
# BRNO UNIVERSITY OF TECHNOLOGY

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

## FACULTY OF ELECTRICAL ENGINEERING AND COMMUNICATION

FAKULTA ELEKTROTECHNIKY
A KOMUNIKAČNÍCH TECHNOLOGIÍ

## DEPARTMENT OF BIOMEDICAL ENGINEERING

ÚSTAV BIOMEDICÍNSKÉHO INŽENÝRSTVÍ


# METHODS FOR COMPARATIVE ANALYSIS OF METAGENOMIC DATA

METODY PRO KOMPARATIVNÍ ANALÝZU METAGENOMICKÝCH DAT

## SUMMARY OF DOCTORAL THESIS

TEZE DIZERTAČNÍ PRÁCE


**AUTHOR** **Mgr. Ing. Karel Sedlář**
AUTOR PRÁCE

**SUPERVIZOR** **prof. Ing. Ivo Provazník, Ph.D.**
ŠKOLITEL



**BRNO 2018**

## ABSTRACT

Modern research in environmental microbiology utilizes genomic data, especially sequencing of DNA, to describe microbial communities. The field studying all genetic material present in an environmental sample is referred to as metagenomics. This doctoral thesis deals with metagenomics from the perspective of bioinformatics that is unreplaceable during the data processing. In the theoretical part of this thesis, two different approaches of metagenomics are described including their main principles and weaknesses. The first approach, based on targeted sequencing, is a well-established field with a wide range of bioinformatics techniques. Yet, methods for comparison of samples from several environments can be highly improved. The approach introduced in this thesis uses unique transformation of data into bipartite graph, where one partition is formed by taxa, while the other by samples or environments. Such a graph fully reflects qualitative as well as quantitative aspect of analyzed microbial networks. It allows a massive data reduction to provide human comprehensible visualization without affecting the automatic community detection that can found clusters of similar samples and their typical microbes. The second approach utilizes whole metagenome shotgun sequencing. This strategy is newer and the corresponding bioinformatics techniques are less developed. The main challenge lies in fast clustering of sequences, in metagenomics referred to as binning. The method introduced in this thesis utilizes genomic signal processing approach. By thorough analysis of redundancy of genetic information stored in genomic signals, a unique technique was proposed. The technique utilizes transformation of character sequences into several variants of phase signals. Moreover, it is able to directly process nanopore sequencing data in the form of native current signal.



## KEYWORDS

metagenomics; targeted sequencing; shotgun sequencing; bipartite graph; microbial network; binning; genomic signal processing

# CONTENTS

# INTRODUCTION

Even though the microbial organisms are not visible to the naked eye, we meet them on every step of our way as they are present everywhere. They live in soil, water, air, on a surface of trees, fruits, or any thinkable habitats including human-made things like doorknobs etc. They even live on and within the bodies of higher eukaryotic organisms, including humans, and form large communities. All the genetic material of these individual organisms within an environment forms a metagenome. There are two main ways to study microbial communities by the means of metagenomics, differing by a sequencing approach. The first way is based on amplicon sequencing of a target gene in the metagenome. Such an approach cannot describe the whole metagenome omitting all the genes, except for the target gene, it contains. On the other hand, it provides an easy way to perform powerful qualitative as well as quantitative analyses of a microbial community by identifying particular taxa while comparing the sequenced genes with a database and by getting the abundance of a taxon while counting copies of the target gene belonging to a taxon. The second way relies on a whole metagenome shotgun sequencing, which allows it to provide a full analysis of a metagenome, including genes, and pathways it contains. However, the data processing is much more complicated.

In this thesis, I am describing the assumptions and expectations of using both of these metagenomic ways, aiming primarily on the bioinformatics approaches and tools and I am presenting improvements and novel techniques in places I found the current tools to be weak. While the preprocessing of amplicon sequencing data is highly automated, techniques for comparison of microbial communities from different habitats and data interpretation are underdeveloped. The technique proposed in the thesis utilizes bipartite graphs. As demonstrated in the thesis, the tool that is nowadays mostly used for analysis of social networks has advantageous properties also for analysis of microbial communities. Not only brings this technique the possibility to reduce the dimensionality of the data without affecting the result of automatic clustering by community detection but it also provides a powerful tool for data visualization that is unreplaceable in data interpretation. On the contrary, the field of processing whole metagenome shotgun data lacks techniques for automatic data clustering and dimensionality reduction. It mostly uses stochastic transformations to be able to process large datasets of sequences. Unlike the current tools, the solution presented in this thesis is based on genomic signal processing. By its application, a completely new, deterministic, and very fast transformation of the metagenomic datasets could have been proposed. It allows the use of a following comparative analysis by automatic clustering as well as human comprehensible visualization for data inspection.

# 1 METAGENOMICS

Metagenomics is closely related to environmental microbiology, the scientific discipline that is more than one hundred years old; yet, the term 'metagenomics' was only introduced by Handelsman et al. [1] in 1998. Its principle lies in studying all genetic material present in an environmental sample without the need for cultivation. As it has been shown, by sequencing of cultivated samples obtained from an environment, a microbial diversity tends to be considerably underestimated as the majority (>99%) of organisms cannot be cultivated by standard techniques [2].

## 1.1 Metagenomic studies

During several years, many previously undescribed microorganisms have been discovered by direct isolation of genetic material from their natural habitat and metagenomics, in addition to microbiology, has also been used in other disciplines and fields such as medicine, veterinary medicine, food industry, ecology, or biotechnology [3]. The wide usage of metagenomics can be supported by the estimation of the total number of microbial cells on Earth, which is $10^{30}$, formed mainly by $10^6$ to $10^8$ separate prokaryotic genospecies [4]. Although the myth claiming that our bodies contain 10 more times bacterial cells than human own cell was busted in 2016, bacteria form still a substantial part of our bodies when the average man is believed to carry $3.8 \cdot 10^{13}$ bacterial cells, which is roughly the same as the total number of human cells [5]. Therefore, the influence of bacteria on human health is being discussed broadly. While many studies aim at gut microbiota composition and its influence on human health [6, 7], also other metagenomes as those in bronchial tracts are being studied [8]. Even though the majority of bacteria are harmless or rather beneficial to a human body, metagenomic approach is also being used in a study of harmful microbes, e.g. those carrying antibiotic resistance genes [9]. In veterinary medicine, metagenomics helps to reveal spreading genes of antibiotic resistance [10] or to prevent economic losses by studying effects of feeding [11] or housing [12] on microbiota composition of farm animals. Applications in ecology usually consist of studying microbial communities in soil or water [13] and biotechnology uses metagenomics to reveal properties of microbial communities utilizable in industry, e.g. biodegradation [14].

A rapid development of DNA sequencing techniques substantially changes the way the metagenomic studies are carried out. Constant improvement of lab techniques puts pressure on the development of increasingly powerful bioinformatics tools. Those tools are in metagenomic research unreplaceable, as the main challenge has moved along

from the data collection to the data analysis. Advances in DNA sequencing caused an immense increase in data volume being processed during a typical metagenome study [15]. Instead of kilobytes and megabytes, a common study starts with gigabytes or even terabytes of raw data. Later, during the data evaluation, datasets are being reduced continuously, ending up in tables of organisms or genes that are human comprehensible and ready to be interpreted. However, such a massive reduction has to be done rigorously, without any negative effects on the final results of the analysis. This is the reason for the rising importance of bioinformaticians being involved in complex metagenomic studies that require the multidisciplinary cooperation of several researchers from different fields.

Two different sequencing strategies can be used [16]. If the main goal is only to taxonomically describe the microbial community, amplicon sequencing of selected gene, so called ***targeted sequencing***, is usually used. Although this approach only allows to determine species, or rather only genera in a sample without any information of their genomes, the data handling is easier and for many applications in microbial studies is such a result sufficient [17]. On the contrary, the second approach allows to study whole genomes of organisms present in a sample while the data processing is much more complicated. This approach is based on ***whole metagenome shotgun sequencing*** (WMS) [18]. The principle of these approaches is summarized in Fig. 1.1.

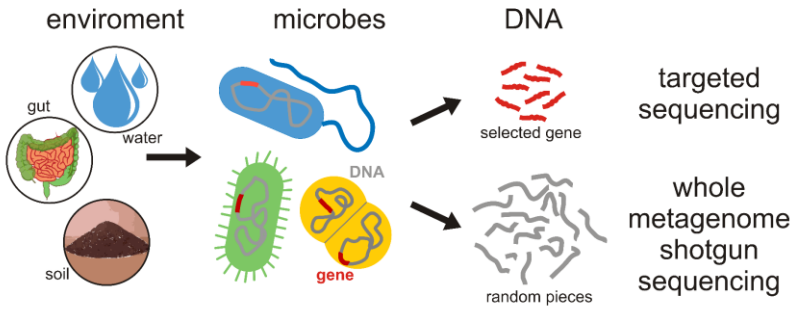


**Fig. 1.1 – Two approaches of studying the microbiota**
Culture independent methods for characterization of microbial communities are based on direct sequencing of isolated DNA to prevent loss of diversity by cultivation. Targeted sequencing captures all organisms in a sample, whole metagenome sequencing provides also information about the functional and metabolic potential of the community.

## 1.2 Targeted sequencing studies

The use of amplicon sequencing for metagenomics has its specifics and a wide range of bioinformatics techniques, tools, or whole pipelines have been already proposed. The bioinformatics strategy for microbial studies can be divided into three main blocks as shown in Fig. 1.2.

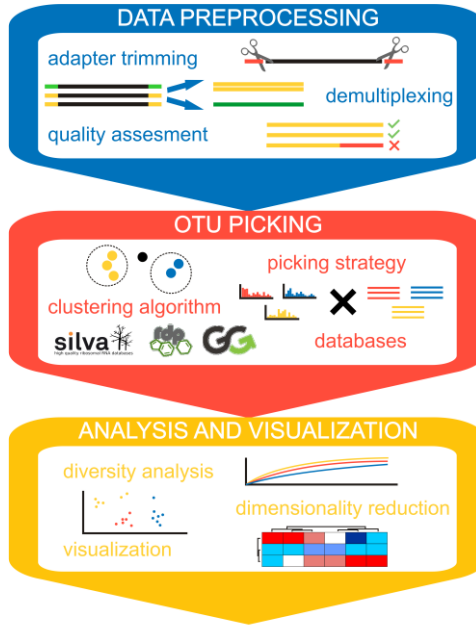


**Fig. 1.2 – Bioinformatics strategy for targeted sequencing**
The whole strategy contains three main blocks. Data preprocessing is similar to other amplicon sequencing projects but remaining blocks are highly specialized for microbial studies.

The first steps in targeted sequencing metagenomics data processing are the same as for any other NGS application and are usually automated. These include adapter trimming, demultiplexing, quality trimming, and pair-end reads merging.

The crucial step in targeted sequencing metagenomics comes after the sequences are demultiplexed and trimmed. During their clustering, referred to as OTU picking, Operational Taxonomic Units are being formed. An OTU represents a cluster of organisms grouped on the basis of their DNA similarity. Such a grouping allows recognition of known organism using a reference database, identification of

novel organisms, data quantification, and application of a wide range of statistical and analytical techniques for description of microbiomes. The purpose of forming particular OTUs is to identify all known and novel organisms in the sample. However, in practice, the achievement of this goal can be quite complicated. There are three main strategies for OTU picking – *de novo*, closed reference, and open reference clustering [19]. The results of OTU picking do not depend only on a selected strategy but are also substantially influenced by a range of other parameters. Among them, clustering algorithm and the level of sequence similarity are the most important. As a default, 97% similarity is set. This value was proposed as the threshold for species delineation based on 16S rRNA gene similarity [20].

Once the data are transformed into OTU table, a novel knowledge can be gathered using a wide range of techniques for analysis of diversity within a single environment or diversity between two environments, relations between community composition and environmental features, and phylogenetic relationships within a community. Possibly, tools primarily meant to analyze diversity can be used to verify the sufficiency of the sequencing depth. The first step in amplicon metagenomic data interpretation usually lies in an analysis of diversity. There are two basic types of diversity: α and β. **Alpha diversity** deals with species within a sample and tries to describe relationships of their co-occurrence, e.g. a number of species and their abundance. On the contrary, **beta diversity** tracks the differences in species composition among various samples, e.g. a number of shared species. The simplest way to describe an OTU table can be found in indexes of diversity, which try to capture a diversity with a single number. They can be understood as an analogy of descriptive statistics, e.g. mean, median, standard deviation etc. There is a wide range of various indexes [21]. While interpretation of some indexes is very similar and their common usage is redundant, combination of contrasting indexes can be highly informative for description of communities.

### *Visualization techniques*

Although microbial data are complex and multidimensional, their simple visualization is possible by a range of techniques. Efficient visualization method can help finding hidden bonds between microbiomes and is crucial in processing and interpretation of metagenomic datasets. There is almost an infinite variety of libraries and tools dedicated for data visualization in scripting languages like Python, R, Matlab etc. or in software for statistical computing like SPSS or Statistica that can be used for microbial studies. Basic techniques for data visualization are commonly available also within pipelines dedicated for metagenomics and include heatmap of OTU abundances [22], scatter plots [23], time series plots [24], and bar plots [25]. Other techniques utilize calculations

of selected parameters and dimensionality reduction. UniFrac-PCoA [26], PCoA biplots [24], heatmaps of correlation coefficients [13] are the most widely used among them.

## 1.3 Whole metagenome shotgun sequencing studies

The steps within WMS sequencing data processing are substantially different from those of targeted sequencing approach. Although they can be also divided into three main blocks as shown in Fig. 1.3, they may not be applied in the exact order and they rather supplement than follow each other. This is caused by different nature of the shotgun data as well as different expectations and demands for the results of an analysis.

Raw shotgun data represent random chunks of genomes. Thus, the first logical step lies in sequence assembly in order to reconstruct original genomes present in a metagenome. However, **metagenome assembly** is a complicated process and separate assembling of related read clusters formed during **binning** process may substantially improve the assembly, i.e. lower the number of contigs while prolonging them. On the other hand, binning of longer sequences usually produces better clusters and some of the binning algorithms require assembled data. Contigs can undergo **descriptive and functional analyses** at once or in separate bins and in the opposite way, reads with additional information acquired during an analysis can help to improve the assembly and the binning process.

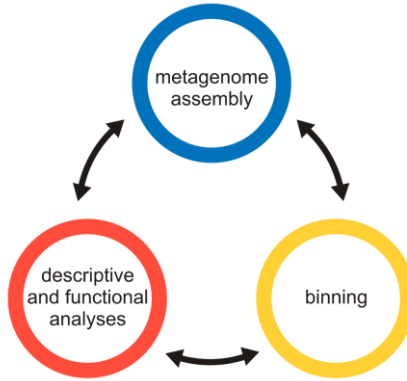


**Fig. 1.3 – Bioinformatics strategy for shotgun sequencing**
The whole strategy contains three main blocks that supplement each other. The processing of raw data may start with any of these and particular step can be applied repeatedly, e.g. contigs produced by metagenome assembly can be binned into separate clusters that are again assembled to produce lesser amount of longer contigs that are searched for genes conditioning the function of the community.

The data handling during the preprocessing step is similar to the targeted sequencing strategy and consists of tasks required for any sequencing data, mainly adapter and quality trimming. The complexity of particular genomes usually requires to be covered by the capacity of a whole sequencing run while multiplexing is not used. Nonetheless, demultiplexing of shotgun data follows the same principles as for amplicon data. Assemblies of metagenomes are extremely demanding tasks that end up far from the complete sequences covering whole genomes of every single microbe present in an environment. Yet, connecting reads into longer sequences is highly advantageous for analyses of metagenomes while it helps to improve following binning processes. Although any standard assembler can be utilized for assembling a metagenomic dataset, specialized tools with optimized parameters for metagenomic data are already available [27]. The main difference between assembly of dataset containing single organism and a metagenome is uneven sequencing depth of particular organisms within a metagenome caused by their different abundance [28]. Metagenomic assemblers are usually developed by a modification of standard assemblers, e.g. MetaVelvet [29], Meta-IDBA [30], and many others [31].

Another task of WMS data processing lies in data sorting, i.e. binning. This step is equivalent to OTU picking in amplicon data processing and serves to characterize the taxonomic diversity of microbial communities. Yet, it is computationally much more challenging. Some of the binning strategies can work directly with short reads while others are optimized for binning longer sequence, i.e. contigs or reads produced by the third generation sequencing (TGS). A certain group of algorithms, clustering the data according to their abundances, requires only contigs as these algorithms calculate with coverage of contigs by short reads. Metagenomic assembly and binning is therefore tightly connected. The basic division of binning strategies distinguishes between methods using a reference database, so called **taxonomy dependent**, and *de novo* methods, so called **taxonomy independent** [32].

Due to the incompleteness of reference databases, a majority of the novel binning algorithms utilizes reference-free approach. My colleagues and I have proposed the division of taxonomy independent techniques into three categories – (i) sequence composition based methods, (ii) abundance based methods, and (iii) hybrid methods [33], see Fig. 1.4. All of these methods use feature vectors inferred from the original data rather than using whole sequences to achieve better computational efficiency. Their main difference comes from the nature of feature extraction. While (i) work with compositional properties of sequences, (ii) work with features based on contig coverage reflecting abundance of given taxa in a microbial sample. Hybrid methods combine both of these features. At the same time, data visualization is

increasingly important to provide user with informative overview of the data and to allow him to perform appropriate corrections, e.g. human augmented binning [34]. Therefore, such binning tools use advanced dimensionality reduction techniques.

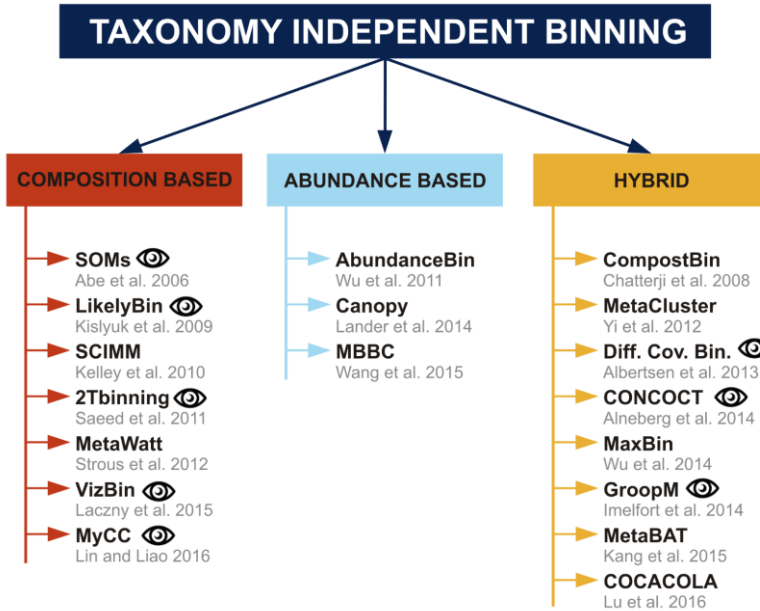


**Fig. 1.4 – Division of taxonomy independent techniques by Sedlar et al.** [33] Current reference-free binning tools can be divided into three subcategories according to the nature of a feature vector used for data clustering. Several techniques (highlighted by an eye symbol) employ dimensionality reduction techniques to provide informative visualization of datasets.

At last but not least, WMS data are subjected to descriptive and functional analyses to infer biological knowledge about the analyzed community. Unlike the targeted sequencing approach, whole metagenome data contain all functional properties of an analyzed microbial community. Nevertheless, genomic approach limits the amount of observable properties to the metabolic and functional potential of a microbial community rather than its real state. To fully describe the function, other omics need to be involved. Functional studies in metagenomics copy procedures applied in a standard genomics. These consist of annotation of sequences and following assignment of a function to annotated elements. Among annotation tools, MG-RAST [35], METAREP [36] or IMG/M [37] are available.

# 2 COMPUTATIONAL BACKGROUND

Deeper understanding of some of computational state-of-the-art methods and tools is important for development of novel, more advanced tools. Those include especially concept of mathematical graphs and techniques for sequences feature extraction and processing. While graphs are suitable for analysis and reduction of large OTU tables and their informative visualizations, feature vectors allow application of very fast algorithms on extensive datasets gathered in whole metagenome shotgun sequencing studies.

## 2.1 Graph theory

The concept of graphs is significantly older than metagenomics itself. In fact, the famous problem of bridges in Königsberg defined in a paper by Leonhard Euler is considered to be the very first paper in the history of graph theory [38]. Euler published the paper in 1736, 262 years before Handelsman introduced the term 'metagenomics'. Since then, graph theory has gone through great development and found application in a wide range of disciplines and fields, from physics through computer science to biology or even humanities.

A graph is a basic object of graph theory or discrete mathematics that can capture pairwise relations between objects. In the most common sense of the term, a graph G is an ordered pair [39]:

$$G = (V, E), \tag{2.1}$$

where $V$ is a set of **vertices** (also called nodes or points) and $E$ is a set of **edges** (also called lines or arcs) connecting these vertices. Thus, $E$ can be understood as:

$$E \subseteq \{\{u, v\} | u, v \in V, u \neq v\}, \tag{2.2}$$

a two-element subsets of $V$. Such a graph is referred to as **undirected** because the association of $u$ and $v$ takes the form of the unordered pair, i.e. $e = \{u, v\}$. However, the pair of $u$ and $v$ may be ordered, i.e. $e = (u, v)$. In that case, the graph is **directed**.

Visual representation of a graph, graph drawing, is its pictorial representation but in fact, a graph is abstract data type that can be defined by several data structures, mostly matrices or lists. The most common structure is an **adjacency matrix**. For simple graph, adjacency matrix is two dimensional, square, Boolean matrix $A$ of type $n \times n$, where $n = |V|$ is a number of vertices.

A special class of graph is a **bipartite graph** [40], which is a graph whose vertices can be divided into two disjoint sets such as:

$$V = V_1 \cup V_2, V_1 \cap V_2 = \emptyset \text{ and } \forall e = \{u, v\}, e \in E: u \in V_1 \wedge v \in V_2 \qquad (2.3)$$

that only two vertices of different kind $V_1$ and $V_2$ can be connected by an edge $e$. $V_1$ and $V_2$ are called partitions. The whole bipartite graph is defined by an biadjacency matrix $B_{m,n}$, whose size is determined by the sizes of both partitions $m$=|$V_1$| and $n$=|$V_2$|.

#### *Biological networks*

A biological network is a network describing any biological system. While vertices, in network theory usually called nodes, represent units of the network, edges stand for interactions between adjacent nodes. Except for the qualitative attributes, e.g. name of a unit or description of an interaction, quantitative attributes are quite common. These may represent a size of a molecule or an abundance of a species in case of nodes or bond strengths, rates of reactions etc. in the case of edges. Thus, biological networks can be regarded as weighted graphs. Applications of networks to biology are almost unlimited and include for example modelling of molecular activity in a cell [41], protein-protein interactions [42], neural networks of brain activity [43], co-occurrence patterns in microbial ecology [44] and many others [45–48].

An analysis of a biological network using graph/network algorithms is a powerful tool to infer hidden biological properties. There are two basic types of analyses. **Dynamic analyses** evaluate changes in attributes of nodes and edges over the time. These mainly include analyses of gene regulatory networks, signaling pathways, or metabolic networks by the means of computational systems biology [49]. On the other hand, a **static analysis** evaluates a topology of a network and, besides systems biology, is widely applied also in bioinformatics. Various analyses of topology of large biological networks revealed that they share many features with other networks, e.g. social networks. One of the widely used parameters evaluating topology of a network is **modularity** $Q$ [50]. It measures the strength of division of a network into particular modules representing different communities or clusters of nodes.

## 2.2 Sequence features

Bioinformatics is a discipline exploring biological sequences that are represented as text strings. From this point of view, bioinformatics may seem as a sub-discipline of stringology, which studies algorithms and data structures used for text strings processing [51]. On the other hand, biological text strings are characterized by strict rules and their information content can be quantified using various statistical,

computational, or signal processing approaches. Utilization of these techniques allows more elaborated, more complex algorithms to be applied for solving selected tasks, e.g. dimensionality reduction, fast binning, etc., that would be hardly achievable by using stringology itself.

The utilization of basic features or vectors of various features helps to overcome the most computationally demanding task – sequence alignment and following similarity calculation. The most basic parameter is GC content. Although percentage of cytosine and guanine in a sequence may seems as an imperfect parameter, a difference in GC content between unrelated populations was proved [52]. Nevertheless, a majority of feature vectors is high dimensional representing *k*-mer frequencies.

Another approach overcoming the string nature of biological sequences is utilization of genomic signal processing [53, 54] techniques (GSP). GSP is a sub-discipline of bioinformatics that applies digital signal processing techniques to biological sequences, DNA and protein. Genomic signals can be understood as advanced features or feature vectors that are obtained by a transformation of the original character string sequence into a form of digital signal or numerical vector. Genomic signals have ability to highlight various features that may occur on a larger scale than particular nucleotides. Thus, a limitation of character-based representation showing only point differences between sequences is suppressed. Various genomic signals were already proved to solve different bioinformatics task, e.g. organism comparison, sequence alignment, gene prediction etc. [55, 56]. The rules according to which a sequence is transformed are referred to as a **numeric map**. There is a range of numeric maps highlighting different features. While some maps are degenerative, i.e. the signal cannot be transformed back into a sequence, others allow reverse transformations to be applied. Another division of numeric maps is derived from the dimensionality of a transformed signal. While native dimensionality of DNA is equal to four (determined by four different nucleotides), various signal representation can reduce it to 3D, 2D, or even 1D.

### *Nanopore signals*

DNA sequences stored in FASTA format using IUPAC codes do not represent a native format of sequencing machines but are rather a product of **base-calling**, which is a process of measured data decoding. Signals representing a nanopore sequencing native data format are commonly referred to as **squiggles** [57]. Currently, novel bioinformatics tools processing directly squiggles are being proposed to overcome the need of stringology in bioinformatics.

# 3 OBJECTIVES OF THE THESIS

The purpose of the thesis is to propose novel methods for comparative analyses of metagenomic data. Due to the differences between targeted and whole metagenome shotgun sequencing data, there are two main objectives with several sub-objectives each:

1. **Develop a method analyzing targeted sequencing data for a comparative analysis of microbiomes using graph and network algorithms, including:**

    1.1. a transformation of OTU table into a graph in a way that qualitative and quantitative information is preserved, while merging of OTU or samples is allowed,

    1.2. application of network algorithms for detection of communities associated with different samples,

    1.3. a human comprehensible and informative data visualization using the most suitable layout and coloring.

2. **Develop a method processing whole metagenome shotgun sequencing data for fast taxonomy independent binning based on genomic signal processing, including:**

    2.1. a selection of the robust numeric map and transformation technique allowing efficient binning,

    2.2. a possibility of processing nanopore data,

    2.3. a human comprehensible and informative data visualization,

    2.4. application of automatic clustering techniques.

The results gathered during fulfilling the first and the second objective are presented in chapter four and five, respectively.

# 4 MICROBIOME BIPARTITE NETWORKS

Two basic approaches to metagenomic studies were described in the theoretical part of the thesis. While only whole metagenome shotgun sequencing approach results in a real study of a metagenome, targeted sequencing studies are still applied to reveal microbial composition of various habitats. Therefore, amplicon sequencing plays an irreplaceable role in microbial studies. Even though a range of advanced techniques for analyses of targeted sequencing data has been already introduced, there is still room for further improvement, especially among techniques allowing an analysis of samples from different habitats, various samples, or different time-points in a single environment.

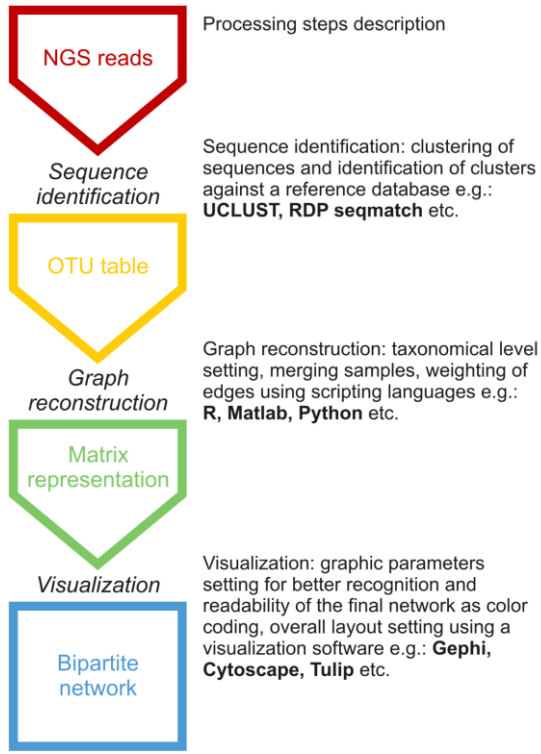


**Fig. 4.1 – Principles of microbial bipartite network reconstruction**
Flowchart describing proposed workflow. Every main step is implementable in several scripting languages/software according to the preferences of users. My colleagues and I introduced the workflow in study by Sedlar et al. [58].

A network representation of microbial communities is not a completely novel idea, for example, the pipeline QIIME allows user to generate so-called OTU networks or bipartite networks that are specifically designed to be visualized using Cytoscape [59]. Nevertheless, such networks are only briefly described and designed mainly for visualization of data rather than its full-fledged analysis. The methodology proposed in this thesis brings detailed description of particular steps needed for transformation of OTU table into a graph with suitable properties. These include, but are not limited to, various transformations allowing community detection, edge weighting, eliminations of low abundant OTUs, sample merging, and others. Rather than introducing another tool with limited usage, in this thesis, I present a comprehensive set of rules for OTU table handling in order to reconstruct informative microbial bipartite networks and their drawings using any programming language or visualization platform. I have introduced these principles in study 'Bipartite Graphs for Visualization Analysis of Microbiome Data' [58] for the first time. The whole workflow is presented in Fig. 4.1.

## 4.1 Microbiome bipartite graph

In fact, OTU table itself can be directly considered as a graph that represents a network showing alpha and beta diversity in and between several microbial samples. Nevertheless, correct identification of particular sets of entities and relations among them helps greatly for further graph refinements that are irreplaceable for meaningful analyses and human comprehensible visualizations.

An OTU table, which is $P$ $m \times n$ matrix, is obtained as a result of sequence identification during an OTU picking procedure. Each of $m$ rows represents different OTUs and each of $n$ columns represents different samples. Taxa as well as samples form together a set of entities that represent vertices of the graph describing a microbiome. Thus, the size of the set $V$ from equation (2.1) is $|V| = m + n$.

All non-zeros values in the OTU table denotes relations between two entities. Thus, they represent edges connecting two vertices. An edge can be established only between the vertex representing a taxon and the vertex representing a sample. Vertices between two taxa or two samples are not possible. Therefore, vertices can be divided into two disjoint sets in the way that only two vertices of different kind $V_1$ and $V_2$ can be connected by an edge $e$. This corresponds to the definition of a bipartite graph.

Biadjacency matrix $B$, which represents a bipartite graph, can be inferred by a simple binarization of an OTU table $P$ in a way that connection between $i^{th}$ OTU and $j^{th}$ environment is created when $i^{th}$ OTU occurs in $j^{th}$ environment. Such a matrix fully represents a graph that stands behind a microbial bipartite network. Before data

transformation from an OTU table to a graph representation is done, additional steps should be performed to ensure the following meaningful analysis. These steps should not change basic features of a graph; however, they should highlight hidden patterns that are not apparent from raw data themselves.

Firstly, a desired taxonomical level has to be chosen by summation of OTU table rows belonging to the same taxonomic unit, e.g. a sum of all bacterial genera belonging to the same class for a graph capturing occurrence of bacterial classes in different samples of whole microbial network. The choice of a higher taxonomic level reduces a number of nodes making the network visualization much clearer. While the number of unidentified sequences lowers with higher taxonomic level, a part of OTUs remains always with unassigned taxonomy. This plays no role for large networks meant primarily for detection of extensive communities of microbes, however, it can be inconvenient for visualization of sparse graphs containing labels that are meant for identification of particular taxa by its name. Thus, the omission of unidentified OTUs may be desirable for selected applications.

Secondly, another reduction of vertices can be done by merging samples. This is highly desirable for comparing different environments by visual inspection, e.g. gut microbiota of group without a treatment, group after a treatment, and a control group. It is very common that every group consists of several biological replicates.While processing of all replicates may serve as a control of reproducibility or as identification of outliers, it is convenient to reduce the amount of data for final presentation. The utilization of median value from all columns of an OTU table belonging to the particular group offers a simple way for merging samples.

Although bipartite graphs represent sparse biological networks from its definition, the utilization of OTUs standing for genera or higher taxonomic levels may increase average degree of vertices against the whole OTUs network rapidly. Many of these edges are represented by low abundant OTUs that are in a given sample presented in only units of sequences. These weak connections can be considered as a background noise that can be filtered out during binarization by application of the threshold *t*:

$$B_{i,j} = \begin{cases} 0, & P_{i,j} \leq t \\ 1, & P_{i,j} > t \end{cases} \tag{4.1}$$

for $1 \leq i \leq m$ and $1 \leq j \leq n$.

Different samples, even samples representing biological replicates from the same environment have different sequencing depths. Therefore, the threshold should be applied on a normalized OTU table that reduces uneven sequencing depth.

Although the Boolean biadjacency matrix $B$ is suitable to filter the weakest connections, it is not able to distinguish the strength of the remaining relations. Rating edges according to the relative abundance of an OTU helps to reveal the hidden patterns. This approach provides suitable results for showing the common and unique parts of microbiota, because all the taxa are given equal priority that is independent of the abundance of the taxon. A weighted biadjacency matrix $W$ can be computed as:

$$W_{i,j} = 10 \frac{P\_rel_{i,j}}{\max(P\_rel_i)}, \tag{4.2}$$

where the weight of the edge ranges from 0 to 10. This value serves mainly for final layout computation, in which the value corresponds to a thickness of an edge.

Detection of main communities forms a powerful tool for analysis of various real-world networks, or graphs standing behind these networks, respectively. The aim of detection is to find clusters of vertices that are densely interconnected by edges within the cluster, while sparsely connected to other clusters. The proposed bipartite graphs can be analyzed by any community detection algorithm including 'Fast unfolding of communities in large networks' by Blondell et al. [60], edge betweenness [61], label propagation [62], walktrap [63], leading eigenvector [64], or optimal distribution [65] algorithms. Possibly, separate detection in both partition can be done by utilization of bipartite graph projection [66].

## 4.2 Results and discussion

### *Basic transformation and network drawing*

The following results are presented using an unpublished dataset containing my own gut microbiome. This dataset was only created for my personal use to find out a possible infection by *Helicobacter pylori*. Nevertheless, inspired by Nobel's prize laureate Barry Marshall, who proved the link between *H. pylori* and stomach ulcers by drinking a broth containing the bacterium [67], I cannot miss the opportunity to present my method using my own body, which fortunately does not contain the bacterium.

The library of V3/V4 variable regions of 16S rRNA gene was not sequenced with a great depth and contained only 1540 non-chimeric sequences of sufficient quality. Moreover, 73 sequences formed singleton OTUs, i.e. an OTU represented by a single sequence, when UCLUST at 97% sequence similarity was used during OTU picking. In total, 186 OTUs were detected with open reference clustering. The resulting microbial network is shown in Fig. 4.2 A. The network was drawn using a random layout, which is very uninformative because of enormous number of overlapping nodes. Moreover,

links represented by straight lines crossed other nodes and are unweighted because the partition containing samples was formed by a single node. Even though any additional analyses are not possible for networks consisting of a single sample, the drawing can be highly improved by removing singletons, by reducing a number of crossing edges using force-directed layout, and by distinguishing between nodes from different partitions, see Fig. 4.2 B. Removing singletons is equal to application of the threshold *t* introduced in the equation 4.1.

On the other hand, remaining 113 non-singleton OTUs formed a network that was still too large for labeling every node. Moreover, more than a half of remaining OTUs were represented only up to five sequences. Thus, it is advantageous to look at the microbiome from higher taxonomic level. All detected OTUs can be grouped into 19 bacterial families from which only two families remained unidentified. Such a reduction of nodes allowed labeling of nodes as presented in Fig. 4.2 C.

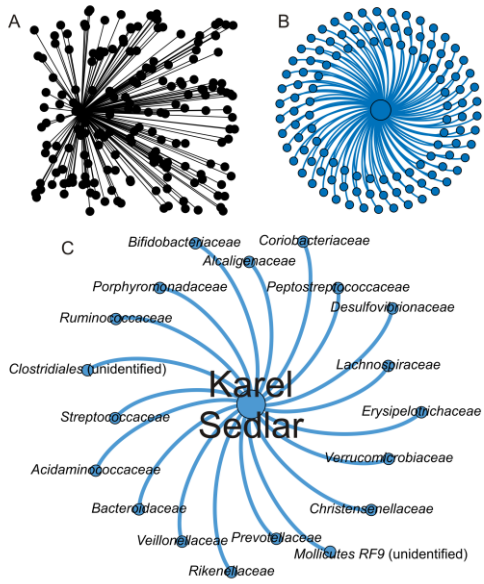


**Fig. 4.2 – Microbiome of Karel Sedlar**
Visualized bipartite network of a single sample representing gut microbiome of the author of the thesis. As all OTUs are given the same priority according to the proposed transformation, no edge weighting can be applied. (A) The network of all detected OTUs is drawn using random layout. (B) The network of non-singleton OTUs is drawn using force-directed layout. (C) The network of bacterial families with labels. Although two families remained unidentified, their identification was possible at least on the higher taxonomical level (order).

The main purpose of a microbial network is to capture and identify communities of microbes in an environment that is explored by more than a single sample. This allows full utilization of the proposed transformation, including edge weighting. A model situation can be introduced by adding two additional random samples. While sample *random 1* has a comparable number of sequences, precisely 1944, sample *random 2* has higher coverage of 5982 sequences. Although all samples have a comparable number of observed species, coefficients of α diversity suggest *random 2* have lower diversity with a dominant taxon. Remaining samples are characterized by higher equitability, see Tab. 4.1.

**Tab. 4.1 Alpha diversity of three samples**

| sample | observed species | Chao1 | Shannon | Simpson |
|---|---|---|---|---|
| Karel Sedlar | 186 | 255 | 5.748 | 0.956 |
| random 1 | 178 | 270 | 5.533 | 0.951 |
| random 2 | 180 | 182 | 3.309 | 0.712 |

The microbial network of these three samples and bacterial families they contain is presented in Fig. 4.3. Edge weighting was applied.

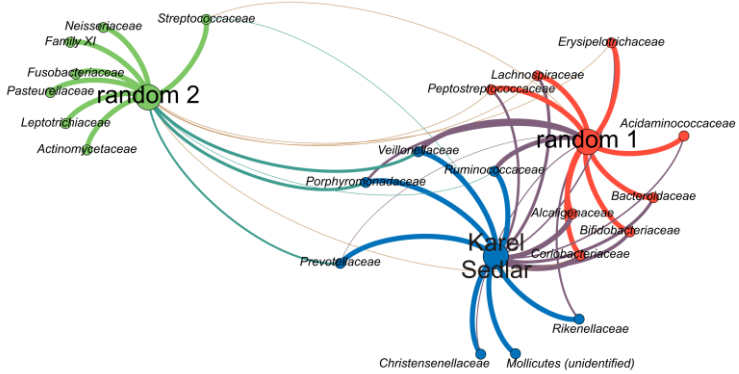


**Fig. 4.3 – Microbial network of three samples**
Visualized bipartite network of three samples. Color coding determines detected communities by fast unfolding algorithm. These communities (red, blue, and green) are formed by samples and their bacteria. Edges colored differently connect nodes from different communities.

The use of filtering by 0.5 % threshold reduced the number of visible taxa from 35 to 22. This is an advantageous step as all taxa have the same priority. Low abundant taxa are usually connected to a single sample and their weights are therefore at maximum value. Despite their possible important biological influence, the network drawing is not able to detect such fact and more nodes make network only less informative.

The reduction of visible nodes affects also alpha diversity. However, reconstructed bipartite network has only descriptive, not predictive, function. Although Chao1 coefficient suggests *random 2* have lower amount of taxa, the number of observed species is comparable for all three samples. In a similar way, the number of observed bacterial families in the presented network, formed by neighborhoods of nodes representing samples, is comparable for all samples. While neighborhoods of samples *Karel Sedlar* and *random 1* are formed by 15 taxa, the neighborhood of samples *random 2* contains 16 taxa.

In the partition of bacterial families, seven nodes have degree equal to one. These families are unique to a single sample. Other six nodes have degree equal to two. Remaining nine nodes are shared by all samples, i.e. their degree is equal to three. The network clearly shows that equitability of samples *Karel Sedlar* and *random 1* is high because edges connecting shared nodes representing bacterial families are thick, i.e. have high weights. On the contrary, weights of edges connecting *random 2* and shared bacterial families are low with the exception of node representing family *Streptococcaceae*. This suggests *Streptococcaceae* form highly abundant taxon that is responsible for lower equitability of *random 2* confirmed by lower Simpson index. This is correct because relative abundance of *Streptococcaceae* in *random 2* is more than 57 %, while its abundance in remaining samples is around 1 %.

### *Merging samples*

To demonstrate the possibilities of merging samples, a real dataset covering different environments by several samples each is needed. A suitable example can be found in already mentioned dataset from study 'Characterization of Egg Laying Hen and Broiler Fecal Microbiota in Poultry Farms in Croatia, Czech Republic, Hungary and Slovenia' [26]. The dataset consists of V3/V4 variable regions of 16S rRNA genes from 45 samples. By merging samples from egg laying hens and from broilers, direct comparison of these two environment can be done. By adding labels to nodes, one can easily assume that microbiomes of hens were more diverse than those of broilers. This could have been summarized in the network comparing these two environments, see Fig. 4.4. Me and my colleagues presented such network for the first time in the conference paper by Sedlar et al. [68]. While the degree of hen vertex was 43, the degree of broiler vertex was only 18, which means that hen microbiome consisted of 43 different bacterial families, while broiler microbiome consisted of only 18 bacterial families.

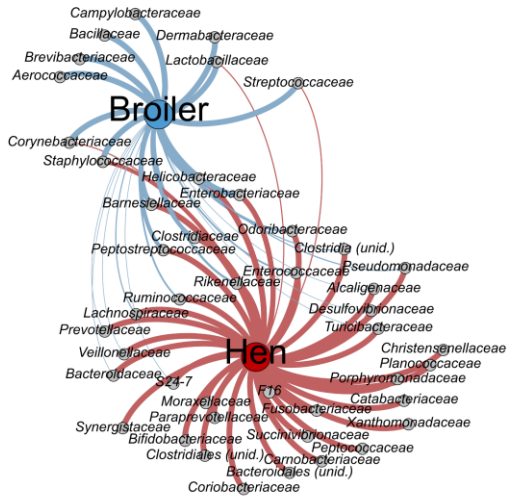


**Fig. 4.4 – Bipartite network showing bacterial families in broilers and hens**
Bipartite network showing bacterial families in broilers and hens. Neighborhoods of nodes representing broilers and hens are drawn in blue and red, respectively. Nodes representing bacterial families are grey. My colleagues and I presented this network in the conference paper by Sedlar et al. [68].

### *Communities in microbial networks*

Merging samples from the same environment utilized previously known information. Nevertheless, in some cases, samples can be obtained from the same environment that can be still divided into several communities. This can be applied for examples on samples obtained from time-series experiments.

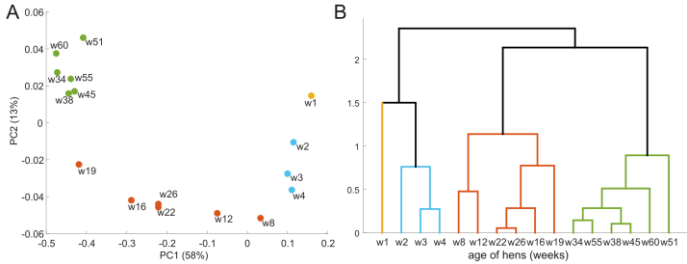


**Fig. 4.5 – Clustering of time-series samples**
(A) 2D PCoA plot of 16 samples using weighted UniFrac distance. Coordinates explains 58% and 13% of original data variability. (B) UPGMA clustering with Mahalanobis distance performed on PCoA data. Four developmental stages of microbiome (yellow, blue, red, and green) were detected.

The situation can be demonstrated on dataset containing microbiota composition in egg-laying hens during 60 weeks of their life, gathered during the study 'Succession and replacement of bacterial populations in the caecum of egg laying hens over their whole life.' [24]. The dataset of 16 samples from the first experiment was clustered using UniFrac-PCoA and hierarchical clustering, see Fig. 4.5. The second experiment added 36 samples. The complete microbial network of all 52 samples is shown in Fig. 4.6.

**Fig. 4.6 – Complete bipartite network of 52 samples**
Complete bipartite network of all samples from the study by Videnska et al. [24] and OTUs they contain introduced in study by Sedlar et al. [58]. The total number of 18,505 nodes, interconnected by 37,356 edges, is divided into four communities (yellow, blue, red, and green) by algorithm of fast unfolding of communities. Color coding respects four developmental stages.

The community colored in yellow consisted mainly of samples from hatched chickens, i.e. from the second experiment. Blue, red, and green community corresponded to the second, the third, and the fourth stage. Unfortunately manual inspection of particular parts of the network is problematic for such large networks and the purpose of the presented workflow is to visualize the data to the naked eye without the need for additional steps. A simple normalization by the use of relative counts is sufficient and suitable way for OTU table preprocessing, because of following edge weighting and data reduction by sample merging. The fact that the simple normalization and reduction do not affect the results of the analysis while improving the overall layout

can be supported by exploring the network on several taxonomic levels. The results presented in study by Sedlar et al. [58], are summarized in Tab. 4.2.

**Tab. 4.2 Parameters of reconstructed networks on several taxonomic levels**

| | whole OTU | genus | family | order | class | phylum |
|---|---|---|---|---|---|---|
| no. of vertices | 18,503 | 280 | 139 | 98 | 73 | 66 |
| no. of edges | 37,356 | 5,776 | 2,268 | 1,155 | 656 | 407 |
| average degree | 4.037 | 41.257 | 32.633 | 23.571 | 17.973 | 12.333 |
| graph density | <0.001 | 0.148 | 0.236 | 0.243 | 0.250 | 0.190 |
| average path length | 3.713 | 2.118 | 2.042 | 2.053 | 1.916 | 1.960 |
| modularity | 0.577 | 0.281 | 0.287 | 0.303 | 0.288 | 0.263 |
| no. of communities | 4 | 4 | 4 | 4 | 4 | 4 |

While moving to higher taxonomic levels reduced only the number of nodes from the partition representing taxa, the division of nodes from the partition representing samples remained satisfactorily consistent. In the yellow cluster, representing stage one, 89.5% of nodes representing samples remained the same during reduction using higher taxonomic levels. The consistency of blue cluster, representing the second stage, was 88.9%. The remaining two clusters, red and green, representing the third and the fourth stage, were even more conserved when 95.4% and 100% of nodes remained in the original cluster.

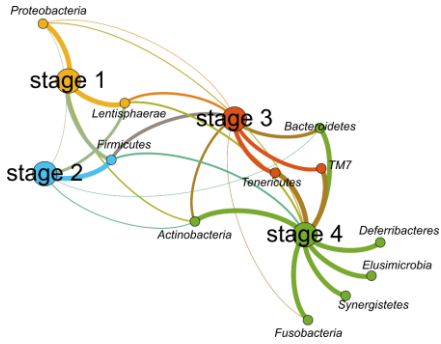


**Fig. 4.7 – Network representing four stages of microbiota development**
Four detected communities detected around nodes representing samples from four stages. The color coding of particular communities (yellow, blue, red, and green) corresponds to the color coding used in Fig. 4.5. Other colors represent intercommunity links.

On the highest taxonomic level, the network consisted of 14 nodes standing for bacterial phyla and 52 nodes standing for samples. Relations between them were described by 407 edges, which was still relatively high number. Nevertheless, another

reduction of nodes and edges can be done by sample merging, this time by newly acquired knowledge about developmental stages, see Fig. 4.7.

Although the reduction by combining samples from the same communities lowers the size of the network, for lower taxonomic levels, the application of threshold $t$ introduced in the equation 4.1 may be desirable. This can be demonstrated using the network of four detected stages and their bacterial families. While the network contains 60 different bacterial families, a majority of them has very low relative abundance that did not exceed 0.5%, see Tab. 4.3 summarizing results from study by Sedlar et al. [58].

**Tab. 4.3 Parameters of reconstructed networks using different threshold**

| threshold | 0 | 0.005 | 0.01 | 0.05 | 0.1 |
|---|---|---|---|---|---|
| no. of vertices | 64 | 21 | 16 | 12 | 10 |
| no. of edges | 156 | 33 | 25 | 16 | 10 |
| average degree | 4.875 | 3.143 | 3.125 | 2.667 | 2 |
| graph density | 0.077 | 0.157 | 0.208 | 0.242 | 0.222 |
| average path length | 2.105 | 2.181 | 2.133 | 2.242 | 2.711 |
| modularity | 0.224 | 0.338 | 0.340 | 0.377 | 0.411 |
| no. of communities | 4 | 4 | 4 | 4 | 4 |

## 4.3 Summary

The methodology and results presented in the chapter four corresponded to the first objective of the thesis – to develop a method for an analysis of targeted sequencing data using graph and network algorithms.

Although graph representation of microbial networks and communities was already known, it was considered to be only a simple visualization technique without any possibility of an additional analysis. The methodology introduced in this thesis brings strict rules for transformation of an OTU table into a bipartite graph representing a microbial network. This unique procedure is able to preserve qualitative and quantitative information and allows additional filtering steps to be applied. The same priority given to every taxon brings the possibility of comparing samples from various environments and studies. Division of networks and identification of different microbial communities typical for an environment or a group of samples can be done by several algorithms utilizing modularity of networks. Extensive possibilities of color coding can highlight various information hidden in data. It is the combination of both partitions, samples and taxa, within the common layout and their common clustering, which makes the presented method unique. Direct visualization of an OTU table in a heatmap, nor PCoA biplot can provide such informative and human comprehensible results.

# 5 METAGENOMIC SIGNAL BINNING

The following chapter is dedicated to processing of whole metagenome shotgun data. The main problem of processing short parts of various genomes lies in their division into clusters, i.e. binning, which represents the base for additional analyses resulting in biological knowledge inference. Different alignment independent techniques use diverse transformations of sequences into feature vectors. Unfortunately, such a transformation of character sequence into several numeric parameters is cumbersome. I see a great potential in utilization of signal processing techniques for fast and efficient description of biological sequences. Only a simple additional step of a sequence transformation into a signal is needed [69].

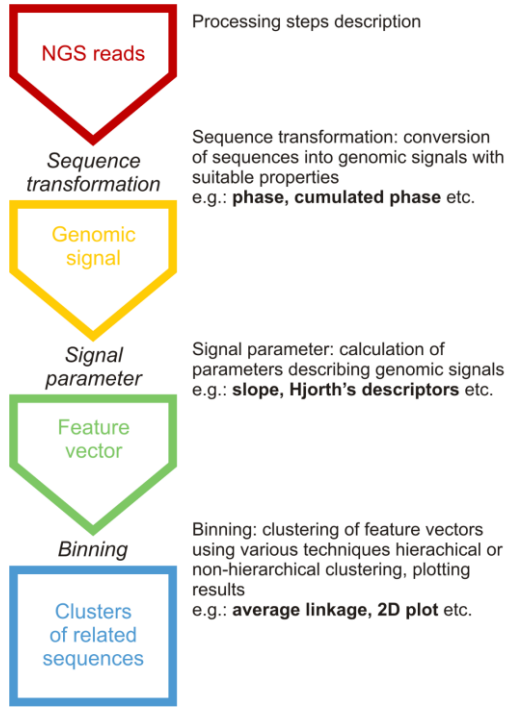


**Fig. 5.1 – Principles of metagenomic signals binning**
Flowchart describing proposed workflow. Every main step allows utilization of several different techniques for data transformation, clustering, and visualization. It is implementable in any scripting language according to the preference of users.

Binning methods themselves are not individual clustering algorithms but rather a unique combination of data transformation, clustering, dimensionality reduction, and visualization techniques. The method I introduce in this thesis meets this definition and combines selected numeric maps and signal parameters with standard clustering techniques in a novel and unique way. The whole workflow is presented in Fig. 5.1. Possibly, the sequence transformation may be omitted when a native current signal from nanopore sequencing is used.

## 5.1 Signal features and clustering

A numeric map assigns real numbers to nucleotides {A, C, G, T} in a sequence. The utilized signal representation is called a phase signal, because this mapping is based on a phase of complex numbers derived from the projection of nucleotide tetrahedron. Tetrahedron is a 3D object; therefore, there are three possible projections in directions of axes *x*, *y*, and *z*. The projection {π/4, -3π/4, 3π/4, -π/4} represents such a rotation of the tetrahedron that takes into account purine-pyrimidines (R-Y) and strong-week (S-W) chemical properties of nucleotides. Remaining projections {3π/4, -3π/4, π/4, -π/4} and {3π/4, -3π/4, -π/4, π/4} represent R-Y and amino-keto (M-K) properties and S-W and M-K properties, respectively. Thus, three different phase signals can be derived from the original character sequence. The conversion of a character sequence can be done in linear time and the reverse transformation is also possible. Moreover, phase can be cumulated or unwrapped along the analyzed sequence.

Unlike the original sequence, the signal representation allows application of spectral analysis by Fourier transform and other signal processing and description techniques, including wavelet transform or Hjorth descriptors. Hjorth descriptors (activity, mobility, and complexity) are based on spectral moments but can also be calculated from time (nucleotide) variances of a given signal, which lowers computational time:

$$\begin{aligned} A = Activity &= \sigma_0^2, \\ M = Mobility &= \frac{\sigma_1}{\sigma_2}, \\ C = Complexity &= \frac{\sigma_2/\sigma_1}{\sigma_1/\sigma_0}, \end{aligned} \tag{5.1}$$

where $\sigma_0^2$ stands for the variance of the genomic signal and $\sigma_1$ with $\sigma_2$ are the standard deviations of the first and the second derivatives of the signal, respectively. Numerical differentiation is used as approximation of the signal derivatives for digital signals.

## 5.2 Results and discussion

### *Redundancy in genomic signals*

DNA sequences carry a huge amount of information. However, a majority of this information is redundant for classification of organisms. Targeted sequencing approach utilizes this fact and aims only on 16S rRNA genes that carry enough of interspecies information. In whole shotgun metagenomics, the information is reduced by utilization of a selected parameter, e.g. tetramer frequency, which omits redundant information. Thus, even the information stored in genomic signals should be redundant.

In study "Set of rules for genomic signal downsampling" [70] me and my colleagues examined the possibility of genomic signal reduction by dyadic wavelet transform (DWT). The test dataset contained 420 sequences and proved that sequences can be massively downsampled without affecting results of a following cluster analysis.

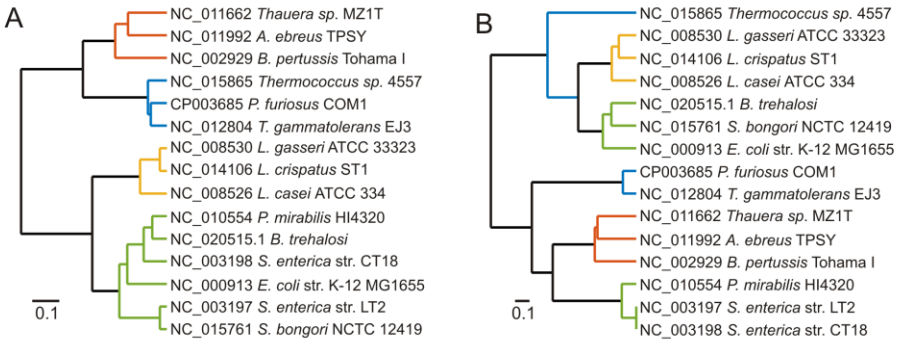


**Fig. 5.2 – Comparison of phylogenies using signals and sequences**
(A) Phylogenetic tree reconstructed from downsampled whole genome signals using Euclidean metrics and average linkage. (B) Phylogenetic tree reconstructed from 16S rRNA genes using proportional sequence distance and average linkage. Color-coding distinguishes between different bacterial classes: *Bacili* (yellow), *Betaproteobacteria* (red), *Gammaproteobacteria* (green), and *Thermococci* (blue). Me and my colleagues presented this analysis in study by Sedlar et al. [71]

Although the average length of a five Mbp genome signal after downsampling by DWT at 14$^{th}$ level is around 300 samples (16384× shorter than the original signal), hierarchical clustering distinguishes among particular taxonomic classes of bacteria. On the contrary, the standard sequence clustering method using 16S rRNA genes, whose length is around 1600 bp, leads to phylogenetic tree with clusters containing more than one bacterial class, see Fig. 5.2.

### *Metagenomic signal binning*

Cumulated phase signals have a linear character. Although whole bacterial genome signals are formed by two or more linear parts with breaks, random short parts of genomes are almost linear without any breaks. Therefore, a value representing the slope of a signal can be regarded as an extreme compression of such signal. At the same time, the slope, or its absolute value, of signals from the same genome should be highly similar and genome-specific. I examined this feature in the conference paper 'Signal Based Feature Selection for Fast Classification of Sequences in Metagenomics' [72]. In that study, I reconstructed unique vectors that fully represent the original DNA sequences using only slope values. Except for cumulated phase signals, I used unwrapped phase signals that also have a linear character. Theoretically, both of these signals have three different variants according to used projection of nucleotide tetrahedron (R-Y and S-W, R-Y and M-K, S-W and M-K). However, unwrapped phases signals representing S-W and M-K are identical to R-Y and M-K signals from their definition. Therefore, only five different slope values can be calculated for every sequence.

For five selected genomes of four species (*E. coli, C. C. ruddii, G. obscurus, R. prowazekii*), 500 reads were simulated as random fragments selected from the genome. Different variants of feature vectors were examined as various combinations of slope values were utilized. These vectors were binned using Ward's hierarchical clustering and Euclidean metrics. The best results were obtained for feature vectors of four values. The slope of cumulated phase representing R-Y and S-W chemical properties was omitted as its inter-species resolution is almost zero. The example of cumulated phase signals with (S-W and M-K) and without (R-Y and S-W) discriminative information is shown in Fig. 5.3.

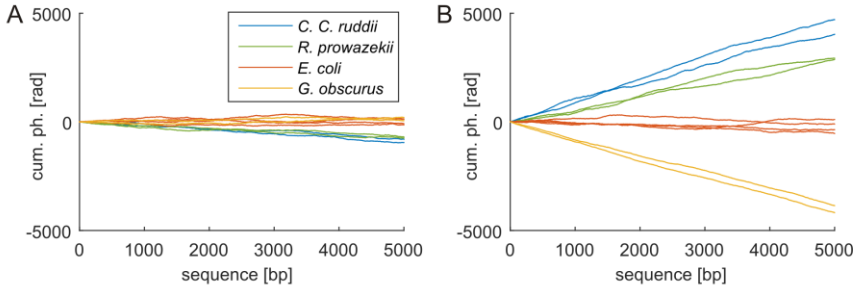


**Fig. 5.3 – Slopes of cumulated phase signals**
Cumulated phase signals for randomly selected sequences from the test dataset presenting (A) S-W and M-K information (B) R-Y and S-W information. I presented this analysis in study by Sedlar [72].

The proposed technique was unable to distinguish between different strains of the same species. This is not an issue, as there is not enough information in short sequences to compare particular strains. Satisfactory results were obtained also for different lengths of simulated reads, see statistics summary in Tab. 5.1 and Tab. 5.2.

**Tab. 5.1 Statistics for slope binning (sensitivity and specificity)**

| Organism | 500 bp | | 1000 bp | | 5000 bp | |
|---|---|---|---|---|---|---|
| | **Sensitivity** | **Specificity** | **Sensitivity** | **Specificity** | **Sensitivity** | **Specificity** |
| *E. coli* | 98,28 | 94,64 | 99,78 | 93,63 | 99,90 | 99,60 |
| *C. C.ruddii* | 75,00 | 98,14 | 97,45 | 97,98 | 100,00 | 99,21 |
| *G. obscurus* | 99,80 | 99,60 | 99,40 | 99,95 | 100,00 | 100,00 |
| *R. prowazekii* | 74,29 | 92,08 | 77,67 | 99,31 | 95,78 | 99,95 |
| **Average** | **86,84** | **96,12** | **93,58** | **97,72** | **98,92** | **99,69** |

**Tab. 5.2 Statistics for slope binning (precision and recall)**

| Organism | 500 bp | | 1000 bp | | 5000 bp | |
|---|---|---|---|---|---|---|
| | **Precision** | **Accuracy** | **Precision** | **Accuracy** | **Precision** | **Accuracy** |
| *E. coli* | 91,60 | 96,00 | 89,80 | 95,84 | 99,40 | 99,72 |
| *C. C.ruddii* | 93,00 | 92,40 | 91,80 | 97,88 | 96,80 | 99,36 |
| *G. obscurus* | 98,40 | 99,64 | 99,80 | 99,84 | 100,00 | 100,00 |
| *R. prowazekii* | 67,60 | 88,84 | 97,40 | 93,88 | 99,80 | 99,08 |
| **Average** | **87,65** | **94,22** | **94,70** | **96,86** | **99,00** | **99,54** |

Although the results seem to be very promising, the diversity of the test dataset is extremely low. The accuracy of the slope binning for real datasets is questionable, yet from the nature of the method comparable to composition based binning techniques using simple parameters like GC content. The main advantage of the method is the length of the feature vector. Vectors of four values can be easily binned using almost any technique for hierarchical or non-hierarchical clustering or any machine learning and datamining algorithms, including *Twister Tries* [73] with linear time complexity. On the other hand, data cannot be directly visualized due to four dimensions.

The analyses of genomic signals summarized in paragraphs above utilized cumulative information of cumulated and unwrapped phase and presumed that a whole genome is one period of an infinite periodic signal. In fact, this periodicity is caused by cumulating purines and pyrimidines that tend to be balanced according to the second Chargaff's rule. Thus, the signal of a whole genome can be considered as a stationary signal. This presumption is not met for random chunks of a genome where purine and pyrimidines are not balanced and cumulated signals have a slope. When the slope is

omitted, i.e. a phase signal is used, the signal seems to be chaotic and non-stationary. Possibly, only slight periodicities may occur to the naked eye. Nevertheless, these periodicities are genome-specific and may be described by techniques for non-stationary signal processing. As phase signals representing R-Y and S-W chemical properties are similar to EEG signals, utilization of Hjorth descriptors seems to be ideal.

I presented the idea of metagenome visualization by Hjorth description in Falling Walls Lab competition. The winning idea of Falling Walls Prague 2016 is based on the projection of metagenomics sequences into 3D space using three Hjorth descriptors: activity, mobility, and complexity. The visualization of the metagenome dataset EqualSet01 defined by Laczny et al. [74] is shown in Fig. 5.4.

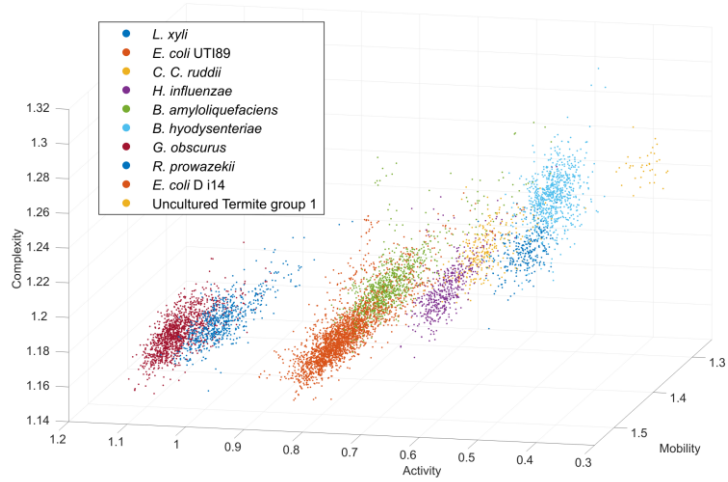


**Fig. 5.4 – Visualization of a metagenome using Hjorth descriptors**
Visualization of EqualSet01 metagenome using Hjorth descriptors extracted from phase R-Y and S-W representation of 5,000 bp long fragments.

Unlike the visualization using dimensionality reduction of *k*-mer frequencies, the visualization using Hjorth descriptors is deterministic and maximally efficient (linear time complexity). Finally, the transformed data in which every sequence is represented by the vector of three values can be automatically binned by any technique. The results with very high accuracy (89.17% - 99.84%) can be achieved by clustering using Gaussian mixture model combined with expectation maximization algorithm as proved in the diploma thesis 'Methods for fast sequence comparison and identification in metagenomic data' I have supervised by Kristyna Kupkova [75].

### *Nanopore data processing*

Squiggles, native nanopore sequencing signals, are current signals similar to other biological signals including EEG. They can be directly processed by Hjorth descriptors as me and my colleagues presented using synthetic metagenome [76], see Fig. 5.5.

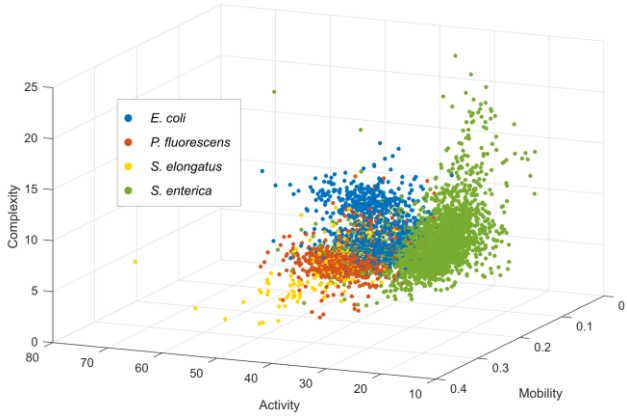


**Fig. 5.5 – Visualization of synthetic nanopore metagenome**
Visualization of PRJEB8716 using Hjorth descriptors extracted from squiggles produced by Oxford Nanopore MinION device. Color-coding distinguishes signals from different species. My colleagues and I presented this analysis in poster by Kupkova et al. [237].

## 5.3 Summary

The methodology and results presented in the fifth chapter corresponded to the second objective of the thesis – to develop a signal processing method for whole metagenome shotgun sequencing data binning.

While genomic signal processing is an established field of bioinformatics, standard character processing techniques are still much more developed. On the other hand, signal processing demonstrated great efficiency for comparison of organisms by sequence alignment dependent as well as alignment independent techniques. Yet, GSP was not utilized in metagenomics until now. The key for its successful application lies in filtering redundant information that does not contribute to species-specific patterns. As analyses presented in this thesis showed, variants of phase signals utilizing chemical similarities of nucleotides are highly redundant from short sequences of a single gene to whole bacterial genomes. Thus, phase signals and their cumulated and unwrapped variant have ideal features for fast comparison and identification of sequences in metagenomes.

# CONCLUSION

It is 20 years since the term ‘metagenomics’ was introduced. During that time, metagenomics underwent a massive development and differentiation into sub-disciplines. Moreover, this rapid development of the field still persists and notes of advancement are often scattered in various review papers. Therefore, in the theoretical part of this PhD thesis, I decided to summarize basic facts regarding metagenomics in order that follows typical analysis of a microbiome from my point of view. I tried to explain the methods using graphic illustrations, examples, and case studies. A majority of illustrations was prepared exclusively for this thesis. Where it was possible, I included examples from studies I authored or co-authored. In several cases, I used the most relevant studies from the field. Except for metagenomics, I included chapter describing computational background whose understanding is necessary for practical part of the thesis that introduces novel methodology for processing microbial data and binning of sequences in a metagenome.

The methodology for processing targeted sequencing data introduced in this thesis presumes that OTU tables fully describe microbial networks formed by organisms living in environments. Therefore, an OTU table can be regarded as a mathematical graph. This thesis brings a unique method for transformation of an OTU table into the form of a bipartite graph representing an analyzed microbial network. One partition is formed by identified taxa and the other by analyzed samples. Transformation preserves qualitative as well as quantitative information that is stored in edges of the graph. The whole method is computationally efficient and allows processing of large datasets thanks to utilization of fast algorithms for community detection.

The second area, based on whole metagenome shotgun sequencing, can be understood as ‘true’ metagenomics because data contains a whole metagenome. The most important step remains in binning, which means clustering of sequences into bins represented by sequences from related genomes. The methodology for taxonomy independent binning introduced in the thesis uses a genomic signal processing approach. Although GSP is an established field of bioinformatics, its utilization in metagenomics is completely new. The proposed methodology is the result of an analysis of phase signals variants. These signals can efficiently filter redundant information that does not contribute to binning. While cumulative signals codes this information into species-specific slope, non-cumulative and non-stationary phase signals contains patterns that can be revealed by Hjorth descriptors. Moreover, the proposed methodology can be applied directly on nanopore sequencing data in the native form of current signals.

# REFERENCES


[1] HANDELSMAN, Jo, RONDON, Michelle R., BRADY, Sean F., CLARDY, Jon and GOODMAN, Robert M. Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chemistry & Biology*. 1998. **5**(10), p. R245–R249.

[2] HUGENHOLTZ, Philip, GOEBEL, Brett M and PACE, Norman R. Impact of Culture-Independent Studies on the Emerging Phylogenetic View of Bacterial Diversity. *Journal of Bacteriology*. 1998. **180**(18), p. 4765–4774.

[3] HANDELSMAN, Jo. Metagenomics: Application of Genomics to Uncultured Microorganisms. *Microbiology and Molecular Biology Reviews*. 2004. **68**(4), p. 669–685.

[4] SIMON, Carola and DANIEL, Rolf. Metagenomic analyses: Past and future trends. *Applied and Environmental Microbiology*. 2011. **77**(4), p. 1153–1161.

[5] SENDER, Ron, FUCHS, Shai and MILO, Ron. Revised Estimates for the Number of Human and Bacteria Cells in the Body. *PLoS Biology*. 2016. **14**(8), p. e1002533.

[6] TILG, Herbert and MOSCHEN, Alexander R. Microbiota and diabetes: an evolving relationship. *Gut*. 2014. **63**(9), p. 1513–1521.

[7] DASH, Sarah, CLARKE, Gerard, BERK, Michael and JACKA, Felice N. The gut microbiome and diet in psychiatry: focus on depression. *Current opinion in psychiatry*. 2015. **28**(1), p. 1–6.

[8] CABRERA-RUBIO, Raúl, GARCIA-NÚÑEZ, Marian, SETÓ, Laia, ANTÓ, Josep M, MOYA, Andrés, MONSÓ, Eduard, et al. Microbiome diversity in the bronchial tracts of patients with chronic obstructive pulmonary disease. *Journal of Clinical Microbiology*. 2012. **50**(11), p. 3562–3568.

[9] HU, Yongfei, YANG, Xi, QIN, Junjie, LU, Na, CHENG, Gong, WU, Na, et al. Metagenome-wide analysis of antibiotic resistance genes in a large cohort of human gut microbiota. *Nature communications*. 2013. **4**, p. 2151.

[10] GERZOVA, Lenka, VIDENSKA, Petra, FALDYNOVA, Marcela, SEDLAR, Karel, PROVAZNIK, Ivo, CIZEK, Alois, et al. Characterization of microbiota composition and presence of selected antibiotic resistance genes in carriage water of ornamental fish. *PLoS ONE*. 2014. **9**(8), p. e103865.

[11] KAEVSKA, Marija, LORENCOVA, A, VIDENSKA, P, SEDLAR, K, PROVAZNIK, I and TRCKOVA, M. Effect of sodium humate and zinc oxide used in prophylaxis of post-weaning diarrhoea on faecal microbiota composition in weaned piglets. *Veterinarni Medicina*. 2016. **61**(6), p. 328–336.

[12] KAEVSKA, Marija, VIDENSKA, Petra, SEDLAR, Karel, BARTEJSOVA, Iva, KRALOVA, Alena and SLANA, Iva. Faecal bacterial composition in dairy cows shedding Mycobacterium avium subsp. paratuberculosis in faeces in comparison with nonshedding cows. *Canadian Journal of Microbiology*. 2016. **62**(6), p. 538–541.

[13] KAEVSKA, Marija, VIDENSKA, Petra, SEDLAR, Karel and SLANA, Iva. Seasonal changes in microbial community composition in river water studied using 454-pyrosequencing. *SpringerPlus*. 2016. **5**(1), p. 409.

[14] SUENAGA, Hikaru, OHNUKI, Tsutomu and MIYAZAKI, Kentaro. Functional screening of a metagenomic library for genes involved in microbial degradation of aromatic compounds. *Environmental Microbiology*. 2007. **9**(9), p. 2289–2297.

[15] IZARD, Jacques. and RIVERA, Maria Amélia Amado. *Metagenomics for microbiology*. 1st Editio. Elsevier Science, 2014. ISBN 9780124104723.

[16] TRINGE, Susannah Green and RUBIN, Edward M. Metagenomics: DNA sequencing of environmental samples. *Nature reviews. Genetics*. 2005. **6**(11), p. 805–14.

[17] MITRA, Suparna, STÄRK, Mario and HUSON, Daniel H. Analysis of 16S rRNA environmental sequences using MEGAN. *BMC Genomics*. 2011. **12**(Suppl 3), p. S17.

[18] THOMAS, Torsten, GILBERT, Jack and MEYER, Folker. Metagenomics - a guide from sampling to data analysis. *Microbial Informatics and Experimentation*. 2012. **2**(1), p. 3.

[19] NAVAS-MOLINA, José A., PERALTA-SÁNCHEZ, Juan M., GONZÁLEZ, Antonio, MCMURDIE, Paul J., VÁZQUEZ-BAEZA, Yoshiki, XU, Zhenjiang, et al. Advancing our understanding of the human microbiome using QIIME. In : *Methods in Enzymology*. 2013. p. 371–444. ISBN 9780124078635.

[20] STACKEBRANDT, Erko and GOEBEL, Brett M. Taxonomic Note: A Place for DNA-DNA Reassociation and 16S rRNA Sequence Analysis in the Present Species Definition in Bacteriology. *International Journal of Systematic and Evolutionary Microbiology*. 1994. **44**(4), p. 846–849.

[21] MAGURRAN, Anne E. *Measuring biological diversity*. Blackwell Pub, 2004. ISBN 978-0-632-05633-0.

[22] HERNANDEZ, Maria E., BECK, David A.C., LIDSTROM, Mary E. and CHISTOSERDOVA, Ludmila. Oxygen availability is a major factor in determining the composition of microbial communities involved in methane oxidation. *PeerJ*. 2015. **3**, p. e801.

[23] YATSUNENKO, Tanya, REY, Federico E., MANARY, Mark J., TREHAN, Indi, DOMINGUEZ-BELLO, Maria Gloria, CONTRERAS, Monica, et al. Human gut microbiome viewed across age and geography. *Nature*. 2012. **486**(7402), p. 222–227.

[24] VIDENSKA, Petra, SEDLAR, Karel, LUKAC, Maja, FALDYNOVA, Marcela, GERZOVA, Lenka, CEJKOVA, Darina, et al. Succession and replacement of bacterial populations in the caecum of egg laying hens over their whole life. *PLoS ONE*. 2014. **9**(12), p. e115142.

[25] GERZOVA, Lenka, BABAK, Vladimir, SEDLAR, Karel, FALDYNOVA, Marcela, VIDENSKA, Petra, CEJKOVA, Darina, et al. Characterization of antibiotic resistance gene abundance and microbiota composition in feces of organic and conventional pigs from four EU countries. *PLoS ONE*. 2015. **10**(7), p. e0132892.

[26] VIDENSKA, Petra, RAHMAN, Md. Masudur, FALDYNOVA, Marcela, BABAK, Vladimir, MATULOVA, Marta Elsheimer, PRUKNER-RADOVCIC, Estella, et al. Characterization of egg laying hen and broiler fecal microbiota in poultry farms in Croatia, Czech Republic, Hungary and Slovenia. *PLoS ONE*. 2014. **9**(10), p. e110076.

[27] GHURYE, Jay S, CEPEDA-ESPINOZA, Victoria and POP, Mihai. *Metagenomic assembly: Overview, challenges and applications*. 2016. Yale Journal of Biology and Medicine. ISBN 1551-4056 (Electronic)r0044-0086 (Linking).

[28] BREITWIESER, Florian P., LU, Jennifer and SALZBERG, Steven L. A review of methods and databases for metagenomic classification and assembly. *Briefings in Bioinformatics*. 2017. P. 1–15.

[29] NAMIKI, Toshiaki, HACHIYA, Tsuyoshi, TANAKA, Hideaki and SAKAKIBARA, Yasubumi. MetaVelvet: An extension of Velvet assembler to de novo metagenome assembly from short sequence reads. *Nucleic Acids Research*. 2012. **40**(20), p. e155–e155.

[30] PENG, Yu, LEUNG, Henry C. M., YIU, S. M. and CHIN, Francis Y. L. Meta-IDBA: A de Novo assembler for metagenomic data. *Bioinformatics*. 2011. **27**(13), p. i94–i101.

[31] VOLLMERS, John, WIEGAND, Sandra and KASTER, Anne Kristin. Comparing and evaluating metagenome assembly tools from a microbiologist's perspective - Not only size matters! *PLoS ONE*. 2017. **12**(1), p. e0169662.

[32] MANDE, Sharmila S., MOHAMMED, Monzoorul Haque and GHOSH, Tarini Shankar. Classification of metagenomic sequences: Methods and challenges. *Briefings in Bioinformatics*. 2012. **13**(6), p. 669–681.

[33] SEDLAR, Karel, KUPKOVA, Kristyna and PROVAZNIK, Ivo. Bioinformatics strategies for taxonomy independent binning and visualization of sequences in shotgun metagenomics. *Computational and Structural Biotechnology Journal*. 2017. **15**, p. 48–55.

[34] LACZNY, Cedric C, STERNAL, Tomasz, PLUGARU, Valentin, GAWRON, Piotr, ATASHPENDAR, Arash, MARGOSSIAN, Houry H., et al. VizBin - An application for reference-independent visualization and human-augmented binning of metagenomic data. *Microbiome*. 2015. **3**(1), p. 1.

[35] MEYER, Foker, PAARMANN, Daniell, D'SOUZA, Mark, OLSON, Robert, GLASS, Elizabeth M, KUBAL, Mike, et al. The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC bioinformatics*. 2008. **9**, p. 386.

[36] GOLL, Johannes, RUSCH, Douglas B., TANENBAUM, David M., THIAGARAJAN, Mathangi, LI, Kelvin, METHÉ, Barbara A., et al. METAREP: JCVI metagenomics reports-an open source tool for high-performance comparative metagenomics. *Bioinformatics*. 2010. **26**(20), p. 2631–2632.

[37] MARKOWITZ, Victor M., IVANOVA, Natalia N., SZETO, Ernest, PALANIAPPAN, Krishna, CHU, Ken, DALEVI, Daniel, et al. IMG/M: A data management and analysis system for metagenomes. *Nucleic Acids Research*. 2008. **36**(SUPPL. 1), p. D534–D538.

[38] BIGGS, Norman., LLOYD, E. Keith. and WILSON, Robin J. *Graph Theory, 1736-1936*. Clarendon Press, 1976. ISBN 9780198539162.

[39] BOLLOBÁS, Béla. *Modern graph theory*. New York, NY : Springer New York, 1999. Graduate Texts in Mathematics. ISBN 0387984887.

[40] ASRATIAN, Armen S, J., Denley Tristan M. and ROLAND, Haaggkvist. *Bipartite graphs and their applications*. Cambridge, U.K. ; New York : Cambridge University Press, 1998. ISBN 052159345X (hardback). Originally published: 1998

[41] SHARAN, Roded and IDEKER, Trey. Modeling cellular machinery through biological network comparison. *Nature Biotechnology*. 2006. **24**(4), p. 427–433.

[42] MASHAGHI, Alireza R., RAMEZANPOUR, Abolfazl and KARIMIPOUR, Vahid. Investigation of a protein complex network. *The European Physical Journal B*. 2004. **41**(1), p. 113–121.

[43] VECCHIO, Fabrizio, MIRAGLIA, Francesca, PILUDU, Francesca, GRANATA, Giuseppe, ROMANELLO, Roberto, CAULO, Massimo, et al. "Small World" architecture in brain connectivity and hippocampal volume in Alzheimer's disease: a study via graph theory from EEG data. *Brain Imaging and Behavior*. 2017. **11**(2), p. 473–485.

[44] BARBERÁN, Albert, BATES, Scott T, CASAMAYOR, Emilio O and FIERER, Noah. Using network analysis to explore co-occurrence patterns in soil microbial communities. *The ISME Journal*. 2012. **6**(2), p. 343–351.

[45] PROULX, Stephen R., PROMISLOW, Daniel E.L. and PHILLIPS, Patrick C. Network thinking in ecology and evolution. *Trends in Ecology and Evolution*. 2005. **20**(6 SPEC. ISS.), p. 345–353.

[46] CROFT, Darren P, KRAUSE, Jens and JAMES, Richard. Social networks in the guppy (Poecilia reticulata). *Proceedings of the Royal Society B: Biological Sciences*. 2004. **271**(Suppl_6), p. S516–S519.

[47] BARABÁSI, Albert László, GULBAHCE, Natali and LOSCALZO, Joseph. Network medicine: A network-based approach to human disease. *Nature Reviews Genetics*. 2011. **12**(1), p. 56–68.

[48] BASCOMPTE, Jordi. Disentangling the web of life. *Science*. 2009. **325**(5939), p. 416–419.

[49] KITANO, Hiroaki. Computational systems biology. *Nature*. 2002. **420**(6912), p. 206–210.

[50] SHEN, Huawei. *Community Structure of Complex Networks*. Springer, 2013. ISBN 978-3-642-31820-7.

[51] CROCHEMORE, Maxime and RYTTER, Wojciech. *Jewels of stringology*. WORLD SCIENTIFIC, 2002. ISBN 9810247826.

[52] LAND, Miriam, HAUSER, Loren, JUN, Se-Ran, NOOKAEW, Intawat, LEUZE, Michael R., AHN, Tae-Hyuk, et al. Insights from 20 years of bacterial genome sequencing. *Functional & integrative genomics*. 2015. **15**(2), p. 141–61.

[53] DOUGHERTY, Edward R, HUANG, Yufei, KIM, Seungchan, CAI, Xiaodong and YAMAGUCHI, Rui. Genomic signal processing. *Current genomics*. 2009. **10**(6), p. 364.

[54] ANASTASSIOU, Dimitrij. Genomic signal processing. *IEEE Signal Processing Magazine*. 2001. **18**(4), p. 8–20.

[55] JEONG, Byeong-Soo, GOLAM BARI, A.T.M., ROKEYA REAZ, Mst., JEON, Seokhee, LIM, Chae-Gyun and CHOI, Ho-Jin. Codon-based encoding for DNA sequence analysis. *Methods*. 2014. **67**(3), p. 373–379.

[56] KUNG, S. Y., LUO, Yuhui and MAK, Man Wai. Feature selection for genomic signal processing: Unsupervised, supervised and self-supervised scenarios. *Journal of Signal Processing Systems*. 2010. **61**(1), p. 3–20.

[57] DEAMER, David, AKESON, Mark and BRANTON, Daniel. Three decades of nanopore sequencing. *Nature Biotechnology*. 2016. **34**(5), p. 518–524.

[58] SEDLAR, Karel, VIDENSKA, Petra, SKUTKOVA, Helena, RYCHLIK, Ivan and PROVAZNIK, Ivo. Bipartite graphs for visualization analysis of microbiome data. *Evolutionary Bioinformatics*. 2016. **12**, p. 17–23.

[59] SHANNON, Paul, MARKIEL, Andrew, OZIER, Owen, BALIGA, Nitin S, WANG, Jonathan T, RAMAGE, Daniel, et al. Cytoscape: A software Environment for integrated models of biomolecular interaction networks. *Genome Research*. 2003. **13**(11), p. 2498–2504.

[60] BLONDEL, Vincent D, GUILLAUME, Jean Loup, LAMBIOTTE, Renaud and LEFEBVRE, Etienne. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*. 2008. **2008**(10), p. P10008.

[61] NEWMAN, M. E. J. and GIRVAN, M. Finding and evaluating community structure in networks. *Physical Review E*. 2004. **69**(2), p. 026113.

[62] RAGHAVAN, Usha Nandini, ALBERT, Réka and KUMARA, Soundar. Near linear time algorithm to detect community structures in large-scale networks. *Physical Review E*. 2007. **76**(3), p. 036106.

[63] PONS, Pascal and LATAPY, Matthieu. Computing Communities in Large Networks Using Random Walks. In : *International Symposium on Computer and Information Sciences*. Springer, Berlin, Heidelberg, 2005. p. 284–293. ISBN 3-540-29414-7, 978-3-540-29414-6.

[64] FORTUNATO, Santo. Community detection in graphs. *Physics Reports*. 2010. **486**(3–5), p. 75–174.

[65] CSARDI, Gabor and NEPUSZ, Tamas. The igraph software package for complex network research. *InterJournal*. 2006. **Complex Sy**, p. 1695.

[66] GUIMERA, Roger, SALES-PARDO, Marta and AMARAL, Luis A. Nunes. Module identification in bipartite and directed networks. *Physical Review E*. 2007. **76**(3), p. 036102.

[67] MARSHALL, Barry J, ARMSTRONG, John A, MCGECHIE, David B and GLANCY, Ross J. Attempt to fulfil Koch's postulates for pyloric campylobacter. *Medical Journal of Australia*. 1985. **142**(8), p. 436–439.

[68] SEDLAR, Karel, SKUTKOVA, Helena, VIDENSKA, Petra, RYCHLIK, Ivan and PROVAZNIK, Ivo. Bipartite graphs for metagenomic data analysis and visualization. In : *2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2015. p. 1123–1128. ISBN 978-1-4673-6799-8.

[69] SEDLAR, Karel. *The use of genomic signal compression for classification and identification of organisms*. Brno: Brno University of Technology, Faculty of Electrical Engineering and Communication. 91 p. Masters's thesis, 2013.

[70] SEDLAR, Karel, SKUTKOVA, Helena, VITEK, Martin and PROVAZNIK, Ivo. Set of rules for genomic signal downsampling. *Computers in Biology and Medicine*. 2016. **69**, p. 308–314.

[71] SEDLAR, Karel, SKUTKOVA, Helena, VITEK, Martin and PROVAZNIK, Ivo. Prokaryotic DNA signal downsampling for fast whole genome comparison. In : *Advances in Intelligent Systems and Computing*. Springer, Cham, 2014. p. 373–383. ISBN 978-3-319-06595-3.

[72] SEDLAR, Karel. Signal Based Feature Selection for Fast Classification of Sequences in Metagenomics. In : *Proceedings of the 22nd Conference STUDENT EEICT 2016*. 2016. p. 558-562. ISBN: 978-80-214-5350–0.

[73] COCHEZ, Michael and MOU, Hao. Twister Tries. In : *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data - SIGMOD '15*. New York, New York, USA : ACM Press, 2015. p. 505–517. ISBN 978-1-4503-2758-9.

[74] LACZNY, Cedric C., PINEL, Nicolás, VLASSIS, Nikos and WILMES, Paul. Alignment-free visualization of metagenomic data by nonlinear dimension reduction. *Scientific Reports*. 2014. **4**(1), p. 4516.

[75] KUPKOVA, Kristyna. *Methods for fast sequence comparison and identification in metagenomic data*. Brno: Brno University of Technology, Faculty of Electrical Engineering and Communication, 74 p. Master's thesis., 2016.

[76] KUPKOVA, Kristyna, PROVAZNIK, Ivo, SEDLAR, Karel, HON, Jiri, MARTINEK, Tomas and ZENDULKA, Jaroslav. *Visualization of Nanopore Sequencing Data in Metagenomics*. poster B-213: ISMB/ECCB 2017 BioVis session, Prague, 2017.

# C U R R I C U L U M V I T A E

## Personal information


| | |
|---|---|
| Name / Surname/ Title | **Mgr. Ing. Karel Sedlář** |
| Address | Brněnská 1424/100, 664 51, Šlapanice, Czech Republic |
| E-mail | kar.sedlar@gmail.com |
| Nationality | Czech |
| Date of birth | 28. 1. 1989 |


## Education

| | |
|---|---|
| **Doctoral study** | 9/2013 – so far, Biomedical Technologies and Bioinformatics |
| Name of the organization | Brno University of Technology, Department of Biomedical Engineering |
| Thesis theme | Methods for comparative analysis of metagenomic data |
| **Graduate study, Title** | 1/2013 – 2/2015, Applied Informatics, Mgr. |
| Name of the organization | Masaryk University, Faculty of Informatics |
| Thesis Theme | Processing of experimental data related to FGFR signaling pathways models |
| **Graduate study, Title** | 6/2011 – 6/2013, Biomedical Engineering and Bioinformatics, Ing. (graduated with honors) |
| Name of the organization | Brno University of Technology, Faculty of Electrical Engineering and Communication |
| Thesis Theme | The use of genomic signal compression for classification and identification of organisms |
| **Undergraduate study, Title** | 6/2009 – 6/2011, Biomedical Technics and Bioinformatics, Bc. (graduated with honors) |
| Name of the organization | Brno University of Technology, Faculty of Electrical Engineering and Communication |
| Thesis Theme | Bootstrap methods in phylogenetics |

## Work experience

| | |
|---|---|
| Dates, Position | 9/2013 – so far, researcher, lecturer |
| Main activities and responsibilities | research and lectures in bioinformatics and systems biology |
| Name of employer | Department of Biomedical Engineering, Brno University of Technology |
| Dates, Position | 6/2011 – 12/2013, trainee, analyst |
| Main activities and responsibilities | Roche 454 Sequencing platform data processing |
| Name of employer | Roche s.r.o, Diagnostics Division (Czech branch) |
| Dates, Position | 1/2011 – so far, intermediary, adviser |
| Main activities and responsibilities | networking opportunities in the field of bioinformatics, providing cooperation between institutions, organization of scientific seminars, next generation sequencing data processing |
| Name of employer | self-employed (Karel Sedlář) |
| Dates, Position | 10/2010 – 6/2011, technical staff |
| Main activities and responsibilities | biosensors for the detection of nucleic acids research, lab work, preparation of solutions, measurement of electrodes by cyclic voltammetry, data processing |
| Name of employer | Department of Microelectronics, Brno University of Technology |

## Scientific Awards

| | |
|---|---|
| 2017 | Supported by Foundation Zdenek et Michaela Bakala (Switzerland) during Biotech 2017 |
| 2016 | 1st place in International Falling Walls Lab 2016 Prague |
| 2015 | Awarded by Rector of Brno University of Technology for outstanding scientific results |
| 2014 | 1st place in Student EEICT 2014, category: Biomedical Engineering and Bioinformatics |
| 2014 | 1st place in ICCT 2014 competition for young scientists up to 35 years |
| 2013 | Awarded by Dean of FEEC BUT for the best diploma thesis |
| 2013 | 1st place in Student EEICT 2013, category: Biomedical Engineering and Bioinformatics |
| 2012 | 2st place in Student EEICT 2012, category: Microelectronics, Technologies and Biomedicine |
| 2011 | Awarded by Dean of FEEC BUT for the best bachelor thesis |
| 2011 | 1st place in Student workshop competition for effective cooperation of biomedical disciplines |
| 2011 | 2st place in Student EEICT 2011, category: Biomedicine and Image Processing |

## International cooperation

| | |
|---|---|
| 5/2014 – 8/2015 | Technical University of Denmark, National Food Institute, Denmark |
| *spreading of antibiotic resistance in livestock animals* | Anses, Hygiene and Quality of Poultry and Pig Products Unit, France<br>Istituto Zooprofilattico Sperimentale delle Venezie, Italy<br>National Veterinary Institute (SVA), Uppsala, Sweden |
| 10/2013 – 12/2014 | Faculty of Veterinary Medicine, University of Zagreb, Croatia |
| *microbiota composition in livestock animals* | Biotechnical Faculty, University of Ljubljana, Slovenia<br>Institute for Veterinary Medical Research, Hungarian Academy of Sciences, Hungary |

## Training courses

| | |
|---|---|
| 10/2017 | Excellence Science Days 2.0, Wroclaw, Poland |
| 4/2017 | EMBL Course: Advanced RNA-Seq and ChiP-Seq Data Analysis 2017, Hinxton, UK |
| 7/2014 | SBSS 2014: Systems Biology Summer School 2014, Nove Hrady, Czech Republic |
| 7/2013 | SBSS 2013: Systems Biology Summer School 2013, Nove Hrady, Czech Republic |
| 2/2013 | ICRC seminar of clinical research, Brno, Czech Republic |
| 10/2011 | Roche Genome Sequencer Software – summer school 2011, Penzberg, Germany |

## Academic/scientific activities

| | |
|---|---|
| Scientific impact | No. of records in WoS: 23, h-index: 6, sum of times cited: 101 |
| Researcher ID | K-1120-2014 |
| ORCID | 0000-0002-8269-4020 |
| Scopus Author | 56309904900 |
| Selected publications | **Sedlar K**, Kolek J, Skutkova H, Branska B, Provaznik I, Patakova P. Complete genome sequence of Clostridium pasteurianum NRRL B-598, a non- type strain producing butanol. *Journal of Biotechnology*. 2015. 214(1): 113-114.<br>Patakova P, Kolek J, **Sedlar K**, Koscova P, Branska B, Kupkova K, Paulova L, Provaznik I. Comparative analysis of high butanol tolerance and production in clostridia. *Biotechnology Advances*. 2017, 35(8): 1-38.<br>**Sedlar K**, Videnska P, Skutkova H, Rychlik I, Provaznik I. Bipartite Graphs for Visualization Analysis of Microbiome Data. *Evolutionary Bioinformatics*. 2016. 12(S1): 17-23.<br>**Sedlar K**, Skutkova H, Vitek M, Provaznik I. Set of rules for genomic signal downsampling. *Computers in Biology and Medicine*. 2015. 64(p1): 1-7. (IF=1.521)<br>Kolek J, **Sedlar K**, Provaznik I, Patakova P. Dam and Dcm methylations prevent gene transfer into Clostridium pasteurianum NRRL B-598: development of methods for electrotransformation, conjugation, and sonoporation. *Biotechnology for Biofuels*. 2016. 9(1): 1-14. (IF=6.444)<br>**Sedlar K**, Kupkova K, Provaznik I. Bioinformatics strategies for taxonomy independent binning and visualization of sequences in shotgun metagenomics. *Computational and Structural Biotechnology Journal*. 2017. 15, 48–55. (IF 4.148) |

| | |
|---|---|
| Projects | Advanced methods for fast identification and visualization of metagenomic data, BUT funding agency no. FEKT/FIT-J-16-3335 (2016) – principal investigator |
| | Relationship between butanol efflux and butanol tolerance of Clostridia, Grant Agency of the Czech Republic GA17-00551S (2017-2019) – key member |
| | High throughput bacterial genome assembly and annotation techniques using genomic signal processing, Grant Agency of the Czech Republic 17-01821S (2017-2019) – key member |
| | Development of interdisciplinary doctoral study programme Biomedical technologies and bioinformatics, OP VVV – Grant of EU CZ.02.2.69/0.0/0.0/16_018/0002582 (2017-2022) - member |
| | Nano-electro-bio-tools for biochemical and molecular biology studies of eukaryotic cells (NanoBioTECell), Grant Agency of the Czech Republic GAP102/11/1068 (2011-2015) – member |
| Academic functions | 11/2014 – 10/2017, senator in faculty academic senate, member of financial committee |
| Supervisor | No. of diploma theses supervised: 4 |
| | No. of bachelor theses supervised: 6 |
| Lecturer | Programming in bioinformatics – lectures and practical exercises (in R language) |
| | Systems biology – lectures and practical exercises |
| | Analysis of biological sequences – practical exercises |
| Invited talks | Bioinformatics in science and research – CEPIN lectures, University of Defense |
| | Bioinformatics for biotechnology research: data mining of Clostridium pasteurianum genome – Biotech 2014, Czech biotechnology society |
| | Visualizations in metagenomics, Lectures in practical bioinformatics, Charles University |

## Bioinformatics, IT

| | |
|---|---|
| Programming skills | general: mainly scripting languages, basics in object oriented paradigm |
| | good knowledge: R, R/Bioconductior, Matlab |
| | basic knowledge: Python, Bash, C++ |
| Statistics | SPSS, WEKA, Statistica |
| Operating systems | Windows, Linux (actively using both 50:50) |
| Other skills | active user of grid computing (Metacentrum - Czech national grid infrastructure) |
| | user knowledge of portable batch computing |
| | maintenance of departmental cluster (12 nodes, 2× Quad Core Intel Xeon E5430 2.66 GHz, 32 GB RAM, 2 x 1 TB HDD RAID0 each) |
| | NGS data processing: fastqc, multiqc |
| | assembly: Celera, Velvet, Trinity, etc. |
| | mapping: bowtie, star, segemehl |
| | RNA-Seq data processing, metagenomics |

## Languages

| | |
|---|---|
| Native language | Czech |
| Other language | English (C1, Cambridge English: Advanced) |
| | French (B1) |
| | German (A2) |

## List of own publications mentioned in the thesis:

**Articles in international journals**

GERZOVA, Lenka, VIDENSKA, Petra, FALDYNOVA, Marcela, **SEDLAR, Karel**, PROVAZNIK, Ivo, CIZEK, Alois and RYCHLIK, Ivan. Characterization of Microbiota Composition and Presence of Selected Antibiotic Resistance Genes in Carriage Water of Ornamental Fish. *PLoS ONE*. 2014, **9**(8), e103865. (IF 2.766)

KAEVSKA, Marija, LORENCOVA, Alena, VIDENSKA, Petra, **SEDLAR, Karel**, PROVAZNIK, Ivo and TRCKOVA, Martina. Effect of sodium humate and zinc oxide used in prophylaxis of post-weaning diarrhoea on faecal microbiota composition in weaned piglets. *Veterinární Medicína*. 2016, **61**(6), 328-336. (IF 0.434)

KAEVSKA, Marija, VIDENSKA, Petra, **SEDLAR, Karel**, BARTEJSOVA, Iva, KRALOVA, Alena and SLANA, Iva. Faecal bacterial composition in dairy cows shedding Mycobacterium avium subsp. paratuberculosis in faeces in comparison with nonshedding cows. *Canadian Journal of Microbiology*. 2016, **62**(6), 538-541. (IF 1.243)

KAEVSKA, Marija, VIDENSKA, Petra, **SEDLAR, Karel** and SLANA, Iva. Seasonal changes in microbial community composition in river water studied using 454-pyrosequencing. *SpringerPlus*. 2016. **5**(1), 409. (IF 1.130)

**SEDLAR, Karel**, KUPKOVA, Kristyna and PROVAZNIK, Ivo. Bioinformatics strategies for taxonomy independent binning and visualization of sequences in shotgun metagenomics. *Computational and Structural Biotechnology Journal*. 2017. **15**, 48–55. (IF 4.148)

VIDENSKA, Petra, **SEDLAR, Karel**, LUKAC, Maja, FALDYNOVA, Marcela, GERZOVA, Lenka, CEJKOVA, Darina, SISAK, Frantisek and RYCHLIK, Ivan. Succession and replacement of bacterial populations in the caecum of egg laying hens over their whole life. *PLoS ONE*. 2014. **9**(12), e115142. (IF 2.766)

GERZOVA, Lenka, BABAK, Vladimir, **SEDLAR, Karel**, FALDYNOVA, Marcela, VIDENSKA, Petra, CEJKOVA, Darina, JENSEN, Annette N, DENIS, Martine, KEROUATON, Annaelle, RICCI, Antonia, CIBIN, Veronica, OSTERBERG, Julia and RYCHLIK, Ivan. Characterization of antibiotic resistance gene abundance and microbiota composition in feces of organic and conventional pigs from four EU countries. *PLoS ONE*. 2015. **10**(7), e0132892. (IF 2.766)

VIDENSKA, Petra, RAHMAN, Md. Masudur, FALDYNOVA, Marcela, BABAK, Vladimir, EILSHEIMER MATULOVA, Marta, PRUKNER-RADOVCIC, Estella, KRIZEK, Ivan, SMOLE-MOZINA, Sonja, KOVAC, Jasna, SZMOLKA, Ama, NAGY, Bela, **SEDLAR, Karel**, CEJKOVA, Darina and RYCHLIK, Ivan. Characterization of egg laying hen and broiler fecal microbiota in poultry farms in Croatia, Czech Republic, Hungary and Slovenia. *PLoS ONE*. 2014. **9**(10), e110076. (IF 2.766)

SKUTKOVA, Helena, VITEK, Martin, **SEDLAR, Karel** and PROVAZNIK, Ivo. Progressive alignment of genomic signals by multiple dynamic time warping. *Journal of Theoretical Biology*. 2015. **385**, 20–30. (IF 1.833)

**SEDLAR, Karel**, SKUTKOVA, Helena, VITEK, Martin and PROVAZNIK, Ivo. Set of rules for genomic signal downsampling. *Computers in Biology and Medicine*. 2016. **69**, 308–314. (IF 2.115)

**SEDLAR, Karel**, VIDENSKA, Petra, SKUTKOVA, Helena, RYCHLIK, Ivan and PROVAZNIK, Ivo. Bipartite graphs for visualization analysis of microbiome data. *Evolutionary Bioinformatics*. 2016. **12**, p. 17–23. (IF 1.877)

**Book chapters**

**SEDLAR, Karel**, SKUTKOVA, Helena, VITEK, Martin and PROVAZNIK, Ivo. Prokaryotic DNA signal downsampling for fast whole genome comparison. In: *Advances in Intelligent Systems and Computing*. Springer, Cham, 2014. p. 373–383. ISBN 978-3-319-06595-3.

**Conference proceedings**

**SEDLAR, Karel,** SKUTKOVA, Helena, VIDENSKA, Petra, RYCHLIK, Ivan and PROVAZNIK, Ivo. Bipartite graphs for metagenomic data analysis and visualization. In: *2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2015. 1123–1128. ISBN 978-1-4673-6799-8.

MADERANKOVA, Denisa, **SEDLAR, Karel**, VITEK, Martin and SKUTKOVA, Helena. The identification of replication origin in bacterial genomes by cumulated phase signal. In: *2017 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*. IEEE, 2017. 1–5. ISBN 978-1-4673-8988-4.

**SEDLAR, Karel**. Signal Based Feature Selection for Fast Classification of Sequences in Metagenomics. In: *Proceedings of the 22nd Conference STUDENT EEICT 2016*. 2016. 558-562. ISBN: 978-80-214-5350–0.

KUPKOVA, Kristyna, **SEDLAR, Karel** and PROVAZNIK, Ivo. Reference-free Identification of Phage DNA Using Signal Processing on Nanopore Data. In: *2017 IEEE 17th International Conference on Bioinformatics and Bioengineering*. IEEE, 2017. 101–105. ISBN 978-1-5386-1324-5.

**Abstracts and posters**

VIDENSKA, Petra, ZWINSOVA, Barbora, SMERKOVA, Kristyna, MICENKOVA, Lenka, **SEDLAR, Karel** and BUDINSKA, Eva. *Comparison of sampling kits and DNA isolation kits from stool samples*. poster

KUPKOVA, Kristyna, PROVAZNIK, Ivo, **SEDLAR, Karel**, HON, Jiri, MARTINEK, Tomas and ZENDULKA, Jaroslav. *Visualization of Nanopore Sequencing Data in Metagenomics*. poster B-213: ISMB/ECCB 2017 BioVis session, Prague, 2017.