basic properties characterizing the species' environments and agrees with biological observations concerning major adaptations.

### **Systematically Improving Existing Annotation**

In the application to five *Corynebacteria*, the developed approach for metabolic network comparison proved to be helpful for detecting missing or erroneous annotations. Combining the comparative metabolic pathway analysis and a homology search for sequence-similar genes into an automated method could be a reasonable approach for designing a new software for systematically searching for genes coding for the enzymes catalyzing missing reactions in a set of organisms. For each metabolic pathway the organisms would be clustered and the resulting clustering dendrograms compared to a tree representing their taxonomic relationship. Whenever the position in the clustering dendrogram for some organism does not match the one in the taxonomic tree, among other reasons, this could be due to missing or erroneous annotations. A homology search for missing genes using gene sequences from the closest taxonomic relatives can be performed for detecting candidate genes in the particular organism. Gene sequences from close relatives are good candidates for a homology search since, according to the theory of evolution, the genome of close relatives is likely to be very similar. On the other hand, gene sequences from organisms implementing similar pathway variants might be good candidates for a homology search as well, since via lateral gene transfer entire operons can be transferred from one organism to another resulting in genes with similar sequence being present in organisms that are not closely related taxonomically. These candidate genes can be found in organisms that are classified into the same group by the automatic clustering procedure. Applying this approach would result in a list of candidate genes that can serve as starting point for wet-lab verifications with the final goal to improve the existing annotation.

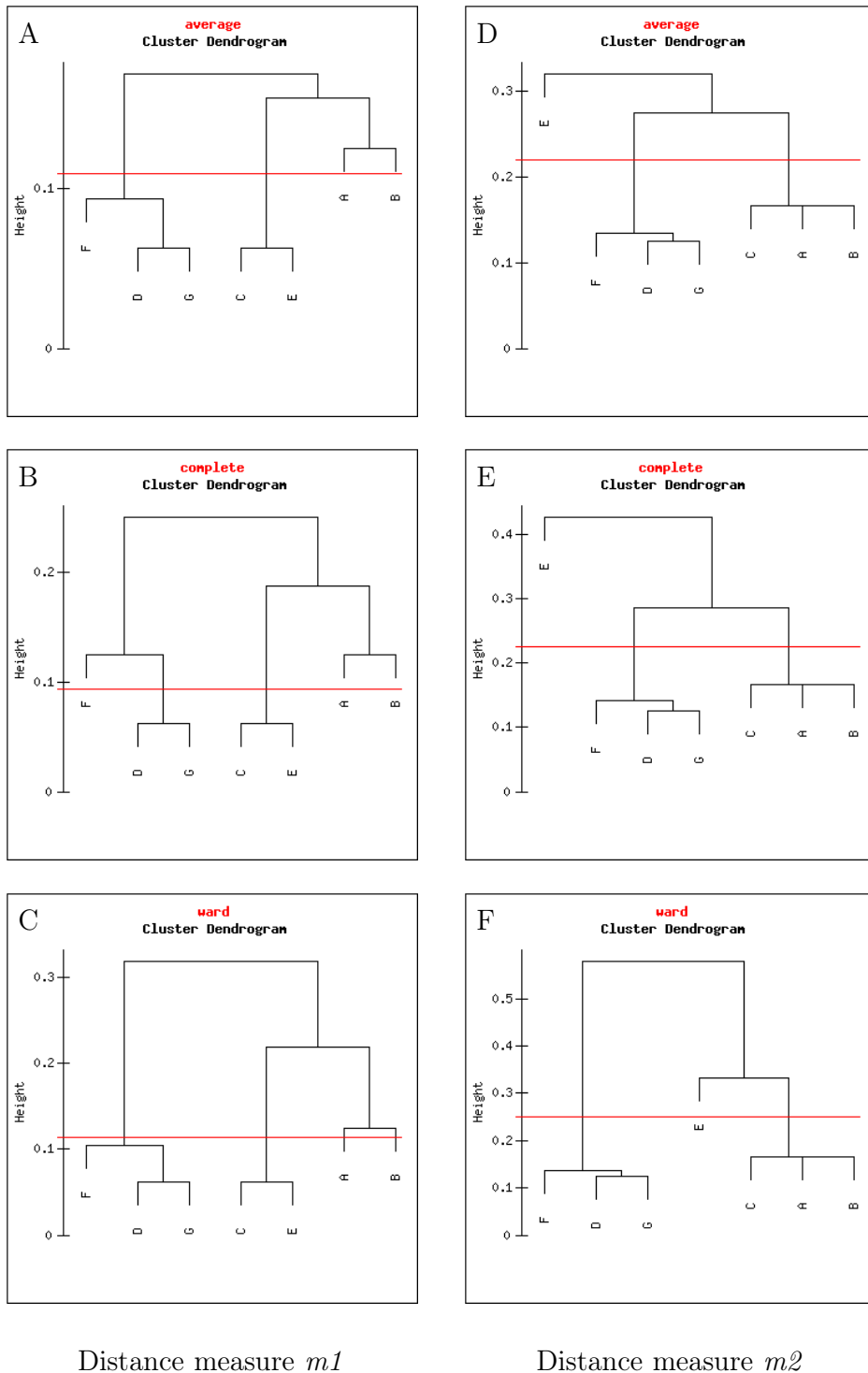


---

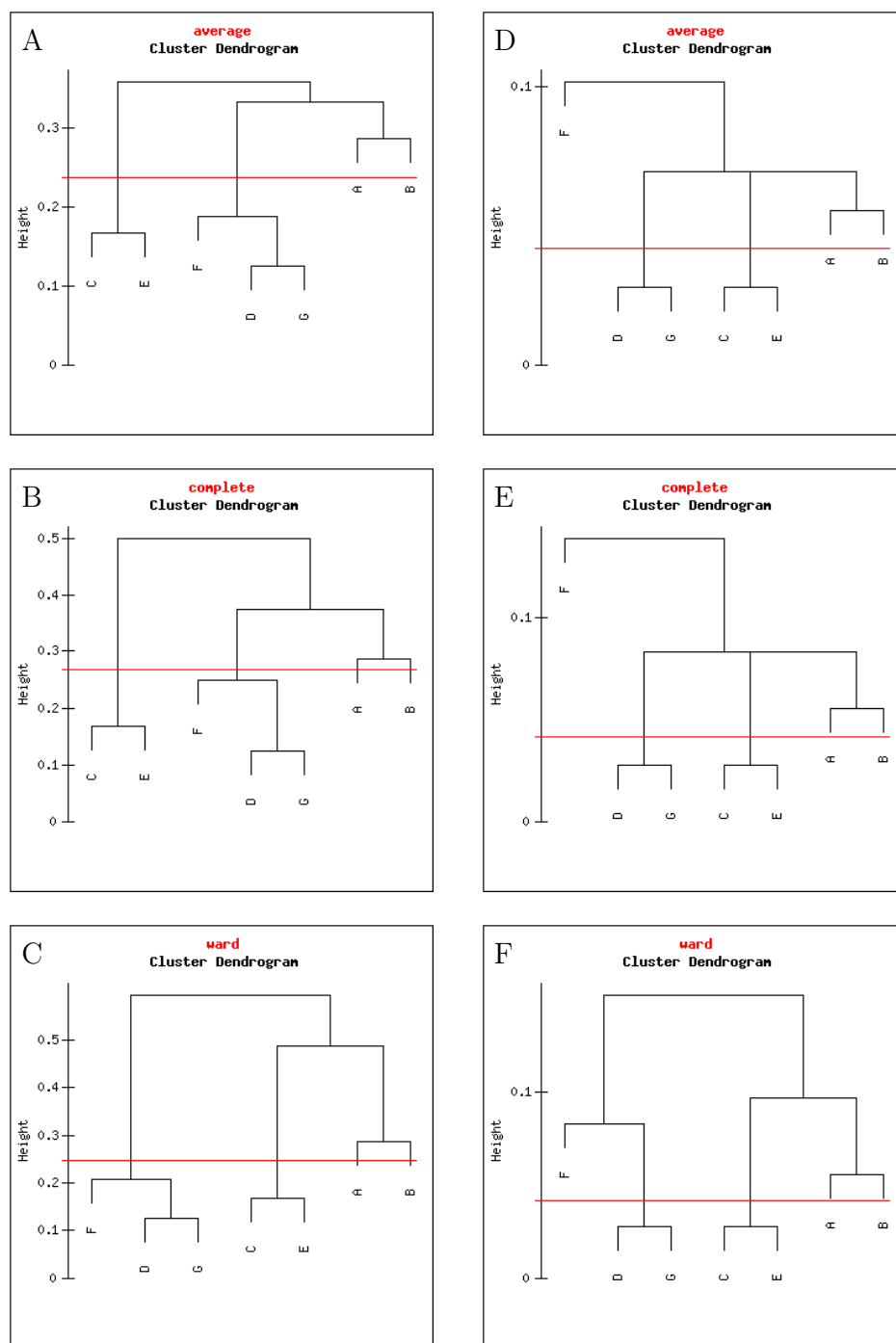
## Clustering Dendrograms

---

For evaluating the developed approach for metabolic network comparison two test scenarios were analyzed, namely a set of artificial organisms on an artificial pathway (see Section 6.1.1 on page 60) as well as a set of real organisms on a subpathway of KEGG's lysine biosynthesis pathway (see Section 6.1.2 on page 67). The analyses resulted in a clustering dendrogram for each combination of a distance measure and a clustering technique. Since these clustering dendrograms are the basis for the automatic classification and thus for the decision as to which distance measure and clustering technique are the best suited ones, these clustering dendrograms are presented here. Furthermore, the developed approach was applied to five *Corynebacteria* on all KEGG pathways as well as on the overall metabolic network of these organisms. Due to the large number of analyzed pathways it is not possible to include the clustering dendrograms of all analyzed pathways. However, since the clustering dendrograms for the overall metabolic network analysis are explicitly discussed (see Section 6.2.2.4 on page 85), they are presented here as well.



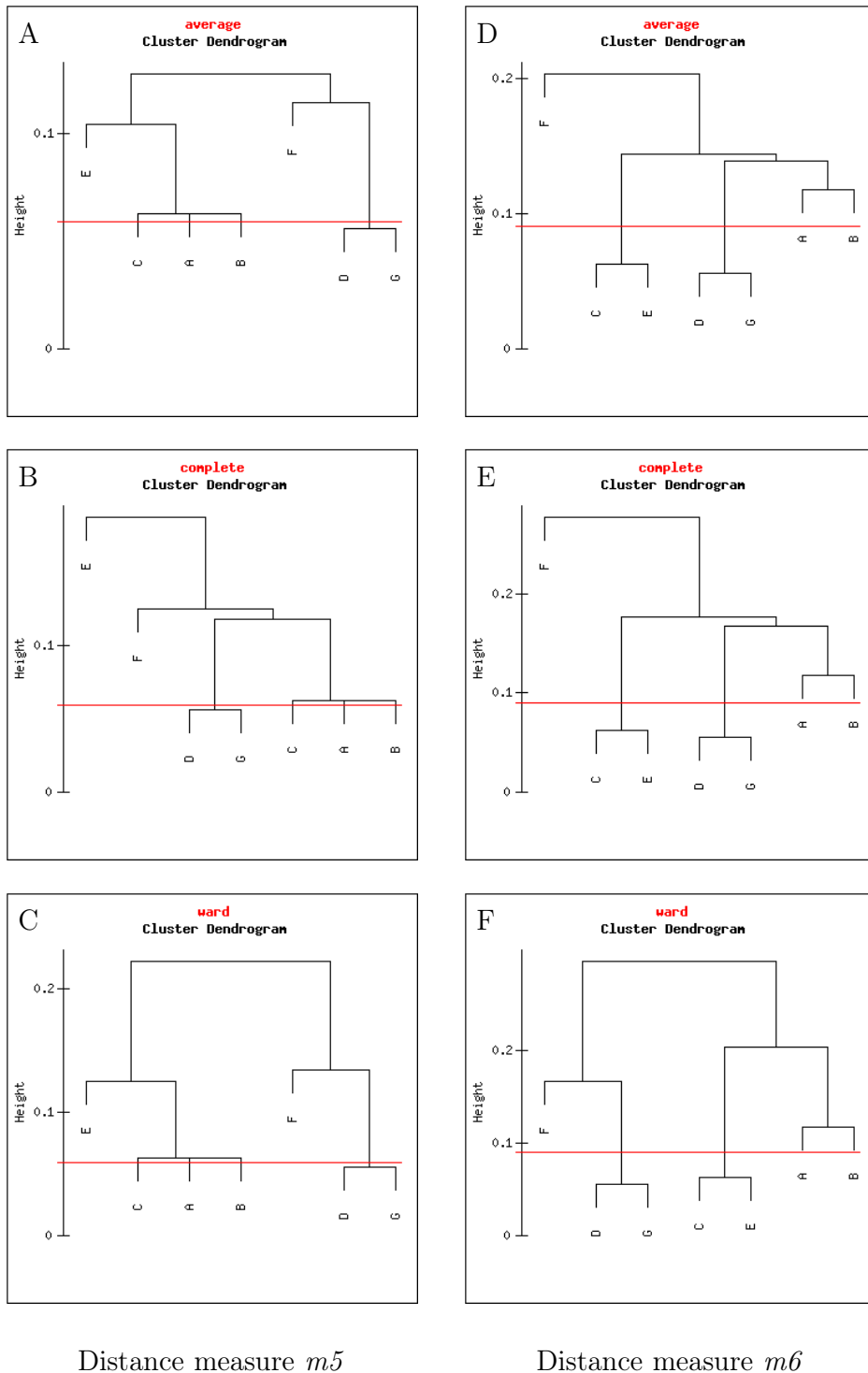
**Figure A.1.:** Clustering dendrograms of seven artificial organisms (A to G) based on the artificial test pathway for distance measure  $m1$  and average (A) and complete (B) linkage agglomerative clustering, as well as the Ward clustering method (C), and distance measure  $m2$  and average (D) and complete (E) linkage agglomerative clustering, as well as the Ward clustering method (F). The red line indicates where the dendrogram is cut in order to yield the classification.



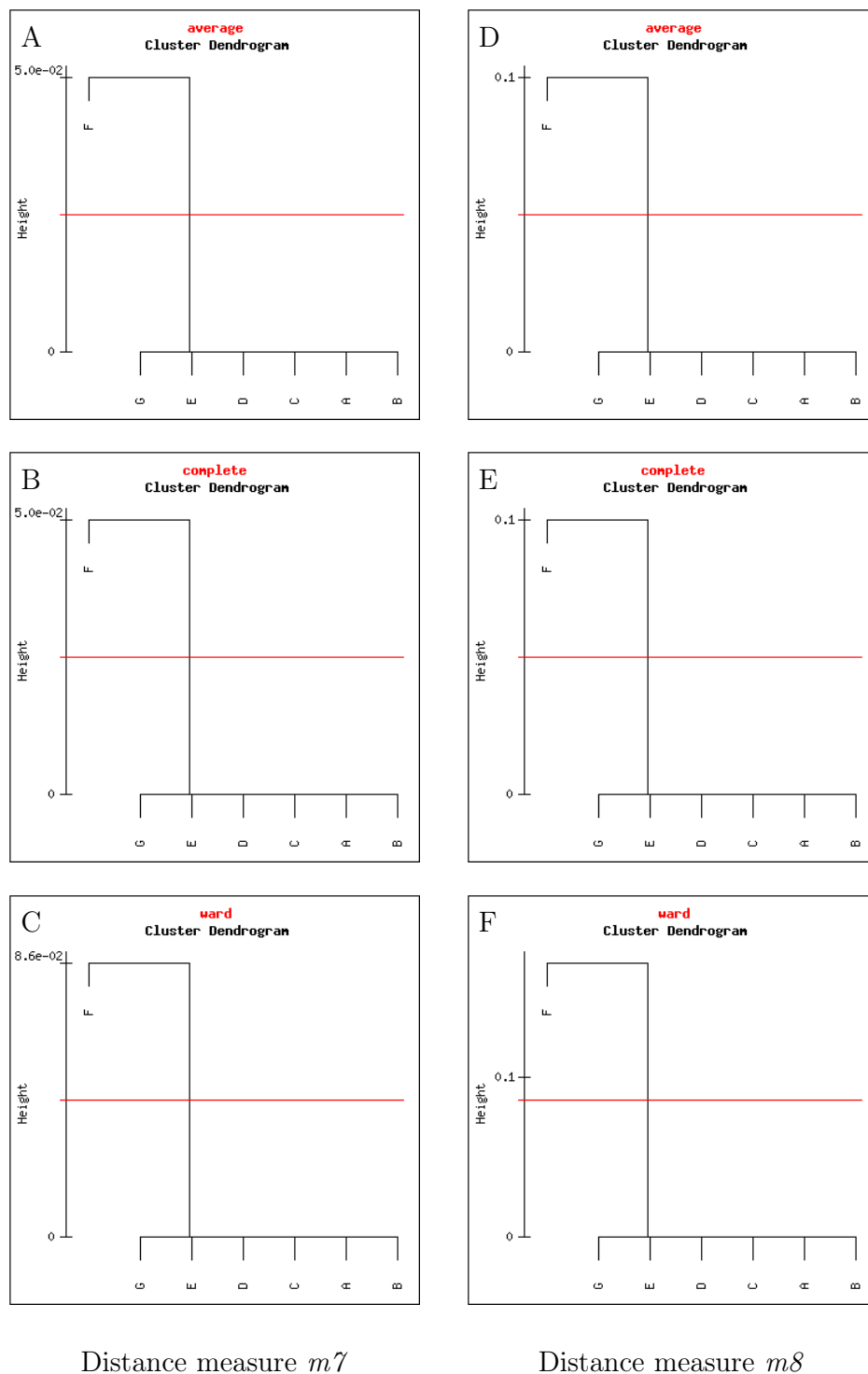
Distance measure  $m_3$

Distance measure  $m_4$

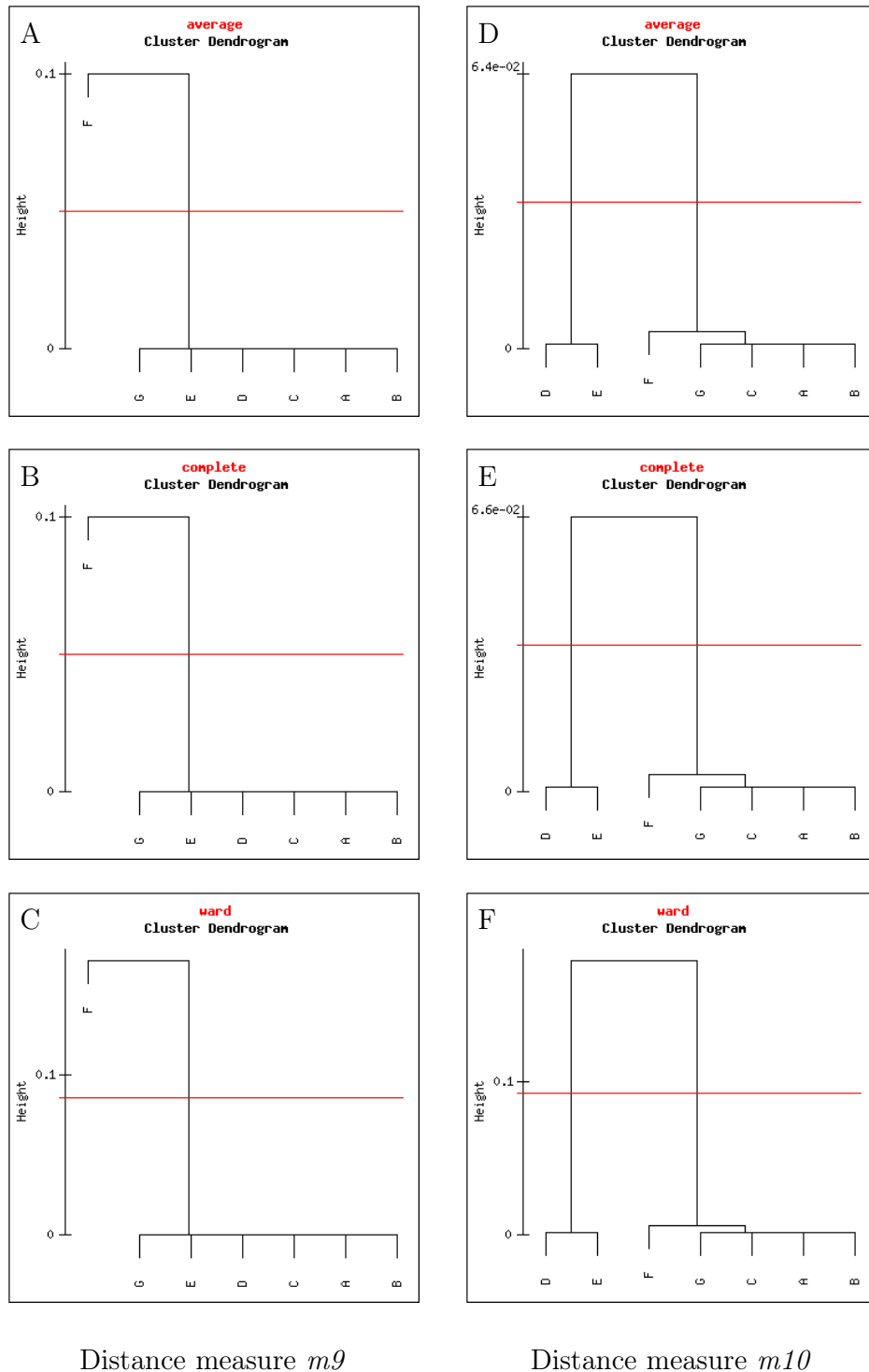
**Figure A.2.:** Clustering dendrograms of seven artificial organisms (A to G) based on the artificial test pathway for distance measure  $m_3$  and average (A) and complete (B) linkage agglomerative clustering, as well as the Ward clustering method (C), and distance measure  $m_4$  and average (D) and complete (E) linkage agglomerative clustering, as well as the Ward clustering method (F). The red line indicates where the dendrogram is cut in order to yield the classification.



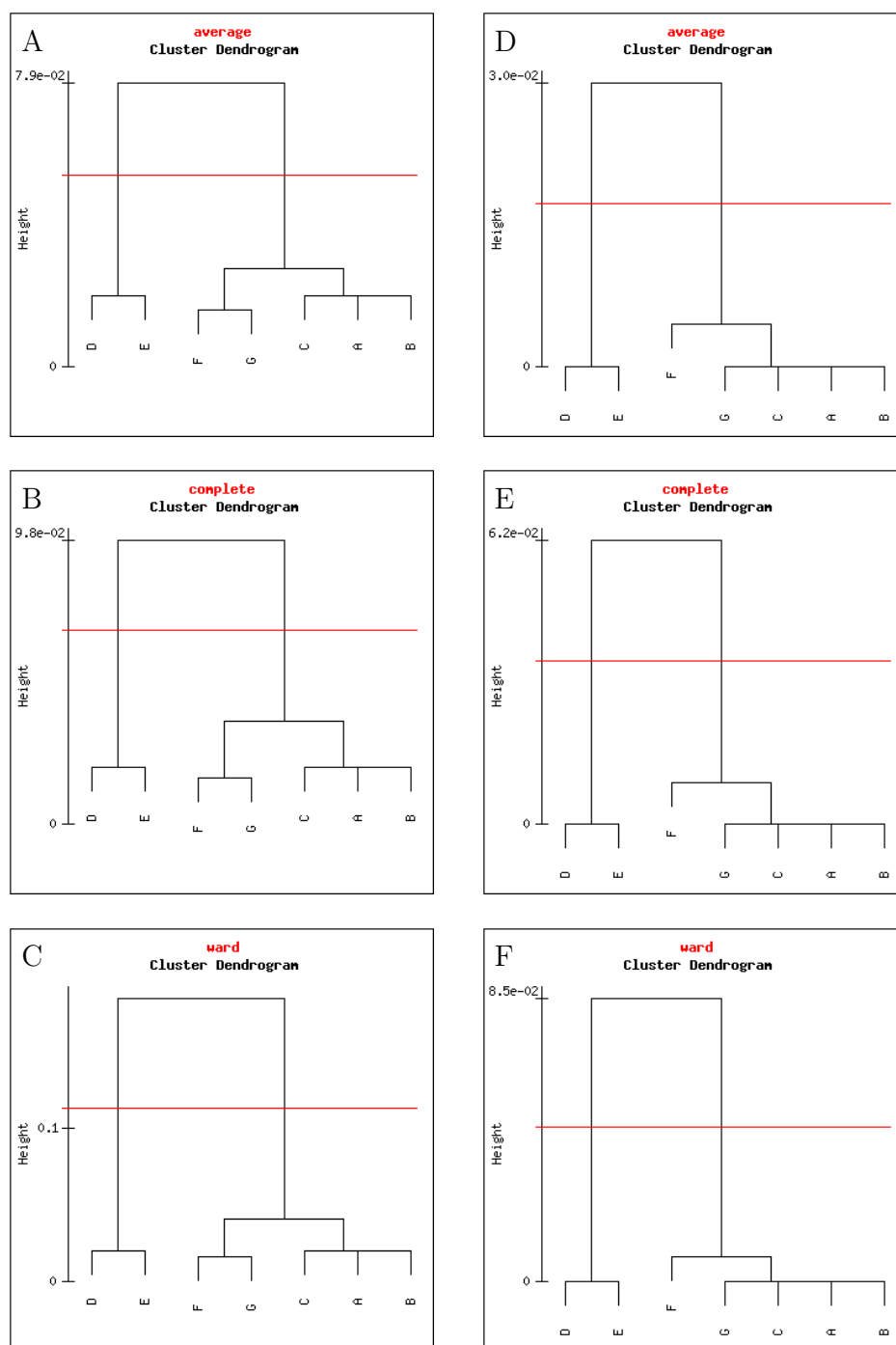
**Figure A.3.:** Clustering dendrograms of seven artificial organisms (A to G) based on the artificial test pathway for distance measure  $m5$  and average (A) and complete (B) linkage agglomerative clustering, as well as the Ward clustering method (C), and distance measure  $m6$  and average (D) and complete (E) linkage agglomerative clustering, as well as the Ward clustering method (F). The red line indicates where the dendrogram is cut in order to yield the classification.



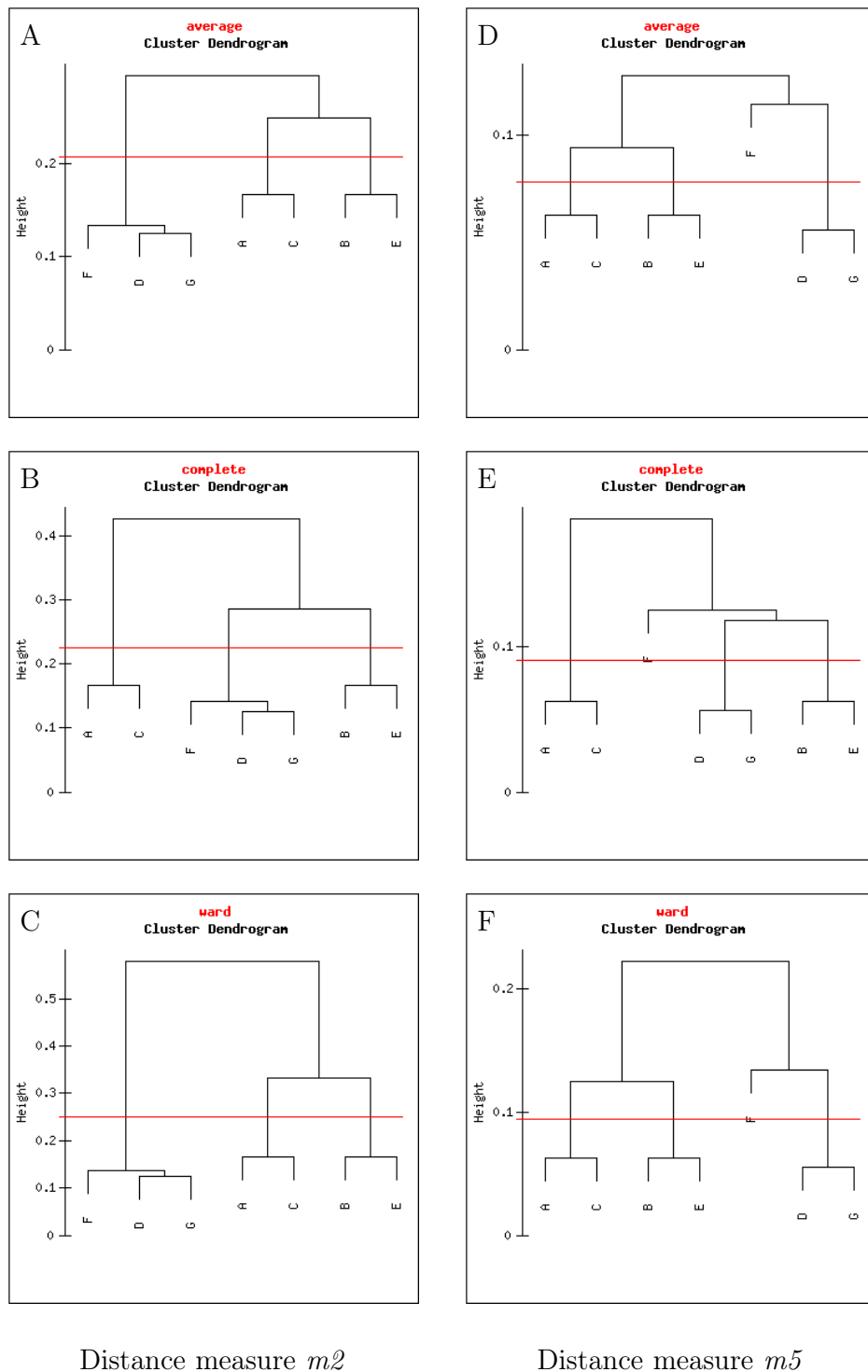
**Figure A.4.:** Clustering dendrograms of seven artificial organisms (A to G) based on the artificial test pathway for distance measure  $m7$  and average (A) and complete (B) linkage agglomerative clustering, as well as the Ward clustering method (C), and distance measure  $m8$  and average (D) and complete (E) linkage agglomerative clustering, as well as the Ward clustering method (F). The red line indicates where the dendrogram is cut in order to yield the classification.



**Figure A.5.:** Clustering dendrograms of seven artificial organisms (A to G) based on the artificial test pathway for distance measure  $m_9$  and average (A) and complete (B) linkage agglomerative clustering, as well as the Ward clustering method (C), and distance measure  $m_{10}$  and average (D) and complete (E) linkage agglomerative clustering, as well as the Ward clustering method (F). The red line indicates where the dendrogram is cut in order to yield the classification.

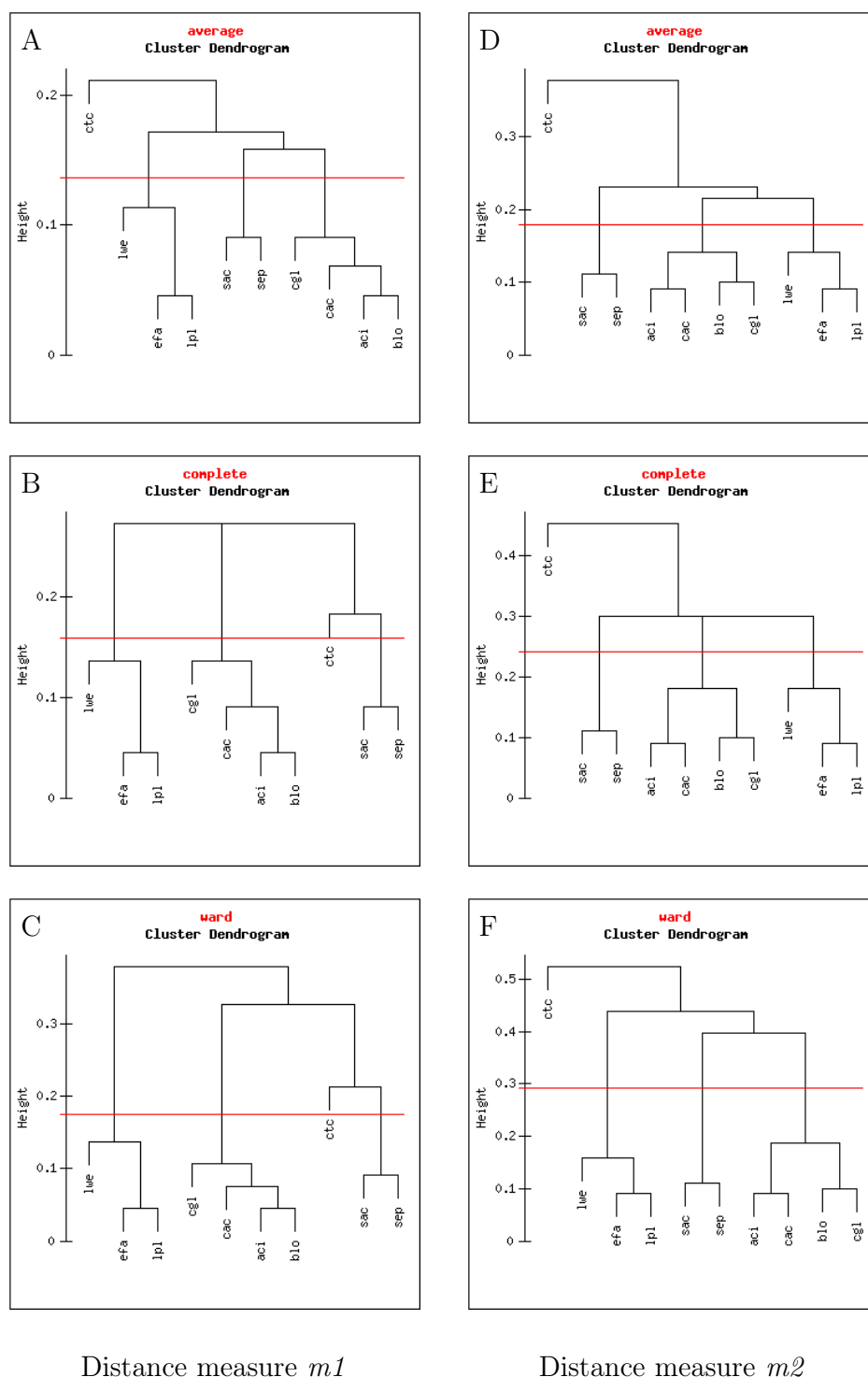
Distance measure  $m11$ Distance measure  $m12$ 

**Figure A.6.:** Clustering dendrograms of seven artificial organisms (A to G) based on the artificial test pathway for distance measure  $m11$  and average (A) and complete (B) linkage agglomerative clustering, as well as the Ward clustering method (C), and distance measure  $m12$  and average (D) and complete (E) linkage agglomerative clustering, as well as the Ward clustering method (F). The red line indicates where the dendrogram is cut in order to yield the classification.

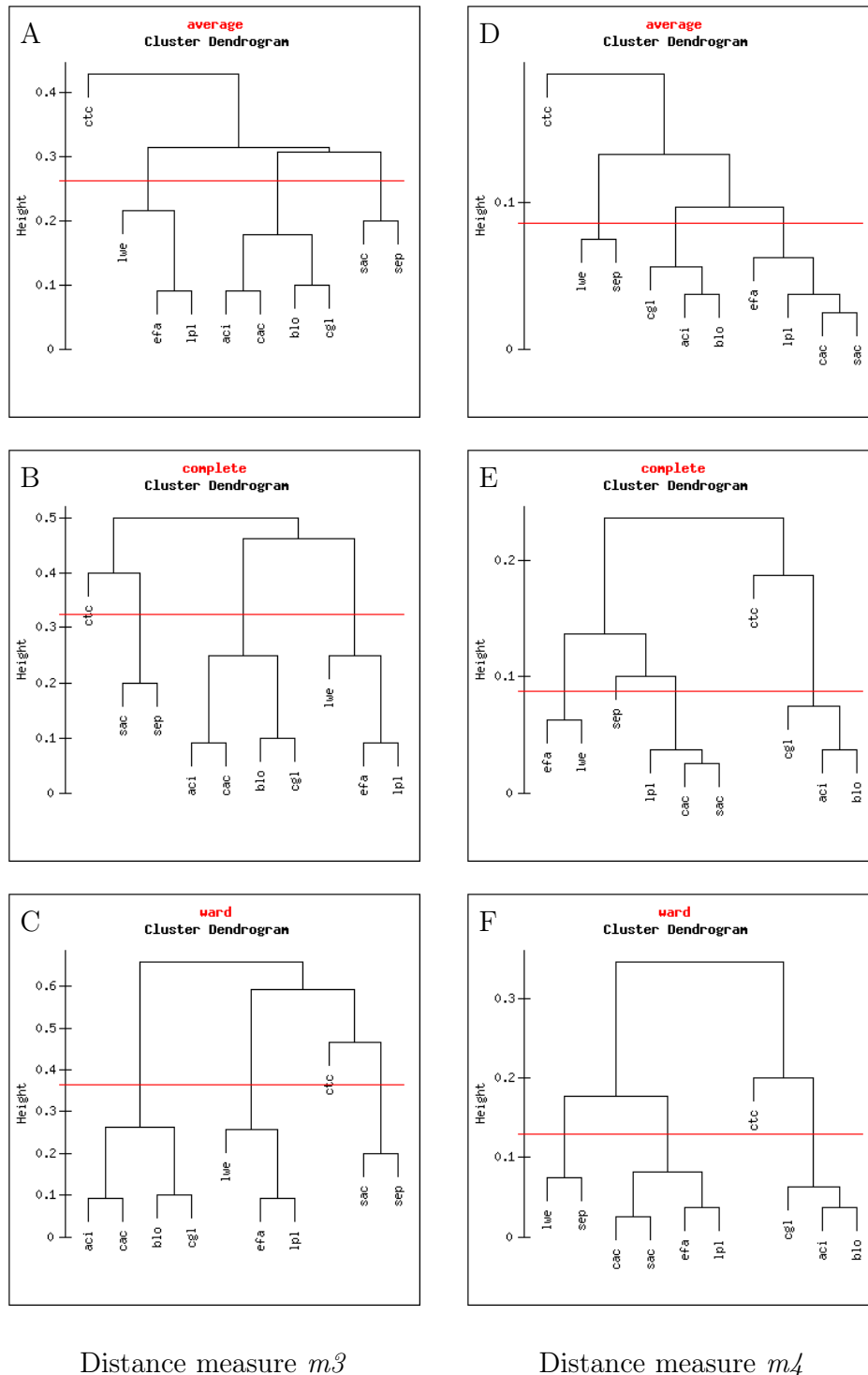


**Figure A.7.:** Clustering dendrograms of seven artificial organisms (A to G) based on the artificial test pathway for distance measure  $m_2$  and average (A) and complete (B) linkage agglomerative clustering, as well as the Ward clustering method (C), and distance measure  $m_5$  and average (D) and complete (E) linkage agglomerative clustering, as well as the Ward clustering method (F). Names of pseudo-organisms  $A$  and  $E$  are exchanged. The red line indicates where the dendrogram is cut in order to yield the classification.

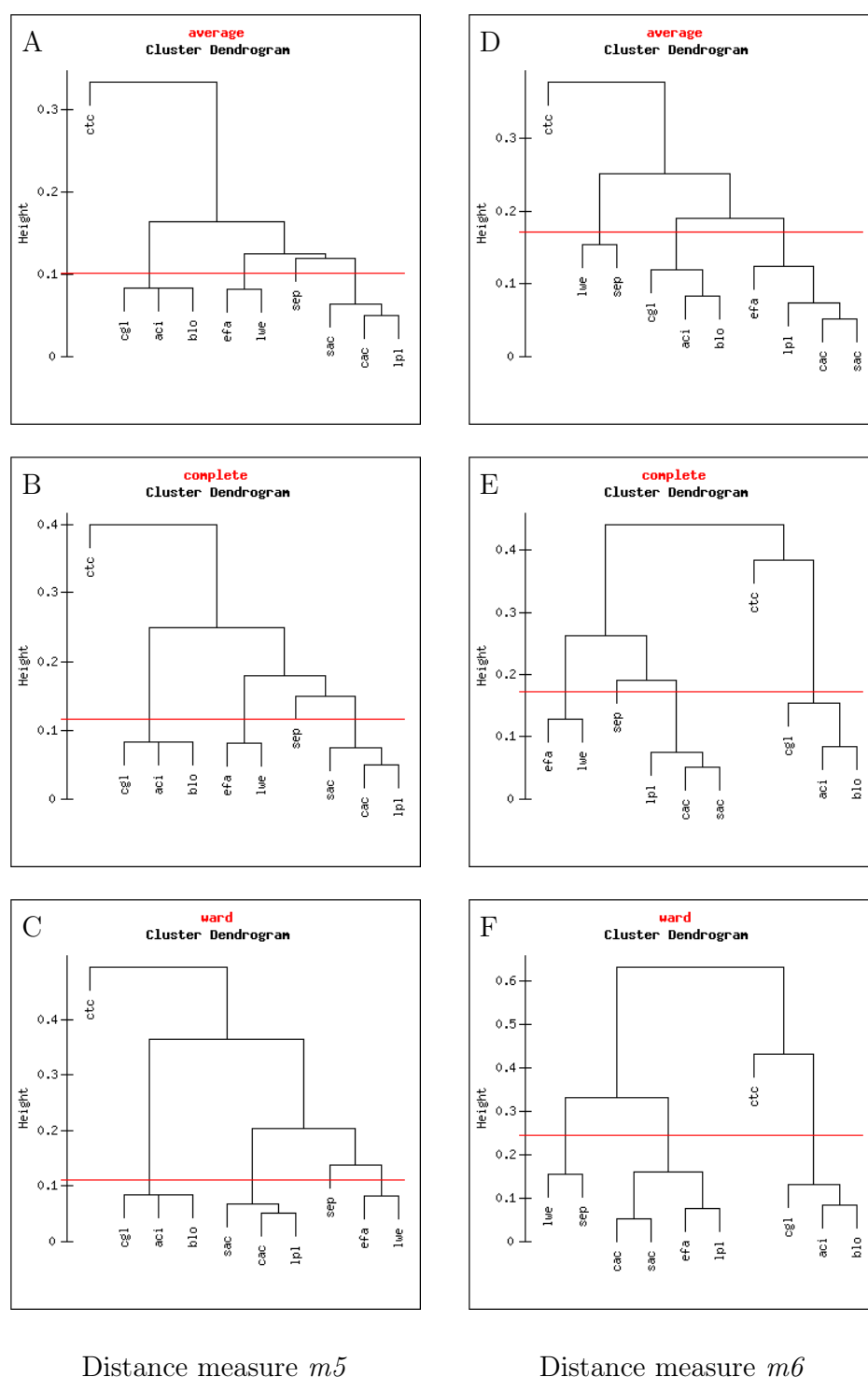


Distance measure  $m_1$ Distance measure  $m_2$ 

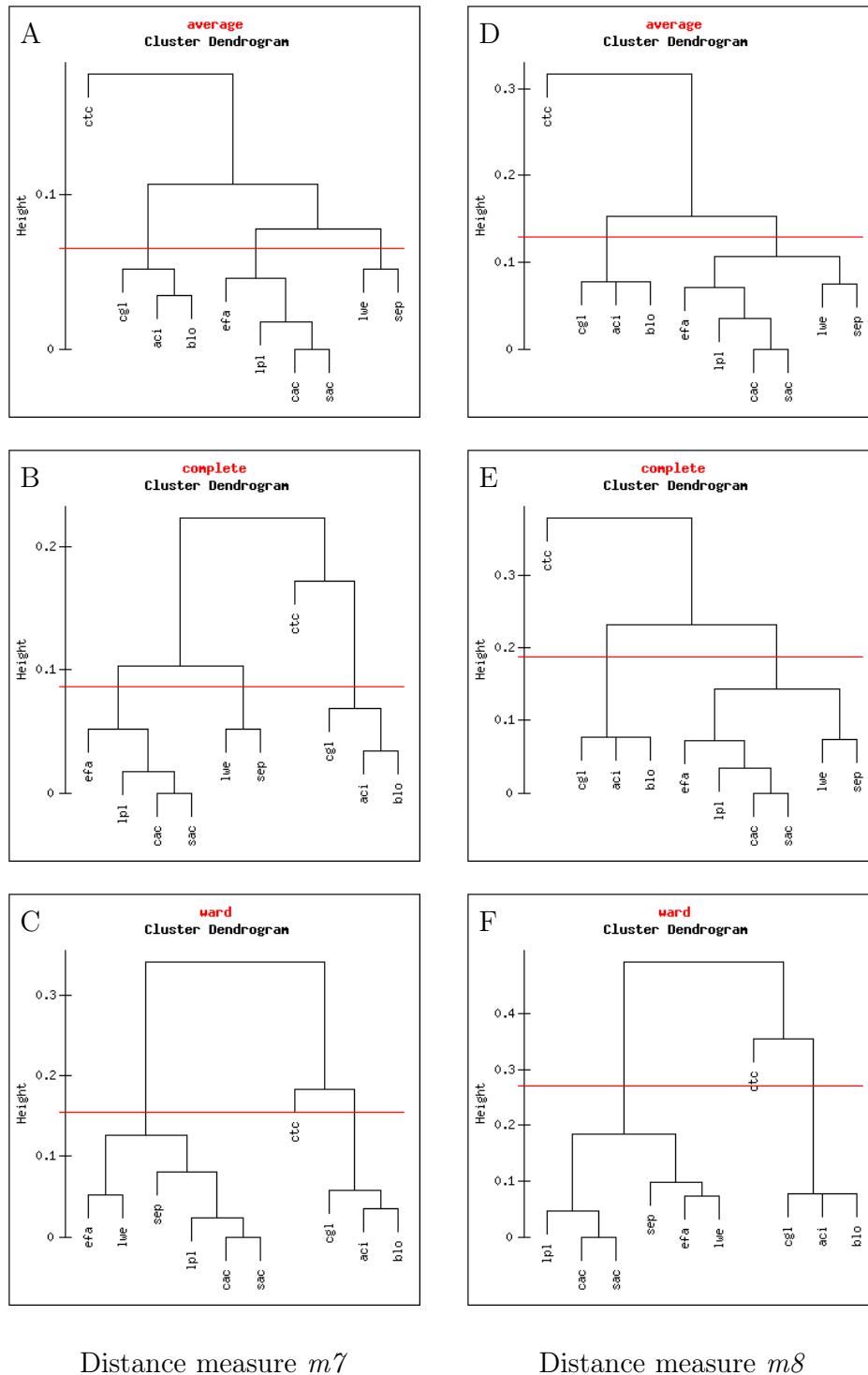
**Figure A.8.:** Clustering dendrograms of various organisms based on a subpathway of the KEGG lysine biosynthesis for distance measure  $m_1$  and average (A) and complete (B) linkage agglomerative clustering, as well as the Ward clustering method (C), and distance measure  $m_2$  and average (D) and complete (E) linkage agglomerative clustering, as well as the Ward clustering method (F). The red line indicates where the dendrogram is cut in order to yield the classification. aci: *Acinetobacter*, blo: *Bifidobacterium longum*, cac: *Clostridium acetobutylicum*, ctc: *Clostridium tetani*, cgl: *Corynebacterium glutamicum*, efa: *Enterococcus faecalis*, lpl: *Lactobacillus plantarum*, lwe: *Listeria welshimeri*, sac: *Staphylococcus aureus*, sep: *Staphylococcus epidermidis*.



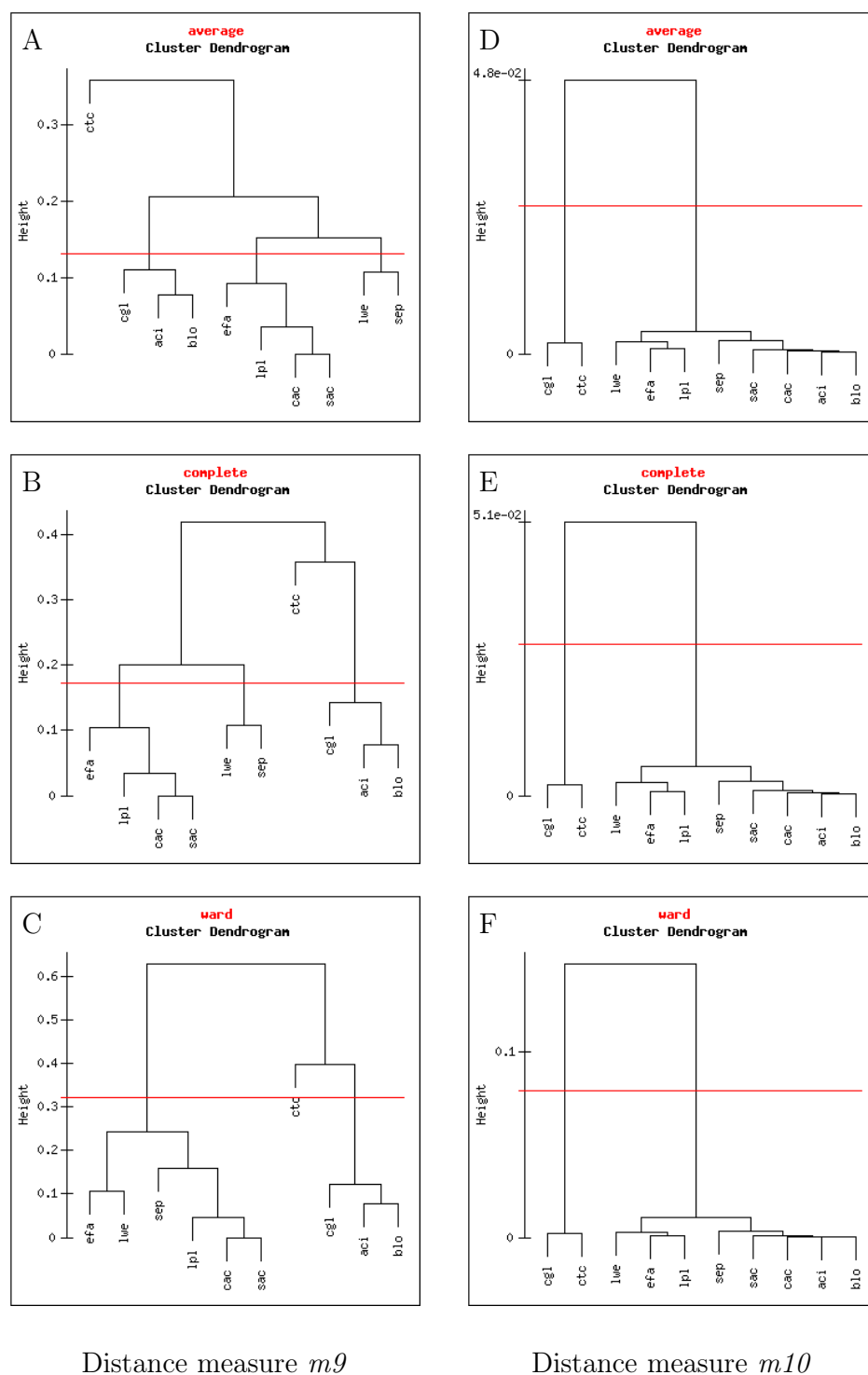
**Figure A.9.:** Clustering dendrograms of various organisms based on a subpathway of the KEGG lysine biosynthesis for distance measure  $m_3$  and average (A) and complete (B) linkage agglomerative clustering, as well as the Ward clustering method (C), and distance measure  $m_4$  and average (D) and complete (E) linkage agglomerative clustering, as well as the Ward clustering method (F). The red line indicates where the dendrogram is cut in order to yield the classification. aci: *Acinetobacter*, blo: *Bifidobacterium longum*, cac: *Clostridium acetobutylicum*, ctc: *Clostridium tetani*, cgl: *Corynebacterium glutamicum*, efa: *Enterococcus faecalis*, lp1: *Lactobacillus plantarum*, lwe: *Listeria welshimeri*, sac: *Staphylococcus aureus*, sep: *Staphylococcus epidermidis*.

Distance measure  $m_5$ Distance measure  $m_6$ 

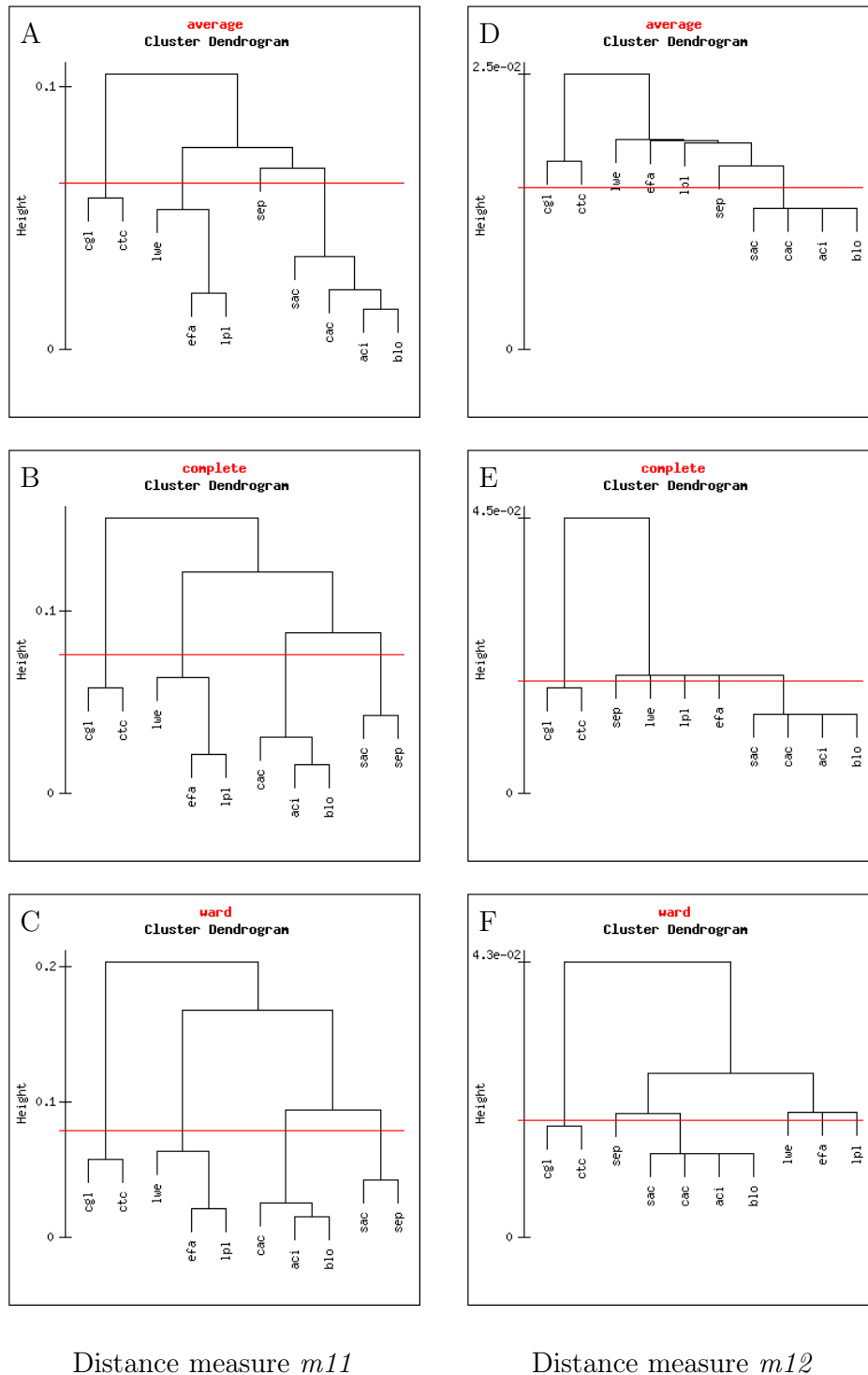
**Figure A.10.:** Clustering dendrograms of various organisms based on a subpathway of the KEGG lysine biosynthesis for distance measure  $m_5$  and average (A) and complete (B) linkage agglomerative clustering, as well as the Ward clustering method (C), and distance measure  $m_6$  and average (D) and complete (E) linkage agglomerative clustering, as well as the Ward clustering method (F). The red line indicates where the dendrogram is cut in order to yield the classification. aci: *Acinetobacter*, blo: *Bifidobacterium longum*, cac: *Clostridium acetobutylicum*, ctc: *Clostridium tetani*, cgl: *Corynebacterium glutamicum*, efa: *Enterococcus faecalis*, lpl: *Lactobacillus plantarum*, lwe: *Listeria welshimeri*, sac: *Staphylococcus aureus*, sep: *Staphylococcus epidermidis*.



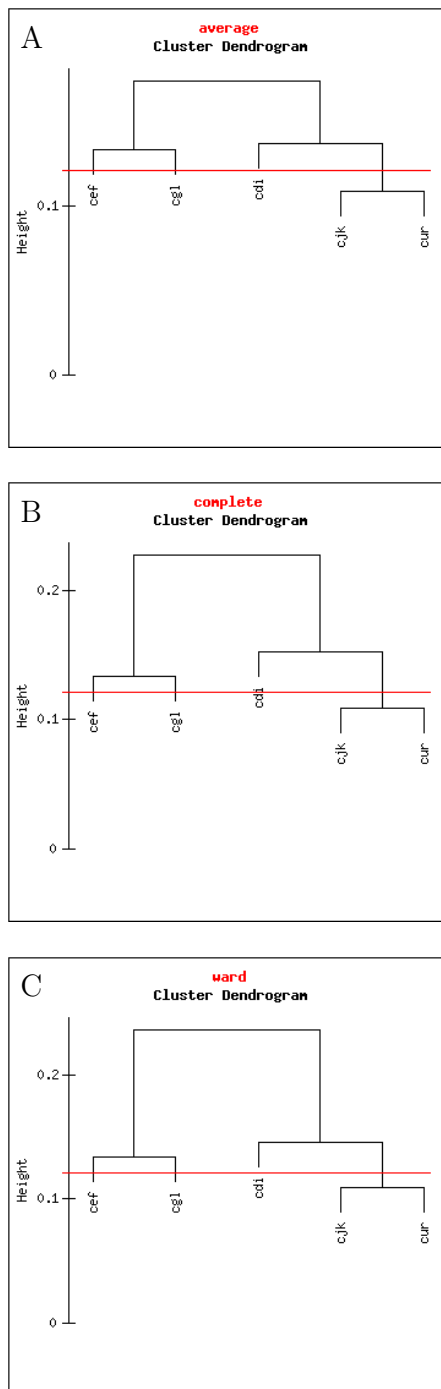
**Figure A.11.:** Clustering dendrograms of various organisms based on a subpathway of the KEGG lysine biosynthesis for distance measure  $m7$  and average (A) and complete (B) linkage agglomerative clustering, as well as the Ward clustering method (C), and distance measure  $m8$  and average (D) and complete (E) linkage agglomerative clustering, as well as the Ward clustering method (F). The red line indicates where the dendrogram is cut in order to yield the classification. aci: *Acinetobacter*, blo: *Bifidobacterium longum*, cac: *Clostridium acetobutylicum*, ctc: *Clostridium tetani*, cgl: *Corynebacterium glutamicum*, efa: *Enterococcus faecalis*, lpl: *Lactobacillus plantarum*, lwe: *Listeria welshimeri*, sac: *Staphylococcus aureus*, sep: *Staphylococcus epidermidis*.



**Figure A.12.:** Clustering dendrograms of various organisms based on a subpathway of the KEGG lysine biosynthesis for distance measure  $m9$  and average (A) and complete (B) linkage agglomerative clustering, as well as the Ward clustering method (C), and distance measure  $m10$  and average (D) and complete (E) linkage agglomerative clustering, as well as the Ward clustering method (F). The red line indicates where the dendrogram is cut in order to yield the classification. aci: *Acinetobacter*, blo: *Bifidobacterium longum*, cac: *Clostridium acetobutylicum*, ctc: *Clostridium tetani*, cgl: *Corynebacterium glutamicum*, efa: *Enterococcus faecalis*, lpl: *Lactobacillus plantarum*, lwe: *Listeria welshimeri*, sac: *Staphylococcus aureus*, sep: *Staphylococcus epidermidis*.



**Figure A.13.:** Clustering dendrograms of various organisms based on a subpathway of the KEGG lysine biosynthesis for distance measure  $m_{11}$  and average (A) and complete (B) linkage agglomerative clustering, as well as the Ward clustering method (C), and distance measure  $m_{12}$  and average (D) and complete (E) linkage agglomerative clustering, as well as the Ward clustering method (F). The red line indicates where the dendrogram is cut in order to yield the classification. aci: *Acinetobacter*, blo: *Bifidobacterium longum*, cac: *Clostridium acetobutylicum*, ctc: *Clostridium tetani*, cgl: *Corynebacterium glutamicum*, efa: *Enterococcus faecalis*, lp1: *Lactobacillus plantarum*, lwe: *Listeria welshimeri*, sac: *Staphylococcus aureus*, sep: *Staphylococcus epidermidis*.



Distance measure  $m1$

**Figure A.14.:** Clustering dendrograms of five Corynebacteria based on the overall metabolic network for distance measure  $m1$  and average (A) and complete (B) linkage agglomerative clustering as well as the Ward clustering method (C). The red line indicates where the dendrogram is cut in order to yield the classification. cdi: *C. diphtheriae*, cef: *C. efficiens*, cgl: *C. glutamicum*, cjk: *C. jeikeium*, cur: *C. urealyticum*.





---

## Bibliography

---

- Alfarano C., Andrade C. E., Anthony K., Bahroos N., Bajec M., Bantoft K., Betel D., Bobechko B., Boutilier K., Burgess E., Buzadzija K., Cavero R., D'Abreo C., Donaldson I., Dorairajoo D., Dumontier M. J., Dumontier M. R., Earles V., Farrall R., Feldman H., Garderman E., Gong Y., Gonzaga R., Grytsan V., Gryz E., Gu V., Haldorsen E., Halupa A., Haw R., Hrvojic A., Hurrell L., Isserlin R., Jack F., Juma F., Khan A., Kon T., Konopinsky S., Le V., Lee E., Ling S., Magidin M., Moniakis J., Montojo J., Moore S., Muskat B., Ng I., Paraiso J. P., Parker B., Pintilie G., Pirone R., Salama J. J., Sgro S., Shan T., Shu Y., Siew J., Skinner D., Snyder K., Stasiuk R., Strumpf D., Tuekam B., Tao S., Wang Z., White M., Willis R., Wolting C., Wong S., Wrong A., Xin C., Yao R., Yates B., Zhang S., Zheng K., Pawson T., Ouellette B. F. F., Hogue C. W. V.: The Biomolecular Interaction Network Database and related tools 2005 update. *Nucleic Acids Research*, 33(Database issue):D418–D424, (2005).
- Andrews S. C., Robinson A. K., Rodríguez-Quiñones F.: Bacterial iron homeostasis. *FEMS Microbiology Reviews*, 27(2-3):215–237, (2003).
- Bader G. D., Betel D., Hogue C. W. V.: BIND: the Biomolecular Interaction Network Database. *Nucleic Acids Research*, 31(1):248–250, (2003).
- Barthelmes J., Ebeling C., Chang A., Schomburg I., Schomburg D.: BRENDA, AMENDA and FRENDA: the enzyme information system in 2007. *Nucleic Acids Research*, 35(Database issue):D511–D514, (2007).
- Borenstein E., Kupiec M., Feldman M. W., Ruppin E.: Large-scale reconstruction and phylogenetic analysis of metabolic environments. *Proceedings of the National Academy of Sciences of the United States of America*, 105(38):14482–14487, (2008).
- Bunke H.: On a relation between graph edit distance and maximum common subgraph. *Pattern Recognition Letters*, 18:689–694, (1997).
- Bunke H.: Error correcting graph matching: On the influence of the underlying cost function. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(9):917–922, (1999).

- Bunke H., Shearer K.: A graph distance metric based on the maximal common subgraph. *Pattern Recognition Letters*, 19(3-4):255–259, (1998).
- Caspi R., Foerster H., Fulcher C. A., Kaipa P., Krummenacker M., Latendresse M., Paley S., Rhee S. Y., Shearer A. G., Tissier C., Walk T. C., Zhang P., Karp P. D.: The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Research*, 36(Database issue):D623–D631, (2008).
- Cerdeño-Tárraga A. M., Efstratiou A., Dover L. G., Holden M. T. G., Pallen M., Bentley S. D., Besra G. S., Churcher C., James K. D., Zoysa A. D., Chillingworth T., Cronin A., Dowd L., Feltwell T., Hamlin N., Holroyd S., Jagels K., Moule S., Quail M. A., Rabinowitsch E., Rutherford K. M., Thomson N. R., Unwin L., Whitehead S., Barrell B. G., Parkhill J.: The complete genome sequence and analysis of *Corynebacterium diphtheriae* NCTC13129. *Nucleic Acids Research*, 31(22):6516–6523, (2003).
- Cox D., James A., Taylor D.: Cosmetic composition. *United States Patent Application 20040180012 (17/12/2003)*.
- Dandekar T., Schuster S., Snel B., Huynen M., Bork P.: Pathway alignment: application to the comparative analysis of glycolytic enzymes. *The Biochemical Journal*, 343(Pt 1):115–124, (1999).
- Day W. H., Edelsbrunner H.: Efficient algorithms for agglomerative hierarchical clustering methods. *Journal of Classification*, 1(1):7–24, (1984).
- Ding C., He X.: K-nearest-neighbor consistency in data clustering: incorporating local information into global optimization. In: *Proceedings of the 2004 ACM Symposium on Applied Computing (SAC 2004)*, Nicosia, Cyprus, pages 584–589, ACM, New York (2004).
- Duran B. S., Odell P. L.: *Cluster Analysis*. Springer Verlag, Berlin (1974).
- Eckes T., Roßbach H.: *Clusteranalysen*. Verlag W. Kohlhammer, Stuttgart (1980).
- Ester M., Kriegel H.-P., Jörg S., Xu X.: A density-based algorithm for discovering clusters in large spatial databases with noise. In: E. Simoudis, J. Han, U. M. Fayyad, eds., *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD 1996)*, Portland, Oregon, pages 226–231, AAAI Press, California (1996).
- Fernández M.-L., Valiente G.: A graph distance metric combining maximum common subgraph and minimum common supergraph. *Pattern Recognition Letters*, 22(6-7):753–758, (2001).
- Forst C., Flamm C., Hofacker I., Stadler P.: Algebraic comparison of metabolic networks, phylogenetic inference, and metabolic innovation. *BMC Bioinformatics*, 7(1):67, (2006).
- Forst C. V., Schulten K.: Phylogenetic analysis of metabolic pathways. *Journal of Molecular Evolution*, 52(6):471–489, (2001).

- Galperin M. Y., Walker D. R., Koonin E. V.: Analogous enzymes: independent inventions in enzyme evolution. *Genome Research*, 8(8):779–790, (1998).
- Gansner E. R., North S. C.: An open graph visualization system and its applications to software engineering. *Software - Practice and Experience*, 30(11):1203–1233, (2000).
- Gasteiger E., Gattiker A., Hoogland C., Ivanyi I., Appel R. D., Bairoch A.: ExpASY: the proteomics server for in-depth protein knowledge and analysis. *Nucleic Acids Research*, 31(13):3784–3788, (2003).
- Gilbert D.: Biomolecular interaction network database. *Briefings in Bioinformatics*, 6(2):194–198, (2005).
- Gill S. R., Pop M., Deboy R. T., Eckburg P. B., Turnbaugh P. J., Samuel B. S., Gordon J. I., Relman D. A., Fraser-Liggett C. M., Nelson K. E.: Metagenomic analysis of the human distal gut microbiome. *Science*, 312(5778):1355–1359, (2006).
- Goesmann A., Haubrock M., Meyer F., Kalinowski J., Giegerich R.: PathFinder: reconstruction and dynamic visualization of metabolic pathways. *Bioinformatics*, 18(1):124–129, (2002).
- Goto S., Bono H., Ogata H., Fujibuchi W., Nishioka T., Sato K., Kanehisa M.: Organizing and computing metabolic pathway data in terms of binary relations. *Pacific Symposium on Biocomputing*, 2:175–186, (1997).
- Green M. L., Karp P. D.: A bayesian method for identifying missing enzymes in predicted metabolic pathway databases. *BMC Bioinformatics*, 5(1):76, (2004).
- Halkidi M., Batistakis Y., Vazirgiannis M.: On clustering validation techniques. *Journal of Intelligent Information Systems*, 17(2/3):107–145, (2001).
- Handl J., Knowles J.: Exploiting the trade-off—the benefits of multiple objectives in data clustering. In: C. A. C. Coello, A. H. Aguirre, E. Zitzler, eds., *Proceedings of the Third International Conference on Evolutionary Multicriterion Optimization (EMO 2005), Guanajuato, Mexico*, vol. 3410 of *Lecture Notes in Computer Science*, pages 547–560, Springer Verlag, Berlin (2005).
- Handl J., Knowles J., Kell D. B.: Computational cluster validation in post-genomic data analysis. *Bioinformatics*, 21(15):3201–3212, (2005).
- Hansen S. K., Rainey P. B., Haagenen J. A. J., Molin S.: Evolution of species interactions in a biofilm community. *Nature*, 445(7127):533–536, (2007).
- Hastie T., Tibshirani R., Friedman J.: *The Elements of Statistical Learning*. Springer Verlag, Berlin (2001).
- Heiner M., Koch I.: Petri net based model validation in systems biology. In: J. Cortadella, W. Reisig, eds., *Proceedings of the 25th International Conference on Applications and Theory of Petri Nets (ICATPN 2004), Bologna, Italy*, vol. 3099 of *Lecture Notes in Computer Science*, pages 216–237, Springer Verlag, Berlin (2004).

- Heymans M., Singh A. K.: Deriving phylogenetic trees from the similarity analysis of metabolic pathways. *Bioinformatics*, 19(Suppl 1):i138–i146, (2003).
- Holyoak T., Sullivan S. M., Nowak T.: Structural insights into the mechanism of PEPCK catalysis. *Biochemistry*, 45(27):8254–8263, (2006).
- Hong S. H., Kim T. Y., Lee S. Y.: Phylogenetic analysis based on genome-scale metabolic pathway reaction content. *Applied Microbiology and Biotechnology*, 65(2):203–210, (2004).
- Jain A. K., Dubes R. C.: *Algorithms for Clustering Data*. Prentice Hall, New Jersey (1988).
- Jain A. K., Murty M. N., Flynn P. J.: Data clustering: a review. *ACM Computing Surveys*, 31(3):264–323, (1999).
- Kai Y., Matsumura H., Izui K.: Phosphoenolpyruvate carboxylase: three-dimensional structure and molecular mechanisms. *Archives of Biochemistry and Biophysics*, 414(2):170–179, (2003).
- Kalinowski J., Bathe B., Bartels D., Bischoff N., Bott M., Burkovski A., Dusch N., Eggeling L., Eikmanns B. J., Gaigalat L., Goesmann A., Hartmann M., Huthmacher K., Krämer R., Linke B., McHardy A. C., Meyer F., Möckel B., Pfefferle W., Pühler A., Rey D. A., Rückert C., Rupp O., Sahn H., Wendisch V. F., Wiegräbe I., Tauch A.: The complete *Corynebacterium glutamicum* ATCC 13032 genome sequence and its impact on the production of L-aspartate-derived amino acids and vitamins. *Journal of Biotechnology*, 104(1-3):5–25, (2003).
- Kalyuzhnaya M. G., Lapidus A., Ivanova N., Copeland A. C., McHardy A. C., Szeto E., Salamov A., Grigoriev I. V., Suci D., Levine S. R., Markowitz V. M., Rigoutsos I., Tringe S. G., Bruce D. C., Richardson P. M., Lidstrom M. E., Chistoserdova L.: High-resolution metagenomics targets specific functional types in complex microbial communities. *Nature Biotechnology*, 26(9):1029–1034, (2008).
- Kanehisa M.: Toward pathway engineering: a new database of genetic and molecular pathways. *Science & Technology Japan*, 59:34–38, (1996).
- Kanehisa M.: A database for post-genome analysis. *Trends in Genetics*, 13(9):375–376, (1997).
- Kanehisa M., Araki M., Goto S., Hattori M., Hirakawa M., Itoh M., Katayama T., Kawashima S., Okuda S., Tokimatsu T., Yamanishi Y.: KEGG for linking genomes to life and the environment. *Nucleic Acids Research*, 36(Database issue):D480–D484, (2008).
- Kanehisa M., Goto S.: KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, 28(1):27–30, (2000).
- Karp P., Arnaud M., Collado-Vides J., Ingraham J., Paulsen I., Saier J., M.H.: The *E. coli* EcoCyc database: no longer just a metabolic pathway database. *ASM News*, 70(1):25–30, (2004).

- Karp P. D., Ouzounis C. A., Moore-Kochlacs C., Goldovsky L., Kaipa P., Ahrén D., Tsoka S., Darzentas N., Kunin V., López-Bigas N.: Expansion of the BioCyc collection of pathway/genome databases to 160 genomes. *Nucleic Acids Research*, 33(19):6083–6089, (2005).
- Karp P. D., Riley M., Paley S. M., Pellegrini-Toole A.: The MetaCyc Database. *Nucleic Acids Research*, 30(1):59–61, (2002).
- Keseler I. M., Bonavides-Martínez C., Collado-Vides J., Gama-Castro S., Gunsalus R. P., Johnson D. A., Krummenacker M., Nolan L. M., Paley S., Paulsen I. T., Peralta-Gil M., Santos-Zavaleta A., Shearer A. G., Karp P. D.: EcoCyc: a comprehensive view of *Escherichia coli* biology. *Nucleic Acids Research*, 37(Database issue):D464–D470, (2009).
- Keseler I. M., Collado-Vides J., Gama-Castro S., Ingraham J., Paley S., Paulsen I. T., Peralta-Gil M., Karp P. D.: EcoCyc: a comprehensive database resource for *Escherichia coli*. *Nucleic Acids Research*, 33(Database issue):D334–D337, (2005).
- Khamis A., Raoult D., Scola B. L.: Comparison between *rpoB* and 16S rRNA gene sequencing for molecular identification of 168 clinical isolates of *Corynebacterium*. *Journal of Clinical Microbiology*, 43(4):1934–1936, (2005).
- Krause L., Diaz N. N., Goesmann A., Kelley S., Nattkemper T. W., Rohwer F., Edwards R. A., Stoye J.: Phylogenetic classification of short environmental DNA fragments. *Nucleic Acids Research*, 36(7):2230–2239, (2008).
- Krings E., Krumbach K., Bathe B., Kelle R., Wendisch V. F., Sahm H., Eggeling L.: Characterization of myo-inositol utilization by *Corynebacterium glutamicum*: the stimulon, identification of transporters, and influence on L-lysine formation. *Journal of Bacteriology*, 188(23):8054–8061, (2006).
- Kulikova T., Akhtar R., Aldebert P., Althorpe N., Andersson M., Baldwin A., Bates K., Bhattacharyya S., Bower L., Browne P., Castro M., Cochrane G., Duggan K., Eberhardt R., Faruque N., Hoad G., Kanz C., Lee C., Leinonen R., Lin Q., Lombard V., Lopez R., Lorenc D., McWilliam H., Mukherjee G., Nardone F., Pastor M. P. G., Plaister S., Sobhany S., Stoehr P., Vaughan R., Wu D., Zhu W., Apweiler R.: EMBL nucleotide sequence database in 2006. *Nucleic Acids Research*, 35(Database issue):D16–D20, (2007).
- Liao L., Kim S., Tom J.-F.: Genome comparisons based on profiles of metabolic pathways. In: E. Damiani, R. J. Howlett, L. C. Jain, N. Ichalkaranje, eds., *Proceedings of the 6th International Conference on Knowledge-Based Intelligent Information and Engineering Systems (KES 2002)*, Crema, Italy, vol. 82 of *Frontiers in Artificial Intelligence and Applications*, pages 469–476, IOS Press, Amsterdam (2002).
- Lipkus A. H.: A proof of the triangle inequality for the Tanimoto distance. *Journal of Mathematical Chemistry*, 26(1-3):263–265, (1999).
- Maltsev N., Glass E., Sulakhe D., Rodriguez A., Syed M. H., Bompada T., Zhang Y., D’Souza M.: PUMA2-grid-based high-throughput analysis of genomes and metabolic pathways. *Nucleic Acids Research*, 34(Database issue):D369–D372, (2006).

- Mavrovouniotis M. L., Stephanopoulos G., Stephanopoulos G.: Computer-aided synthesis of biochemical pathways. *Biotechnology and Bioengineering*, 36(11):1119–1132, (1990).
- Michal G.: *Biochemical Pathways*. Spektrum Akademischer Verlag, Heidelberg (1999).
- Nishio Y., Nakamura Y., Kawarabayasi Y., Usuda Y., Kimura E., Sugimoto S., Matsui K., Yamagishi A., Kikuchi H., Ikeo K., Gojobori T.: Comparative complete genome sequence analysis of the amino acid replacements responsible for the thermostability of *Corynebacterium efficiens*. *Genome Research*, 13(7):1572–1579, (2003).
- Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (NC-IUBMB): *Enzyme nomenclature 1992 : recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the nomenclature and classification of enzymes*. Academic Press, San Diego (1992).
- Oehm S., Gilbert D., Tauch A., Stoye J., Goesmann A.: Comparative Pathway Analyzer – a web server for comparative analysis, clustering and visualization of metabolic networks in multiple organisms. *Nucleic Acids Research*, 36(Web Server issue):W433–W437, (2008).
- Ogata H., Fujibuchi W., Goto S., Kanehisa M.: A heuristic graph comparison algorithm and its application to detect functionally related enzyme clusters. *Nucleic Acids Research*, 28(20):4021–4028, (2000).
- Overbeek R., Larsen N., Pusch G. D., D’Souza M., Selkov Jr E., Kyrpides N., Fonstein M., Maltsev N., Selkov E.: WIT: integrated system for high-throughput genome sequence analysis and metabolic reconstruction. *Nucleic Acids Research*, 28(1):123–125, (2000).
- Pereira G. A., Pimenta F. P., dos Santos F. R. W., Damasco P. V., Júnior R. H., Mattos-Guaraldi A. L.: Antimicrobial resistance among Brazilian *Corynebacterium diphtheriae* strains. *Memorias do Instituto Oswaldo Cruz*, 103(5):507–510, (2008).
- Qian Y., Lee J. H., Holmes R. K.: Identification of a DtxR-regulated operon that is essential for siderophore-dependent iron uptake in *Corynebacterium diphtheriae*. *Journal of Bacteriology*, 184(17):4846–4856, (2002).
- Radmacher E., Alderwick L. J., Besra G. S., Brown A. K., Gibson K. J. C., Sahn H., Eggeling L.: Two functional FAS-I type fatty acid synthases in *Corynebacterium glutamicum*. *Microbiology*, 151(7):2421–2427, (2005).
- Rodionov D. A., Vitreschak A. G., Mironov A. A., Gelfand M. S.: Comparative genomics of the vitamin B12 metabolism and regulation in prokaryotes. *The Journal of Biological Chemistry*, 278(42):41148–41159, (2003).
- Romesburg H. C.: *Cluster Analysis for Researchers*. Belmont, California (1984).
- Schilling C. H., Letscher D., Palsson B. Ø.: Theory for the systemic definition of metabolic pathways and their use in interpreting metabolic function from a pathway-oriented perspective. *Journal of Theoretical Biology*, 203(3):229–248, (2000).

- Schomburg I., Chang A., Ebeling C., Gremse M., Heldt C., Huhn G., Schomburg D.: BRENDA, the enzyme database: updates and major new developments. *Nucleic Acids Research*, 32(Database issue):D431–D433, (2004).
- Schuster S., Hilgetag C.: On elementary flux modes in biochemical reaction systems at steady state. *Journal of Biological Systems*, 2(2):165–182, (1994).
- Selkov E., Basmanova S., Gaasterland T., Goryanin I., Gretchkin Y., Maltsev N., Nenashev V., Overbeek R., Panyushkina E., Pronevitch L., Selkov E., Yunus I.: The metabolic pathway collection from EMP: the Enzymes and Metabolic Pathways database. *Nucleic Acids Research*, 24(1):26–28, (1996).
- Selkov E., Grechkin Y., Mikhailova N., Selkov E.: MPW: the Metabolic Pathways Database. *Nucleic Acids Research*, 26(1):43–45, (1998).
- Seressiotis A., Bailey J. E.: MPS: An artificially intelligent software system for the analysis and synthesis of metabolic pathways. *Biotechnology and Bioengineering*, 31(6):587–602, (1988).
- Sharma V., Gupta P., Dixit A.: In silico identification of putative drug targets from different metabolic pathways of *Aeromonas hydrophila*. In *Silico Biology*, 8(3-4):331–338, (2008).
- Späth H.: *Cluster Analysis Algorithms*. Ellis Horwood, Chichester (1980).
- Tatusov R. L., Koonin E. V., Lipman D. J.: A genomic perspective on protein families. *Science*, 278(5338):631–637, (1997).
- Tauch A., Kaiser O., Hain T., Goesmann A., Weisshaar B., Albersmeier A., Bekel T., Bischoff N., Brune I., Chakraborty T., Kalinowski J., Meyer F., Rupp O., Schneiker S., Viehoveer P., Pühler A.: Complete genome sequence and analysis of the multiresistant nosocomial pathogen *Corynebacterium jeikeium* K411, a lipid-requiring bacterium of the human skin flora. *Journal of Bacteriology*, 187(13):4671–4682, (2005).
- Tauch A., Trost E., Tilker A., Ludewig U., Schneiker S., Goesmann A., Arnold W., Bekel T., Brinkrolf K., Brune I., Götter S., Kalinowski J., Kamp P.-B., Lobo F. P., Viehoveer P., Weisshaar B., Soriano F., Dröge M., Pühler A.: The lifestyle of *Corynebacterium urealyticum* derived from its complete genome sequence established by pyrosequencing. *Journal of Biotechnology*, 136(1-2):11–21, (2008).
- Tohsato Y., Matsuda H., Hashimoto A.: A multiple alignment algorithm for metabolic pathway analysis using enzyme hierarchy. In: R. Altman, T. L. Bailey, P. Bourne, M. Gribskov, T. Lengauer, I. N. Shindyalov, L. F. T. Eyck, H. Weissig, eds., *Proceedings of the 8th International Conference on Intelligent Systems for Molecular Biology (ISMB 2000)*, La Jolla/San Diego, California, vol. 8, pages 376–383, AAAI Press, Menlo Park, California (2000).
- Tyson G. W., Chapman J., Hugenholtz P., Allen E. E., Ram R. J., Richardson P. M., Solovyev V. V., Rubin E. M., Rokhsar D. S., Banfield J. F.: Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature*, 428(6978):37–43, (2004).

- UniProt Consortium: The universal protein resource (UniProt). *Nucleic Acids Research*, 36(Database issue):D190–D195, (2008).
- Valiente G.: *Algorithms on Trees and Graphs*. Springer Verlag, Berlin (2002).
- Venter J. C., Remington K., Heidelberg J. F., Halpern A. L., Rusch D., Eisen J. A., Wu D., Paulsen I., Nelson K. E., Nelson W., Fouts D. E., Levy S., Knap A. H., Lomas M. W., Nealson K., White O., Peterson J., Hoffman J., Parsons R., Baden-Tillson H., Pfannkoch C., Rogers Y.-H., Smith H. O.: Environmental genome shotgun sequencing of the Sargasso Sea. *Science*, 304(5667):66–74, (2004).
- Wagner A., Fell D. A.: The small world inside large metabolic networks. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 268(1478):1803–1810, (2001).
- Ward Jr J. H.: Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301):236–244, (1963).
- Willett P., Barnard J. M., Downs G. M.: Chemical similarity searching. *Journal of Chemical Information and Computer Sciences*, 38(6):983–996, (1998).
- Ye Y., Osterman A., Overbeek R., Godzik A.: Automatic detection of subsystem/pathway variants in genome analysis. *Bioinformatics*, 21 Suppl 1:i478–i486, (2005).