Visual Exploration of Microbial Populations

Sara Johansson Fernstad† C-Research, Linköping University, Sweden
Jimmy Johansson‡ C-Research, Linköping University, Sweden
Suzi Adams¶ Unilever Discover Port Sunlight, UK
Jane Shaw‡ Unilever Discover Port Sunlight, UK

ABSTRACT
Studies of the ecology of microbial populations are increasingly common within many research areas as the field of microbiomics develops rapidly. The study of the ecology in sampled microbial populations generates high dimensional data sets. Although many analysis methods are available for examination of such data, a tailored tool was required to fulfill the need of interactivity and flexibility for microbiologists. In this paper, MicrobiVis is presented. It is a tool for visual exploration and interactive analysis of microbial populations. MicrobiVis has been designed in close collaboration with end users. It extends previous interactive systems for explorative dimensionality reduction by including a range of domain relevant features. It contributes a flexible and explorative dimensionality reduction as well as a visual and interactive environment for examination of data subsets. By combining information visualization and methods based on analytic tasks common in microbiology as a means for gaining new and relevant insights. The utility of MicrobiVis is demonstrated through a use case describing how a microbiologist may use the system for a visual analysis of a microbial data set. Its usability and potential is indicated through positive feedback from the current end users.

Index Terms: H.4.m [Information Systems Applications]: Miscellaneous; J.3 [Life and Medical Science] — Biology and genetics.

1 INTRODUCTION
Microbiology is an enduring and rapidly developing research area and as new and more efficient methods for DNA analysis are becoming available, huge amounts of high dimensional data are gathered. The ability to determine the microbiome without a need to culture provides detection of many more microbes than are found by traditional methods. Thus opens a new realm of exploration for microbiologists wishing to understand the full present population of microbes within an environment. Studies occur within many environments, with a demand for innovative and specialized analysis tools as the sequencing capability and available data output increases rapidly. Within consumer product development, as performed by our industrial partner, studies are carried out to understand and define the ecology in microbial populations, and thereby develop new and innovative products for home and personal care. Extracted microbial DNA from sites of the human body is analysed to examine the processes driving differences between microbial populations in varying sample sites or in different subjects, and to define which processes may be associated with a measured outcome. The microbial diversity in the human body is rich and a single sample may often include hundreds of different microbial species. Thus, data sets are large in terms of dimensionality, and efficient methods for analysis of high dimensional data are important to gain relevant insights. Although many tools are available for analysis and reduction of high dimensional data, as well as tools developed for systems biology, none of these fulfill the need of explorative and interactive analysis as expressed by microbiologists in this process. Hence, we developed a system to meet this need, based on previous research in high dimensional data analysis.

This paper presents MicrobiVis, an interactive system for visual exploration of microbial populations, which has been designed in close collaboration with end users. Prior to implementing MicrobiVis, a previously developed tool for interactive dimensionality reduction, as presented in Johansson and Johansson [21], was demonstrated to the expert team. This tool was designed for generic data analysis tasks, not focusing on the specific issues involved in analysis of microbial data. However, the microbiologists appreciated the potential usefulness of the system within their domain. MicrobiVis is a further development and a substantial improvement of this system. Throughout MicrobiVis development, iterative design decisions have been taken, based on needs and feedback from end users, alongside the application of information visualization principles and a focus on effective use of task-driven visual analytics. It is worth emphasizing that MicrobiVis is designed to provide a tool for initial, rapid interactive data exploration to generate hypotheses, and is intended to be used alongside other exploratory and confirmatory analyses used by microbiologists, to complement the current toolbox.

MicrobiVis utilizes a combination of quality metrics and introduces an overall measure of interestingness using a ranking approach. Quality metrics and rank are visually represented in an interactive environment where several methods for explorative visual analysis and dimensionality reduction are available. Combining information visualization methods with a range of features and quality metrics specific for the microbiology domain, as a means for gaining new and domain relevant insights. It also provides possibilities of exploring subsets of interesting groups of microbes identified through algorithmic guidance combined with user domain knowledge. Additionally, the system provides possibilities of analysing microbes classified at three taxonomic levels commonly used in microbiology. This automatically provides a domain-relevant hierarchical dimensionality reduction, which may reveal patterns otherwise hidden within the data.

To assist analysis further, a range of interactive aids and visual representations are available, facilitating identification of patterns and microbial relationships. Moreover, the paper contributes a use case and summary of user feedback demonstrating the utility of the system. The use case describes how a microbiologist may use MicrobiVis for analysis of a microbial data set from a study of bacterial populations.

The remainder of the paper is organized as follows. Section 2 provides an introduction to microbial data in consumer product development and in section 3, related research is presented. Section 4 describes the MicrobiVis system in detail and section 5 includes a use case and feedback from microbiologists. Finally, the paper is concluded in section 6.

† e-mail: sara.johansson@it.liu.se
‡ e-mail: jimmy.johansson@it.liu.se
¶ e-mail: suzi.adams@unilever.com
§ e-mail: jane.shaw@unilever.com
‖ e-mail: david.taylor@unilever.com
2 Data and Tasks

A general goal of studying microbial data in consumer product development is to examine microbial populations in samples taken from different parts of the human body relating to that product [2, 15]. The aim is to understand and define microbial ecology and the processes driving it. In these studies, samples may be divided into different groups, for instance, varying sample locations within subjects, or due to subjects belonging to different groups. DNA is extracted from the samples and resulting data sets classified into Operational Taxonomic Units (OTUs), these being a close approximation to bacterial species detected in each sample, with associated detected counts.

Microbiology describes a hierarchy of organisms, by means of the biological classification system, with species being the basic unit. Thus an OTU will have an associated taxonomy of species, genus, family, and so on. Microbiologists often find it useful to examine patterns at different taxonomic levels, since there may be much individual variation in function at one level, but a consistent pattern visible at a higher grouping. OTU classification at different taxonomic levels is acquired through the Ribosomal Database Project (RDP) [8] which also provides confidence values of the classification. OTUs may also be unclassified. For illustration in this work, three taxonomic levels of interest were chosen: phylum, genus and species. Phylum provides a high level classification for microbes, whilst genus and species are the two lowest levels commonly available.

Common objectives of microbial studies are to explore which processes may drive differences between two or more groups and which processes may cause a certain outcome. This may, for instance, be studied by examining differences in OTU counts between groups and by identifying OTUs or combinations of OTUs that may drive differences and outcomes. Methods used currently by this project for analysing data from these kinds of studies include statistical tests, multivariate approaches and information visualization.

For the latter, parallel coordinates can be used as profile plots of the OTU counts or proportions across different samples for a selected subset of OTUs. As a method for exploration and communication of results, graph software Cytoscape [34], can be used to show bipartite networks of OTUs linked to the samples in which they are found, adapting a method by Ley et al. [24]. Visual representations of phylogenetic trees are sometimes also used to provide biological context to relationships between OTUs. To identify and display cluster patterns, Principal Components Analysis (PCA) [9] using biologically-relevant distance measures are employed. A commonly used measure is the UniFrac distance [25], designed for microbial communities and based on phylogenetic information. Using PCA provides group separation on a sample basis, however, interpretation of the principal components is not always straightforward. Additionally, the QIME pipeline [7] was used to process the sample data and provided a range of output that established the prevalence of genera in the samples. QIME provides a diverse set of outputs for visualization, for example, pie-charts, heatmaps, PCA plots, trees and networks. These visual outputs are not linked, and the user must generally view each separately.

Throughout this paper, data will be referred to using the following terminology; samples correspond to data items, and variables are referred to as OTUs at species level and genera at genus level.

3 Related Work

The domain of biological visualization is growing rapidly with initiatives such as VizBi [28] and BioVis [18] emerging to support the rapid expansion of biological data. This section will present a selection of the work within this domain which is most relevant for MicrobiVis.

3.1 Bioinformatics Visualization

A recent supplement issue of Nature Methods [27] presents reviews on visualization of biological data, and we refer to these for good overviews on visualization of biological data in general. More specifically, the Gehlenborg et al. [12] paper reviewing visualization methods for omics data and systems biology is relevant for the work presented in this paper. Microbial data are often visualized using graph layout software such as Cytoscape [34] and VisANT [17], thus work in this area is relevant to ours, although MicrobiVis does not include any graph visualization. A wide range of graph visualization tools are available and several reviews in the context of bioinformatics have been presented. For good overviews, we especially refer to Saraiya et al. [32], Suderman and Hallet [37] and Pavlopoulos et al. [29]. In addition to network visualization tools, more generic multiple linked view systems such as Spotfire [3] are also commonly used. Saraiya et al. [31] evaluates five such visualization tools used for gene expression data, concluding that the effectiveness of tools clearly depends on data type and task. A method commonly used in microarray data analysis is hierarchical clustering. Eisen et al. [11] present a visualization system for this kind of cluster analysis, being one of the first using a combination of heatmaps and dendograms as visual representations in this context. The Hierarchical Clustering Explorer [33] uses similar visual representations with support for dynamic querying and includes a range of interactive features. Several systems use traditional information visualization methods for analysis of biological data and our work has been inspired by this. Some examples are GeneSpring [13] which includes a range of representations for visual analysis, and Hochheiser et al. [16] describing how a system for time series data analysis including parallel coordinates [20] and histogram displays can be used for analysis of microarray data. Parallel coordinates has also been used by for instance Dietzsch et al. [10], who present an application for analysis of gene expression data, using common visualization methods to display gene expression values under different experimental conditions together with statistical measures. Another example is GeneShelf, presented by Kim et al. [22], which is a web-based application for exploration of microarray data using multiple views of combined parallel coordinates and histograms. Barsky et al. [5] combines parallel coordinates and interaction graphs in Cerebral, a visualization plugin for Cytoscape providing analysis of biological systems at multiple experimental conditions.

3.2 High Dimensional Data Visualization

Biological data are often high dimensional, and the efficiency of traditional information visualization methods, such as parallel coordinates, scatter plot matrix [6] and table lens [30], is rapidly reduced as the number of variables increase. A common way of dealing with high dimensional data is to perform dimensionality reduction. Automated methods such as Self-organizing Maps [23], Multi-Dimensional Scaling and PCA [9] effectively project data sets including hundreds of variables onto a low dimensional space. However, fully automated methods are limited in not utilizing the users’ knowledge. In many cases, semi-automated methods and interactive systems may offer advantageous environments for explorative analyses. Several interactive systems for dimensionality reduction are available, using different metrics for preserving and identifying structures. A full review of systems related to MicrobiVis can be found in Johansson and Johansson [21]. Additionally, several recent papers [4, 36, 38] have presented various approaches for projection of high dimensional data using quality metrics. In terms of dimensionality reduction, the basis of MicrobiVis development has been the concept of combining several quality metrics and providing guidance through algorithmic methods, facilitating an explorative analysis process. Previous work on combining quality metrics for dimensionality reduction has been presented by Jo-
hansson and Johansson [21], where several quality metrics are combined using a weighted sum of metrics. Through a visual display the user is guided in terms of selecting the appropriate number of variables to retain. Ingram et al. [19] presents DimStiller, which takes the concept of user-guidance in the dimensionality reduction process further by carrying out dimensionality reduction and analysis as a chain of stepwise transformations controlled by the user.

Feedback from end users has indicated the usefulness of providing a single value as a summary for the most interesting variables. However, using a weighted sum of metrics the user may find it difficult to estimate an appropriate set of weights as they must consider the relative importance of a set of statistical concepts. They may also be interested in exploring the effect of individual metrics. Based on this the MicrobiVis system combines a set of quality metrics into a single measure using a ranking approach, and provides visual representation of quality metrics and ranks to provide guidance in an explorative dimensionality reduction process. Furthermore, it is tailored to specific tasks required by microbiologists, incorporating biologically-motivated quality metrics and visual features specific for the domain.

4 MicrobiVis

The main concept of MicrobiVis is to provide visual exploration and dimensionality reduction guided, but not restrained, by algorithmic analysis. The work flow of MicrobiVis can be separated into two main parts; Algorithmic Analysis and Visual Analysis, as displayed in figure 1. The algorithmic analysis is preceded by a user-defined and task guided selection of quality metrics used for analysis of the data set. Two default metric setups are available, based on domain relevant tasks, which are easily modified by the user through a selection interface. MicrobiVis was implemented using C# and DirectX.

During algorithmic analysis, the data set is analysed according to the selected quality metrics, obtaining a set of interestingness measures for the OTUs of the high dimensional data set. The OTUs are ranked, based on the set of measures, to obtain a single measure of interestingness for each OTU. The visual analysis is based on the results of the algorithmic analysis and provides interactive exploration and visual overview of structures within the full data set as well as subsets of it. Through linked visual representations a user may interactively select OTU subsets to explore using a range of features, some of which are domain specific. Selected OTU subsets as well as quality metrics and rank can be exported for further analysis using other tools.

4.1 Quality Metrics and Ranking

A key principle of quality metric-driven dimensionality reduction is providing a reduction using several measures of interestingness. The quality metrics act individually as measures of each OTU's involvement in a specific structure within the data set. During quality metric analysis, one quality value is computed for each metric and each OTU in the high dimensional data set. The metrics relating to general statistical properties, as described in detail in Johansson and Johansson [21], are retained in MicrobiVis and extended with a set of domain specific metrics based on measures used within microbiology and selected in collaboration with microbiologists, as follows.

**Abundance:** represents the total amount of an OTU for all samples. It is often represented using a logarithmic scale and hence the abundance quality value of OTU $x_j$ is computed as $Q_{abundance}(x_j) = \log_{10}(\sum_{i=1}^{N} x_{i,j})$, where $N$ is the number of samples.

**Prevalence:** represents the number of samples in the data set in which the OTUs are detected. Prevalence is represented as a percentage of the total number of samples and the prevalence quality value of an OTU is hence computed as the relative number of samples for which the count is above zero.

As described in section 2, classes of samples are often defined, and differences between classes may be of interest. Due to this, two class-based metrics are available, aiming to assign high quality values to OTUs where the difference between classes is significant. The **prevalence class difference** of an OTU is computed as the mean difference between the prevalence of all sample classes for that OTU. **Abundance class difference** is computed similarly, but using the logarithm of the difference. These are examples of simple descriptive statistics for an initial implementation, but others could be used. Recent and ongoing work on distributions for microbiome data [35] should inform future implementations for the designer to supply appropriate statistical descriptors.

**Confidence in ID:** represents a confidence value whether an OTU is correctly classified. Confidence values are obtained through classification using RDP [8], which generates classification information at each taxonomic level as well as confidence values for each OTU. The RDP output can be loaded into MicrobiVis and the information is used both as a confidence in ID quality metric and to create a hierarchical structure based on classification at phylum, genus and species level.

**Pre-calculated metrics:** can be loaded to make use of the computational strengths of existing software packages. Two kinds of tables are loadable, either including fully computed quality values with one value per OTU for each loaded metric, or a matrix of pairwise relationships between OTUs. In the second case individual quality values are extracted from the matrix, defining the quality value of OTU $x_j$ as $Q_{pairwise}(x_j) = \sum_{k=1}^{M} v(x_j, x_k)$ for $k \neq j$, where $v(x_j, x_k)$ is the value of the pairwise relationship of OTUs $x_j$ and $x_k$, $k = 1, \ldots, M$ and $M$ is the number of OTUs.

**Ranking:** Although examination of several quality metrics individually may be both useful and straightforward, the extraction of an overall measure of interestingness using a combination of quality metrics has the advantage of providing a single value as a summary for the most interesting OTUs. Within MicrobiVis, a ranking algorithm based on the non-domination principle [14] is used to extract an overall measure of interestingness based on the quality values, and is formally defined as follows. If $(\forall m)(q_{m,j} \leq q_{m,k}) \land (\exists m)(q_{m,j} < q_{m,k})$, where $\vec{q}_j$ and $\vec{q}_k$ are vectors containing the quality values of OTUs $x_j$ and $x_k$, then it is said that $\vec{q}_j$ dominates $\vec{q}_k$. The algorithm assigns high ranks to OTUs with high quality metrics and assigns low rank to OTUs with the same rank have a level of equivalence in their metric profile. For illustration, four OTUs $(\vec{x}_1, \vec{x}_2, \vec{x}_3, \vec{x}_4)$ are assigned ranks based on two metrics $(m_1, m_2)$, as displayed in figure 2. OTU $\vec{x}_4$ has the lowest values for both
$m_1$ and $m_2$, and receives a rank of one. In the remaining unranked points, $\vec{x}_b$ has the lowest value of $m_1$, and $\vec{x}_c$ the lowest value of $m_2$, but neither has the smallest value in both metrics. Thus $\vec{x}_b$ and $\vec{x}_c$ both receive rank two. OTU $\vec{x}_d$ has higher values in both metrics and receives a rank of three.

The computationally heavy part of the work flow is the quality metrics analysis. Depending on data set size and quality metrics used the computation time may typically vary between a few seconds and up to a couple of minutes using a laptop with an Intel i5 2.53GHz CPU and 4GB RAM. All subsequent interactions and computations are performed within milliseconds.

4.2 Visualization

The visual analysis environment of MicrobiVis enables explorative dimensionality reduction and visual exploration of the microbial population. It includes a primary window containing two views, as displayed in figure 3(a). In the bottom view of the primary window is the Data view, displaying the microbial population using a visual representation selected by the user. The default representation is parallel coordinates, in which axes represent OTUs and polylines represent samples. Selectable representations are scatter plot matrix and table lens. When the data set is reduced, the data view is instantly updated, displaying only the selected subset of OTUs, as displayed in figure 3(b). The Ranking and Quality view (RaQ view) in the top part of the primary window displays quality values and ranks of OTUs, using parallel coordinates where the polylines represent OTUs. Hence, one axis in the data view corresponds to one polyline in the RaQ view. The leftmost axis in the RaQ view represents OTU ranks and the remainder of the axes represent quality metrics. The RaQ view provides an easily interpreted overview of the quality metric profiles of the OTUs and acts as a visual guidance for identification of OTUs differing from the overall structures or OTUs with specific desirable or undesirable properties. To emphasize the relationship between the views the same colour scheme is used for polylines in the RaQ view and axes in the data view. Additionally, a phylogenetic tree can be loaded into MicrobiVis utilizing the commonly used Newick format. It is displayed in the Tree view below the data view, as shown in figure 3(b), each leaf node corresponding to one OTU. While selecting OTU subsets, leaves corresponding to selected OTUs are highlighted, preserving a biologically relevant context for selected OTUs.

4.3 User Interface and Interaction

MicrobiVis includes a range of features for interactive and exploratory dimensionality reduction. It furthermore includes a set of features for visual exploration.

4.3.1 Dimensionality Reduction

Since microbial data often may include hundreds of OTUs, the dimensionality reduction features of MicrobiVis are an important part of the system. Interactive selection of subsets of OTUs, guided by quality metrics and ranking, facilitates identification of potentially interesting structures and the forming of hypotheses. The selected subset of OTUs is instantly displayed in the data view, providing visual feedback and fast confirmation as to whether it may be of interest to analyse the current subset further. The primary OTU selection is provided through filtering along the axes in the RaQ view, retaining only OTUs fulfilling the query defined by the combined thresholds on quality metrics and rank. However, by manually selecting OTUs known to be of specific interest, for instance, due to some specific property or based on domain knowledge, these are retained within the selected subset although not within the filter thresholds. Selected OTUs are highlighted in blue in all views. Microbial data may often be successfully analysed at different taxonomic levels, since structures sometimes may be more visible at certain levels. If data analysed using RDP have been loaded into MicrobiVis the system automatically creates a hierarchical structure based on OTU classification at phylum, genus and species level. Analysis on a higher rank automatically provides a domain specific dimensionality reduction since phylum and genus classes include a set of OTUs. Using this method for dimensionality reduction, the OTUs belonging to the same class are merged and represented by their mean in the data view. Quality values and ranks are then recomputed, employing the algorithms described in section 4.1 and by using the average values of any pre-calculated metrics. As will be demonstrated in the use case, the analysts may be interested in examining all OTUs belonging to a specific genus. Due to this, an additional dimensionality reduction based on RDP classification is available. For example, by selecting one or a set of genera from a menu, all OTUs not belonging to selected genera are removed from the data view, providing a quick and interactive selection of OTU subsets based on domain relevant classification.

4.3.2 Subset Exploration

In addition to dimensionality reduction, MicrobiVis includes a set of features for analysis of data subsets. Three different visual representations (parallel coordinates, scatter plot matrix and table lens) are available in the data view, enabling analysis of different structures in the selected data subset. When sample class information is available, this information can be visually represented using one colour for each class in the data view, enabling fast identification of class differences. Furthermore, the average profiles of the sample classes can be displayed through additional dashed lines in the parallel coordinates, as visible in figure 3(b). Sample class information can also be displayed through tooltips while hovering over samples in the data view. If several sample classifications exist, such as group, subject ID and sample number, the information of all classifications can be displayed concurrently in the tooltip. Since much analysis is based on genus-level classification, MicrobiVis lets the user interactively select at OTU level whether the axis names in the data view should display OTU IDs or genus classification. In addition, classification information and OTU IDs are displayed as tooltips while hovering the axis headers in the data view or the leaf nodes in the tree view. Below the data view, two check boxes are available, one used to normalize the data using a global maximum and minimum value for all of the displayed OTUs, providing a straightforward means for comparing abundances, and the other used to normalize the OTUs according to domain specific intensity values which may be separately loaded. The variable order has a large impact on how easily structures are perceived in a visual representation and to facilitate analysis a set of OTU ordering algorithms are available in MicrobiVis. OTUs can be ordered in descending order according to the abundance or prevalence metrics or, if a phylogenetic tree is loaded, according to the structure of the tree data.

5 Result

This section will describe the usability of MicrobiVis through a use case and summary of user feedback. The use case describes how a microbiologist may analyse a high dimensional microbial data set using MicrobiVis, and gain new insights from the data by doing so.
5.1 Use Case

The data set analysed here is from a study examining oral bacterial populations [1]. The objective of the study was to compare bacterial populations in plaque collected from individual sites in individual volunteers belonging to two study populations of interest, referred to as group 1 and group 2. In this use case, data represents levels of 227 Operational Taxonomic Units (OTUs) which comprise 95% of the cumulative total population found in seventy-two samples taken from four sites in the mouth of nineteen subjects. Four samples were not included in the analysis set. The data have been anonymized in the use case due to confidentiality. The goal of analysing this data is to examine differences in microbial profiles between the groups to identify microbial processes driving differences and thereby develop innovative products for oral care. Prior to analysis using MicrobiVis the data had been processed through internal proprietary processes. The QIIME pipeline [7] was used and provided a range of output that established the prevalence of genera in the two populations.

5.1.1 Initial Analysis Using Rank

The data set is loaded into MicrobiVis together with a phylogenetic tree and an RDP file giving genus and phylum classification for the OTUs, and a level of confidence in those classifications. The analyst’s main interest is to examine differences between the two subject groups, and metrics for group separation tasks are suggested: an overall Pearson correlation metric, metrics for positive and negative Pearson correlation, a cluster metric, the abundance and prevalence metrics, and the group difference metrics for abundance and prevalence. In addition, a confidence in ID metric is automatically extracted from the RDP file. Figure 3(a) displays the primary view of MicrobiVis, showing the full data set of sample profiles across all OTUs. Since group differences are the main interest of the analyst, the samples are coloured according to group and the average lines of the groups are also displayed, as dashed lines, green representing group 1 and orange representing group 2.

As described previously, an overall measure of interestingness based on selected quality metrics is acquired automatically through non-dominated ranking. By filtering on rank alone, using the leftmost axis in the RaQ view, a subset of the potentially most interesting OTUs is shown to the analyst. She makes use of this to retain only the top two ranks, as displayed in figure 3(b). The data view is set to display the genus of the OTU as axis headers, since the analyst has some knowledge regarding general behaviours in genera based on theory and previous analyses of the data. While filtering on rank, an OTU belonging to Genus 16 stands out with high counts for group 1 samples, and low counts for group 2. Previous analysis using the Metastats analysis tool [39] has shown a significant difference in Genus 16 between the two groups [1]. It is unknown to the analyst whether the group difference was driven by all OTUs from Genus 16, or just a subset. Through the initial analysis here, one Genus 16 OTU where differences are clearly visible between groups is immediately picked out. The analyst selects the Genus 16 OTU by clicking on its header in the data view, which highlights it in all views (as displayed in figure 3(b)). The RaQ view shows that the selected OTU is moderately high in both prevalence and abundance. This is important to the analyst, as high abundance and prevalence provide some confidence that learnings may be generalizable to a whole population. Through a tooltip she quickly identifies the selected Genus 16 OTU. By use of the reduction menu to show all Genus 16, as displayed in figure 4, the microbiologist can...
investigate whether this functionality is generally found in all OTUs within this genus, or whether it is peculiar to this one. As visible from figure 4 this is a common functionality in Genus 16 but it is not found in all OTUs. Among the highlighted Genus 16 OTUs in the tree view is an unsampled leaf node (marked with a blue arrow). This is an unclassified OTU but its position in the tree indicates that it possibly should be classified as Genus 16.

5.1.2 Selection of Commonly Occurring OTUs

Since abundance and prevalence are important measures for the analyst, she interactively selects various subsets of OTUs through different filter combinations of abundance and prevalence together with corresponding group difference metrics in the RaQ view, as displayed in figure 3(c). Through this, she may interactively identify individual OTUs which are commonly occurring and where the microbial counts appear to differ significantly between group 1 and group 2. Whilst filtering, the analyst spots four OTUs with low confidence values, although high in abundance and prevalence, as highlighted in figure 3(c). They represent OTUs commonly occurring in a majority of samples, but for which the genus classification is not as confident. One of them is unclassified whereas the others belong to Genus 21 and Genus 4. The analyst makes a note about the four OTUs to investigate their classification further.

In general, it is important for the analyst to identify unclassified OTUs or OTUs where the classification confidence is low while abundance and prevalence are high, since they are commonly occurring and it may be possible to find a more confident classification. Large clusters of unclassified OTUs, which may be visible in the tree view, that are high in prevalence and abundance may represent microbes not yet identified and are hence highly interesting for a microbiologist to examine further. The analyst points out that this kind of interactive analysis providing quick feedback during an exploratory stage is a great help in prioritizing subsequent analyses.

5.1.3 Examination at Genus Level

At this point, the analyst decides to examine the data at genus level. This is the level at which previous analysis has been performed using Metastats, and through this, the analyst has knowledge of statistically significant differences between the groups in some genera. Furthermore, she has certain hypotheses based on theory regarding some genera and co-association. Figure 3(d) displays the data set at genus level. As before, the analyst filters on abundance and prevalence to provide a less cluttered view of the data for examination of group separation. While interactively adding and removing genera from the displayed subset, the analyst identifies some genera in which the separation between group 1 and group 2 is clearly visible; for instance, Genus 10 and Genus 29 (ninth axis from the left and rightmost axis in the data view, marked with blue arrows), as shown in figure 3(d). She can also examine different genera which are known to co-habit, to see if any patterns emerge in the data set. This knowledge makes it interesting and important for the analyst to be able to examine the OTUs of a specific genus more closely, to create some hypotheses on which OTUs are important for driving a functional difference between the groups. MicrobiVis provides visual exploration by rapid movement between phylum, genus and OTU level, which makes it possible to step interactively between the levels while identifying possible behaviours of interest to see their extent across a whole taxonomic level and possible phylogenetic conservation of function.

Whilst viewing the sample profiles at genus level, the analyst notices that the maximum sample counts vary a lot between the genera. Genus 13 has an unexpectedly high maximum count. The analyst turns on the global axis scaling, as displayed in figure 5, to compare all axes scaled at the global maximum and minimum count. From this view, it clearly appears that the maximum count of Genus 13 is much higher than all others, and that the maximum counts of Genus 5 and Genus 29 are also relatively high. These are values not expected by the analyst given previous knowledge.

To examine the samples with high count in more detail, she hovers the polylines in the data view to display tooltips showing such as Subject ID and Sample number. From this, it is seen that most of the samples high in Genus 13, and also some of the samples high in Genus 29, are from the same subject. This is displayed in figure 5, which shows the samples of this subject highlighted in black. This is critical knowledge for the analyst, since it means that some results relating to Genus 13 may not be generalizable to a whole population but are skewed by a subject with atypical profile for this group. The analyst points out that this possibility of quickly checking anomalies is an important strength of MicrobiVis, and although it could be checked using other methods or by manually examining the data it would be a more time-consuming process. MicrobiVis has allowed rapid checking as soon as an anomaly is noticed. She makes a note about the Subject ID to find out peculiarities in the metadata when compared to other subjects.

5.1.4 Looking for Co-association Between Genera

Genus 4 (marked with a red arrow in figure 3(d)) is thought to be involved in separating group 1 and group 2, in terms of co-habitating with Genus 10 and Genus 24. However, at the genus level, it appears as if both group 1 and group 2 may have high counts of Genus 4. To explore Genus 4 OTUs, the analyst returns to displaying data at OTU level. Here she uses the reduction menu of MicrobiVis to display only the Genus 4 OTUs, as displayed in the top part of figure 6. As visible from the polylines and the dashed group average lines in the figure, some Genus 4 OTUs are generally high in group 2, such as the fourth from the right, whereas others are not. Not all OTUs of the same genus are involved in the same processes, and since the goal of this study is to identify likely OTUs that might be present in group 1 or group 2, this is the kind of information that may be highly relevant in suggesting new research investigations to the analyst. By hovering the polylines, the analyst comes to the conclusion that this is not due to obvious atypical subject profiles.

To continue exploring the Genus 4, Genus 10, Genus 24 group, these genera are selected together in the reduction menu, as displayed in the bottom part of figure 6. One OTU identified as Genus 24 (leftmost axis) is seen to have an unexpected distribution, with high counts for group 1. Examination using tooltips suggests that this is unlikely to be a subject anomaly, and the analyst makes a note about the OTU ID to investigate. The analyst again points out how the interactivity of MicrobiVis and the possibility of quickly identifying OTUs with varying features helps in determining sub-
sequent analyses. The analyst then switches to displaying the OTUs in the data view using a scatter plot matrix, as displayed in figure 7. In this view, cells representing highly correlated OTUs are coloured in blue, through this co-occurring OTUs are easily identified. The analyst makes a note of co-occurring OTUs, comparing them with theoretical knowledge, to see if patterns in previous analysis are borne out by this data. This view allows extremely rapid exploration of correlations between pairs of OTUs and provides discussion amongst colleagues about the patterns seen.

5.1.5 Use Case Conclusions

The analyst leaves the session with a list of OTUs and samples of interest, with high priority to examine further, including OTUs that may be involved in group 1 and group 2 states, as well as OTUs that do not follow expected patterns. During the analysis process, the analysts also noted additional data that may be relevant to introduce in future analysis, such as other measurements on samples, or calculated diversity metrics \[7, 26\]. MicrobiVis is designed to allow easy inclusion of additional metrics and data during the interaction process as ideas occur to the analyst. The interactivity and features of MicrobiVis helps the analyst in identifying patterns that may be interesting to analyse further, and hence helps in prioritizing workload. Its use alongside other microbiomics visualization and analysis tools, plus standard clinical statistics approaches, allows a visual exploration of the data. Furthermore, the analyst has been able to gain insights about the generalizability of results.

5.2 User Feedback

As described, MicrobiVis is a further development of a system built for generic data exploration tasks and has been developed and designed in close collaboration with microbiologists working for our industrial partner. The design process has used an iterative approach, driven by regular feedback from the intended end users. Given the relatively small number of domain specialists in the team, no formal usability evaluation of the system has been performed, as their involvement in the design limits objective evaluation with them as participants. An evaluation with another team of microbiologists who were not involved in the design process is possible, and may be the subject of future work.

Feedback from the current end users as to the useful output generated from the exploratory tasks undertaken by MicrobiVis has been positive and is an important indication of the potential of the system. A benefit of MicrobiVis, as pointed out by the end users, is that through the design, with focus on the structure of samples, OTUs and genera found in the microbiomics field, they can easily think in a biological context during exploration and do not have to adapt to a less familiar model. A comparison was made with principal components analysis and network visualization, where the interpretation is not necessarily as intuitive. Users appreciated the possibility of seeing a full OTU profile across samples in a single view, even if this view is initially cluttered, with a range of metrics for filtering it according to different tasks. The RaQ view provides a quick, automatic route for generating an initial interesting view, after which the users quickly develop their own lines of exploration and can work with individual metrics to adjust the view to their particular interest. Through analysis using MicrobiVis, they are able to step their exploration view instantly and interactively between OTU, genus and phylum; this seems to be less straightforward to facilitate in other tools, and, as described in the use case, not all OTUs, and by inference, species, of the same genus are necessarily involved in the same processes. Another feature appreciated was the possibility of spotting subjects with atypical profiles and other outliers, as well as the immediate visual linking of OTU to their abundance and prevalence levels, providing confidence in terms of generalizability. The microbiologists have emphasized that creating hypotheses as to which OTUs are involved in certain processes is normally not an easy task, but using MicrobiVis initial ideas can be generated with ease via initial data exploration, compared against output of other microbiomics tools, and followed up by further designed experiments and statistical analysis.

6 Conclusions and Future Work

This paper introduces MicrobiVis, a system for visual exploration of microbial population data. The system has been developed in collaboration with microbiologist end users and contributes through its design an interactive environment facilitating analysis and acting as a means for developing new and domain relevant insights. MicrobiVis combines information visualization methods with features commonly used in microbiology. Thus provides flexible and explorative dimensionality reduction guided by algorithmic methods, utilizing features and classifications in a biologically relevant context. The usefulness of the system has been demonstrated through a use case describing how a microbiologist could use MicrobiVis to gain new insights and understanding of a microbial data set. Feedback from microbiologists indicates that some of the main contributions of MicrobiVis are the ability to use rank as a filter which can immediately show something interesting. It allows the analyst to think in the biological context without having to adapt to a different model and that the interactivity provides quick identification of anomalies and interesting structures. Some of these would be difficult to identify using other approaches. This greatly helps prioritizing subsequent analysis. Interactive features considered specifically useful includes the ability to explore at various taxonomic levels, to pick out and examine all OTUs of a specific genera for identification of which are involved in certain processes, to be able to consider subjects with atypical profiles quickly when anomalies are found, all of which are normally not easily done. Future work includes further implementation of additional features as suggested by users, incorporation of improved statistical theory based on work exploring appropriate distributions for OTU data, scalability testing and
Acknowledgements

Our thanks to colleagues at Unilever Discover Port Sunlight for useful discussions, feedback and data preparation, particularly S. Grimshaw, P. Helme and A. Smith. This research is partly funded by Unilever Discover Port Sunlight, by the Visualization Programme coordinated by the Swedish Knowledge Foundation and by the Swedish Research Council in the Linnaeus Centre CADICS.

References


