August 2012

Optimal clustering techniques for metagenomic sequencing data

Erik T. Cameron
The University of Western Ontario

Supervisor
L. M. Wahl
The University of Western Ontario

Graduate Program in Applied Mathematics

A thesis submitted in partial fulfillment of the requirements for the degree in Master of Science

OPTIMAL CLUSTERING TECHNIQUES FOR METAGENOMIC SEQUENCING DATA

(Thesis format: Integrated Article)

by

Erik Cameron

Graduate Program in Applied Mathematics

A thesis submitted in partial fulfillment
of the requirements for the degree of
Masters of Science

The School of Graduate and Postdoctoral Studies
The University of Western Ontario
London, Ontario, Canada

THE UNIVERSITY OF WESTERN ONTARIO
School of Graduate and Postdoctoral Studies

CERTIFICATE OF EXAMINATION

Examiners:

Supervisor:

.....
Dr. G. Wild

.....
Dr. L. M. Wahl

.....
Dr. X. Zou

Supervisory Committee:

.....
Dr. G. Gloor

The thesis by

Erik T Cameron

entitled:

Optimal clustering techniques for metagenomic sequencing data

is accepted in partial fulfillment of the
requirements for the degree of
Masters of Science

.....
Date

.....
Chair of the Thesis Examination Board

Abstract

Metagenomic sequencing techniques have made it possible to determine the composition of bacterial microbiota of the human body. Clustering algorithms have been used to search for core microbiota types in the vagina, but results have been inconsistent, possibly due to methodological differences. We performed an extensive comparison of six commonly-used clustering algorithms and four distance metrics, using clinical data from 777 vaginal samples across 5 studies, and 36,000 synthetic datasets based on these clinical data. We found that centroid-based clustering algorithms (K-means and Partitioning around Medoids), with Euclidean or Manhattan distance metrics, performed well. They were best at correctly clustering and determining the number of clusters in synthetic datasets and were also top performers for predicting vaginal pH and bacterial vaginosis by clustering clinical data. Hierarchical clustering algorithms, particularly neighbour joining and average linkage, performed less well, failing unequivocally on many datasets.

Keywords: vaginal microbiota, bacterial vaginosis, metagenomics, cluster analysis, distance metric

Co-Authorship

The work presented in Chapter 2 has been submitted for publication with the title *Optimal clustering techniques for metagenomic sequencing data are predictive of clinical measures in the vaginal microbiota* and is coauthored by E. T. Cameron and L. M. Wahl. The original draft for the article was prepared by the author and revisions were done by the author and L. M. Wahl. The analytical work using Matlab was performed by the author under the supervision of L. M. Wahl.

Acknowledgements

I would like thank my supervisor, Dr. Lindi Wahl, for her guidance in completing this Thesis. You are the perfect combination of scientist, teacher, and friend and I feel lucky to have worked with you. Your knowledge of a broad range of topics and disciplines has been invaluable and I could not have completed this work without you. Thank you.

I would also like to thank Dr. Gregor Reid and Dr. Gregory Gloor as well as their research groups, especially Jean Macklaim for providing access to the clinical data, and Andrew Fernandes for advice regarding the clustering algorithms and distance metrics.

Finally, I would like to thank Dr. Geoff Wild whose instruction and advice with the use of LaTeX and BibTeX have been extremely helpful in the completion of this project.

Contents

List of Figures

List of Tables

List of Appendices

Chapter 1

Introduction

1.1 Human Microbiota

The human microbiome is an ecosystem of microbes that colonize the human body. These microbes can be found throughout the organs of the body such as in the intestine and vagina, and on the surface of the skin. They play significant roles in our metabolism, helping to digest food and synthesize vitamins (Guarner and Malagelada, 2003), or protect the body from infection (Gupta et al., 1998). The living elements of a microbiome are known as microbiota. When referring to microbiota within this text we will specifically be referring to the bacterial elements thereof.

Previous techniques for sequencing human microbiota have relied on culturing bacteria before sequencing genetic information (Hugenholtz, 2002). This meant that bacteria which did not survive in cultures were missed and results were biased towards bacteria that thrived in cultures. More recent metagenomic techniques amplify genetic material directly from samples, without the need for an intermediate culturing step, resulting in better representation of the microbiota associated with a sample (Hugenholtz, 2002).

High-throughput sequencing methods, such as Illumina and 454 sequencing, are common metagenomic methods (Pareek et al., 2011; Hall, 2007). They use a set of genes referred to as 16S rDNA as a sequencing target (Case et al., 2007). This set of genes is universal and highly conserved among bacteria and codes for the 16S ribosomal RNA subunit, which forms part

of the structure of bacterial ribosomes (Woese, 1987). 16S rDNA in the sample is amplified and sequenced. Bacteria can be identified by matching their unique 16S rDNA sequences to a reference database of bacterial genome sequences (Hugenholtz, 2002).

Terminal Restriction Fragment Length Polymorphism (T-RFLP) is a technique which can be used to identify the bacteria in a sample by measuring the size of certain fragments of their 16S rDNA (Liu et al., 1997). 16S rDNA in a sample is amplified and tagged with a fluorescent dye. *Restriction enzymes* are added to the solution. These are enzymes which cut DNA at a specific sequence known as a *restriction site*. The chosen restriction sites are highly conserved among most bacteria, but they occur at varying distances along the 16S gene. The result is that each sequence is cut into a fragment whose size is characteristic of its parent bacteria (Liu et al., 1997). The taxonomic identity of the bacteria can then be determined by comparing the lengths of these fragments to a database.

The lengths of the fragments from a sample are determined by electrophoresis on agarose gel. Occasionally, different bacteria produce similar length fragments when cut at a particular restriction site. These bacteria can be distinguished by using multiple restriction enzymes marked with different dyes in separate runs, and comparing the results using software (Liu et al., 1997).

The results of both T-RFLP and high-throughput sequencing methods are absolute counts for each operational taxonomic unit (OTU) identified in the sample. An OTU contains a group of sequences which have been identified to a certain taxonomic level such as species or genus. The level of taxonomic identification can differ throughout a dataset. The counts are normalized to give an abundance profile, indicating the relative abundance of each OTU. In practice, many sequences in a given sample will have no match in the reference database, resulting in a certain proportion of unknown bacteria in each abundance profile (Hall, 2007). High-throughput sequencing and T-RFLP have been used in the literature to produce abundance profiles of microbiota in parts of the human body such as the gut (Arumugam et al., 2011), vagina (Ravel et al., 2010; Hummelen et al., 2010; Martinez et al., 2009, 2008; Zhou et al.,

2007), stomach (Bik et al., 2006), and mouth (Aas et al., 2005).

Vaginal microbiota are of particular interest to researchers because of the role they play in women's health. Healthy vaginal microbiota maintain an acidic pH level of around 4.5, which can help prevent urinary tract infections (Gupta et al., 1998), as well as the transmission of human immunodeficiency virus (HIV) (Lai et al., 2009). The maintenance of this acidity is generally attributed to the presence of lactic acid-producing bacteria in these biota (Boskey et al., 2001).

Bacterial vaginosis (BV) is a common condition affecting about 30% of women worldwide (Martinez et al., 2009). The condition causes unpleasant discharge and odour as well as increased susceptibility to sexually transmitted infection (Fredricks et al., 2005). The relationship between BV and vaginal microbiota has been studied by Ravel et al. (2010) who used high-throughput 454 sequencing on 16S rRNA to profile the vaginal biota of 396 white, black, Hispanic and Asian women living in North America. Five major community types were identified. Communities high in *Lactobacillus* bacteria were associated with healthy biota while those dominated by other taxa, including *Gardnerella* and *Atopobium* were associated with BV. Similar results were reported by Hummelen et al. (2010), who used high-throughput Illumina sequencing to profile the vaginal biota of 132 HIV positive women in Tanzania. These authors detected eight community types and identified *Lactobacillus iners* and *Lactobacillus crispatus* as being associated with healthy biota while communities containing *Gardnerella vaginalis* were associated with BV.

Racial differences in vaginal microbiota have been studied by Zhou et al. (2007) who profiled the composition of vaginal biota in 144 North American Caucasian and black women using T-RFLP. The authors identified 8 kinds of vaginal communities and found large differences in the community compositions between the two racial groups, with *Lactobacillus* dominated communities being rarer in black women. This difference in vaginal communities was offered as a potential explanation for the increased susceptibility of black women to BV, which is consistent with findings that communities dominated by lactobacilli are more resistant

to BV (Ravel et al., 2010; Hummelen et al., 2010).

A similar study by Zhou et al. (2010) examined the abundance profiles of 73 Japanese women using T-RFLP. Seven community types were identified, all of which were similar to those found in black and white women in the previous study. Japanese women were more likely than black women to have biota dominated by lactobacilli and were more resistant to BV, supporting the vaginal community explanation for racial difference in BV susceptibility. The researchers cited genetic differences in immune function which affect the composition of the microbiota, as noted by Dethlefsen et al. (1987). However, for both articles (Zhou et al., 2007, 2010) each racial group studied was represented for each vaginal community type. This suggests that while race and genetics play primary roles in determining the biota of an individual, the same community types are shared across several geographic regions. The similarity of these data supports the validity of combining and comparing data between studies, which could potentially offer new insights.

1.2 Clustering

Clustering algorithms are a set of tools that are commonly used to analyze microbiota. They aggregate abundance profiles into groups with similar bacterial compositions. An abundance profile can be visualized as a point on a simplex with dimensionality equal to the total number of unique OTUs identified. Profiles with similar bacterial compositions are close to each other in the space of this simplex.

A wide variety of clustering algorithms exist for handling a large range of data types. Clustering of abundance profiles has to date mostly involved the use of hierarchical clustering algorithms (Ravel et al., 2010; Zhou et al., 2010; Hummelen et al., 2010), although centroid-based methods have been used as well (Arumugam et al., 2011). Hierarchical methods are probably the most familiar to researchers in the field because of their frequent application in the construction of phylogenies. Each data point is treated as a separate cluster, and the two closest clusters

are joined. This is repeated until the desired number of clusters is produced. Centroid-based methods place a number of centroids in a dataset and assign each data point to the cluster associated with the closest centroid. The positions of the centroids are chosen to minimize some objective function, such as the sum of the squared distances from each data point to its centroid.

Studies of the vaginal microbiome have used hierarchical algorithms rather than centroid-based techniques. Algorithms used in the literature include Ward's method (Zhou et al., 2010), UPGMA (unweighted pair group method with arithmetic mean, also called average linkage clustering) (Zhou et al., 2007), complete linkage clustering (Ravel et al., 2010), and neighbour joining (Hummelen et al., 2010). A study of gut microbiota (Arumugam et al., 2011) used Partitioning around Medoids (PAM), a centroid-based algorithm.

UPGMA clustering (Sokal and Michener, 1958) determines the closest clusters by measuring the distance between every pairwise combination of points in the two clusters, and averaging. The two clusters with the smallest average distance are combined. Complete linkage clustering (Sorensen, 1948) instead measures the distance between two clusters by choosing one data point from each cluster such that the distance between the two points is maximized. Neighbour joining (Saitou and Nei, 1987) calculates a value between two clusters, i and j as,

$$D(i, j) = (n - 2)d(i, j) - \sum_{k=1}^{n} d(i, k) - \sum_{k=1}^{n} d(j, k), \tag{1.1}$$

where $d(i, j)$ is the distance between the two clusters if they are single points, or is defined below in Equation (1.2) for clusters formed by combination at earlier steps, and n is the current number of clusters. This value is computed for each pair of clusters to produce the matrix D. Clusters a and b are then combined if $D(a, b)$ is the minimal non-diagonal entry of D. The first term on the right side of Equation (1.1) causes clusters farther from each other to be less likely to be combined. The next two terms cause clusters distant from the majority of the dataset to be more likely to be combined. Each time a new cluster, c is created by joining two clusters a and b, the distance from c to a given cluster k is defined,

$$d(c,k) = \frac{d(a,k) + d(b,k) - d(a,b)}{2}. \tag{1.2}$$

This distance is calculated for all remaining clusters $k \neq c$ after each combination step so that the matrix D can be calculated according to Equation (1.1) in the next step.

Ward's method (Ward, 1963) considers each possible 'next step' of combined clusters. For each, it calculates the sum of the squared distances (SSD) from the each data point, j, to the center (mean) of its cluster, m_j. For a dataset with N abundance profiles this is expressed as,

$$SSD = \sum_{j=1}^{N} d(j, m_j)^2. \tag{1.3}$$

At each step, Ward's method combines the two clusters which will reduce the SSD of the dataset by the greatest amount.

Centroid-based methods place a set of K centroids in the space of a dataset to produce an aggregation of K clusters. They assign each data point to the cluster associated with the closest centroid, and calculate the SSD from each data point, j, to its cluster centroid c_j, using the same calculation as in Equation (1.3) while replacing m_j by c_j. The solution when using the K-means (Lloyd, 1982) and PAM (Kaufman and Rousseeuw, 1990) clustering algorithms is the placement of centroids which minimizes this SSD. While there is a single optimal solution to these clustering results, calculating it directly is computationally complex (MacQueen, 1967). Instead, heuristic algorithms which randomly place the initial centroids and recursively move them through the space of the dataset are used. Each step moves a single centroid so that the SSD of the system is reduced, and the algorithm ends when the SSD cannot be reduced in a further step (MacQueen, 1967). K-means clustering (Lloyd, 1982) moves the centroids through continuous space while PAM (Kaufman and Rousseeuw, 1990) moves them only into positions occupied by data points.

The heuristic algorithms used to solve K-means and PAM clustering are non-deterministic. If they find a local minimum, a solution which cannot lower its SSD through one centroid

movement step, they will return it as the solution to the clustering problem even if it is not the absolute minimum solution (MacQueen, 1967). To achieve a result close to the absolute minimum, several runs of the clustering algorithm are typically used, each with a different random initial placement of centroids. The run which produces the lowest SSD is taken as the best solution.

When a clustering algorithm is used, a distance metric must be chosen to define the distance between two points in the space of a dataset. The most familiar distance metric is Euclidean distance, which defines the distance between two vectors, v and u, using a generalization of the Pythagorean Theorem,

$$\text{Euclidean distance}\,(v,u) = \sqrt{(v-u)\cdot(v-u)}. \tag{1.4}$$

The Euclidean distance metric has been used for clustering in research on the microbiota of the human gut (Arumugam et al., 2011) and vagina (Zhou et al., 2007). Similar distance metrics include Manhattan distance, which defines the distance between two points as the sum of the difference in position along each axis, similar to a measure of distance for a trip along city streets which form a grid. For vectors v and u in n dimensions it uses the formula

$$\text{Manhattan distance}\,(v,u) = \sum_{i=1}^{n} |v_i - u_i|\,. \tag{1.5}$$

Distances can also be defined by the angle between two vectors. If the angle between v and u is θ, then the cosine distance between the vectors is defined as,

$$\text{Cosine Distance}\,(v,u) = 1 - \cos(\theta). \tag{1.6}$$

This distance metric has been used for clustering in research of the microbiota of the human vagina (Hummelen et al., 2010). Finally we consider a distance metric based on the Pearson

Correlation between the elements of two vectors, Correlation (v, u),

$$\text{Correlation Distance } (v, u) = 1 - \text{Correlation } (v, u), \tag{1.7}$$

which has also been used in research clustering human vaginal microbiota (Zhou et al., 2010).

Cluster optimization is the practice of determining the number of clusters in a real dataset. Clustering algorithms produce a number of clusters determined *a priori* by the investigator. There are several objective techniques that can be used to select a number of clusters to produce from a real dataset. Research on vaginal microbiota has used objective methods such as the Pseudo-F index of Calinski and Harabasz (1974) in some studies (Zhou et al., 2007, 2010) while others have not indicated the methods of optimization used (Ravel et al., 2010; Hummelen et al., 2010). This issue has been studied in detail by Abdo et al. (2006), who recommended three objective techniques for cluster optimization of abundance profile data. Of these three we use the Pseudo-F index of Calinski and Harabasz (1974). This technique requires the user to create a large range of numbers of clusters, and produces a score for each aggregation which indicates how well the spatial variance is explained with as few clusters as possible (Calinski and Harabasz, 1974). The user keeps the aggregation with the highest score.

High-dimensional data can cause clustering algorithms to fail to produce meaningful results. This 'curse of dimensionality' occurs when data points are distributed over a space with a very large number of dimensions. The distance between any two points in a space will increase with the number of dimensions and the distance between two clusters which differ in only a few dimensions becomes relatively small (Steinbach et al., 2003). This obfuscates clusters in the data and many clustering algorithms will not properly detect them (Aggarwal et al., 1999). Solutions include using feature selection to remove noisy dimensions from the data before clustering, or employing clustering algorithms that project clusters into the relevant dimensions (Aggarwal et al., 1999).

Principal Component Analysis (PCA) is often used with high dimensional data to reduce

dimensionality and group correlated variables, and is often applied to data before clustering (Ding and He, 2004). It is useful for removing noise from data and has been shown to improve the performance of the K-means clustering algorithm on some datasets, helping to find solutions that are closer to optimal (Ding and He, 2004). Pre-treatment of data with PCA has been used in one study of human gut microbiota which was then clustered using PAM (Arumugam et al., 2011).

In summary, to cluster a set of data, a clustering technique and distance metric must first be chosen. The number of clusters can then be determined using an optimization technique. The data may or may not be pre-treated with techniques such as PCA.

1.3 Our Contribution

As the tools of microbiome analysis improve and a growing amount of research examines the composition of the human vaginal microbiome, it is clear that using effective data analysis techniques is of increasing importance. Many studies of the human microbiome use cluster analysis to group similar abundance profiles, including several studies of vaginal microbiota (Zhou et al., 2010, 2007; Ravel et al., 2010; Hummelen et al., 2010) and a landmark study investigating microbiota of the human gut (Arumugam et al., 2011). Clustering is used to group subjects with similar abundance profile composition. In the gut microbiota study by Arumugam et al. (2011) the investigators found three clusters which they called enterotypes.

Similar efforts with vaginal microbiota have yielded as few as five (Ravel et al., 2010) or as many as 12 (Zhou et al., 2007) clusters, as shown on Table 1.1. This research has relied on a variety of methodologies, including differences in clustering algorithms, distance metrics, and cluster optimization. Our goal is to recommend a single, consistent technique for the treatment of these data.

Table 1.1: Summary of previous studies clustering vaginal microbiota.

Study	# Profiles	Sequencing Technique	Cluster Technique, Distance Metric	Cluster Optimization	# Clusters
Zhou et al. 2007	144	T-RFLP	UPGMA, Euclidean	Calinski-Harabasz	12
Zhou et al. 2010	73	T-RFLP	Ward's Method, Correlation	Calinski-Harabasz	9
Ravel et al. 2010	396	High-through (454)	Complete Linkage, Euclidean	Not Declared	5
Hummelen et al. 2010	132	High-through (Illumina)	Neighbour joining, Cosine	Not Declared	8

Our primary objective is to find a data analysis technique that groups patients into biologically relevant clusters. For example, several studies of vaginal microbiota have had an emphasis on subjects with BV (Martinez et al., 2008; Hummelen et al., 2010; Ravel et al., 2010). A tool that consistently clusters subjects into groups which are predictive of BV status would be useful. Similarly, the clusters we find should be able to predict vaginal pH, which is an indicator of vaginal health (Gupta et al., 1998).

We prefer to recommend techniques that are easy to execute and interpret. For this reason we focus on clustering techniques that are widely available in software packages such as R and Matlab, and can be carried out quickly on large datasets using personal computers. We focus on sharp clustering techniques, which assign patients definitively to clusters on a one-to-one basis, because such classifications are convenient to work with for both mathematicians and biologists.

We use data collected in five clinical studies of vaginal microbiota in Tanzania, Brazil and Canada, totalling 777 abundance profiles. These data included women with a variety of health conditions such as HIV (Hummelen et al., 2010), BV or Vulvovaginal candidiasis (VVC) (Martinez et al., 2009), as well as post-menopausal women (Hummelen et al., 2011), pregnant women (unpublished data) and women suffering from toxin shock (unpublished data).

The abundance profile data that we studied had 260 unique OTUs (dimensions) but only a few of these OTUs composed a large proportion of any abundance profiles, hence the data were only widely distributed over the space of a few dimensions. Other papers in the literature report similar distributions in abundance profile data (Ravel et al., 2010; Martinez et al., 2008). Preliminary trials with these data indicated that basic clustering techniques used for low dimensional data, such as K-means, could produce meaningful clustering results. For this reason our investigation did not focus on clustering techniques designed to treat high-dimensional data, which are most useful where basic clustering algorithms fail and when data are widely distributed over many dimensions (Aggarwal et al., 1999). Figure 1.1 shows the distribution of the data we used for the six most common OTUs.

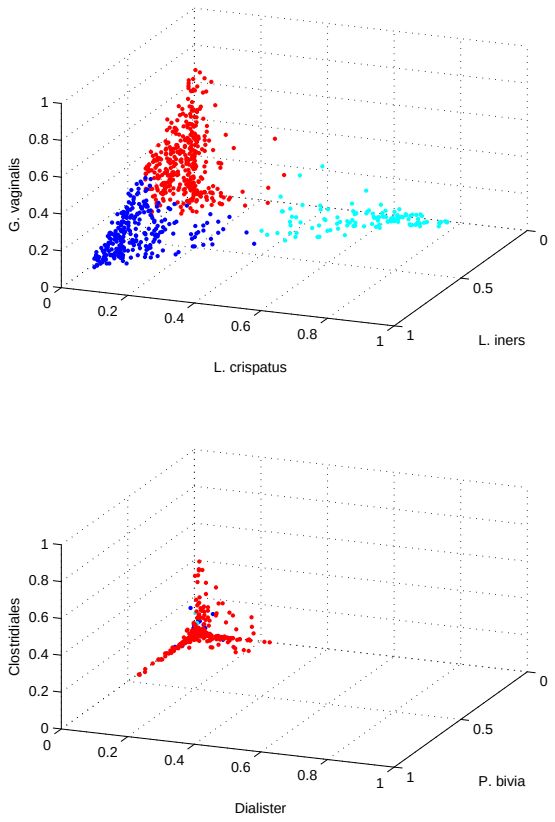

Figure 1.1: Proportion of biota composed by the six most common OTUs for 777 abundance profiles. Note that the first three OTUs are dominant while other OTUs compose a much smaller proportion of the biota. Data points have been separated into three differently colored clusters dominated by the three most common OTUs for visualization using the K-means technique.

PCA has been used in the literature to treat abundance profile data before clustering (Arumugam et al., 2011) and has been shown to improve the performance of K-means in some cases (Ding and He, 2004). Preliminary tests on our clinical data showed that for the pooled set of 777 abundance profiles, the clustering result for K-means with 10 replicates was identical with or without the application of PCA to the data. We chose not pre-treat our data with PCA before clustering because it did not have an impact on results, and to reduce the complexity of our data analyses.

Ultimately we tested six clustering algorithms and four distance metrics. We tested the UPGMA, Ward's method, neighbour joining and complete linkage hierarchical clustering algorithms, and the K-means and PAM centroid-based clustering methods. We used the Euclidean, Manhattan, cosine and correlation distance metrics. Clustering requires a choice of algorithm and distance metric, giving us a total of 24 algorithm-metric combinations which we refer to as *processes*.

We generated 36,000 synthetic abundance profiles based on our clinical data. Frequency distributions for OTUs in the real data tended to be single peaked within clusters but some were double peaked over the entire dataset (examples can be seen in Figures 1.2 to 1.4). We chose to use the beta distribution to generate these data. The beta distribution has two positive parameters, α and β, and generates a value from 0 to 1. It can generate single peaked distributions which form a 'hump' similar to a Gaussian, or are monotonically decreasing or increasing with a peak at 0 or 1 respectively, similar to an exponential distribution. It can also generate double peaked distributions which tend towards values of 0 and 1. Figure 1.5 shows example probability density functions (PDFs) for the beta distribution.

We found that the single peaked beta distributions emulated common OTUs well, as was the case in Figure 1.3. They emulated rare OTUs like that in Figure 1.2 well when they took on an exponential-like shape with a maximum at 0. For OTUs which had frequency distributions with two peaks, a double peaked beta distribution was effective as shown in Figure 1.4.

We fit beta distributions to the proportion values for each OTU over the 777 abundance

profiles. We fit these distributions to each OTU over the dataset as a whole to generate noise. We also partitioned our dataset into three clusters using K-means and fit beta distributions for each OTU within the separate clusters. These three sets of distributions were used to generate clustered data. Figures 1.2 to 1.4 show the frequency distributions for some OTUs in real data, and compares them to the distributions for synthetic OTUs which were based on them. For additional details on the production of synthetic data see the Methods section in the following chapter.

We tested which processes were best at determining the true number of clusters, and correctly clustering these synthetic data. For our clinical data, we tested which processes were most predictive of a subject's BV status and vaginal pH level.

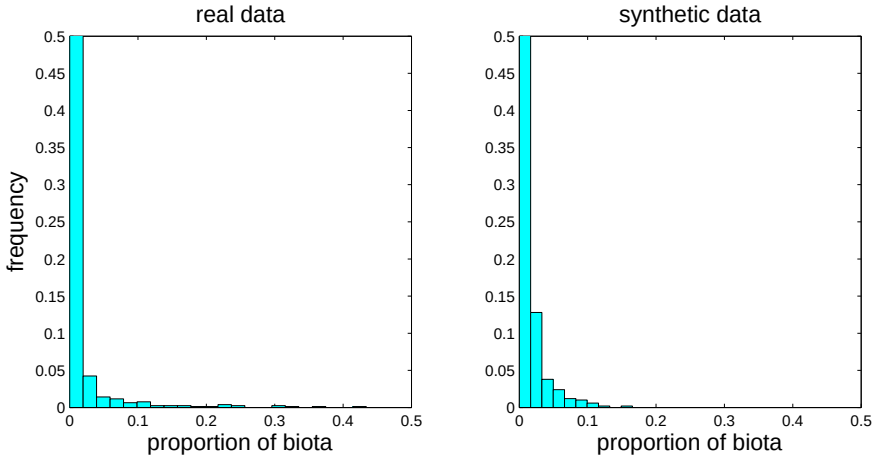

Figure 1.2: Proportion of biota composed by a rare OTU over all profiles

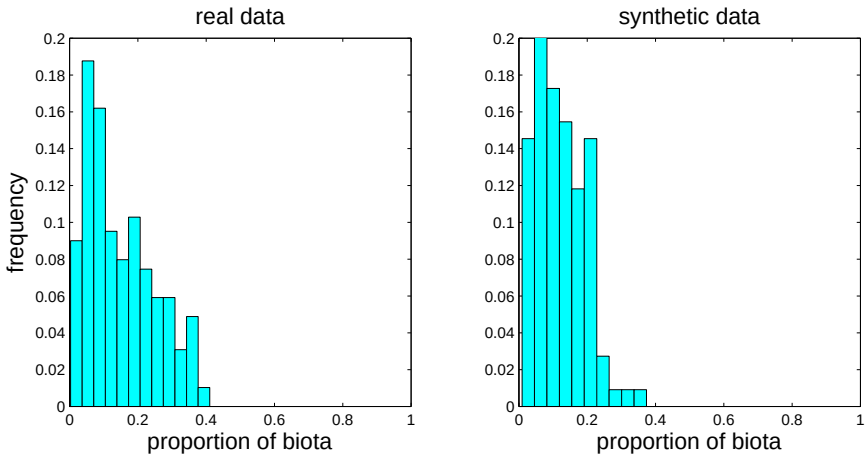

Figure 1.3: Proportion of biota composed by L. iners for one cluster

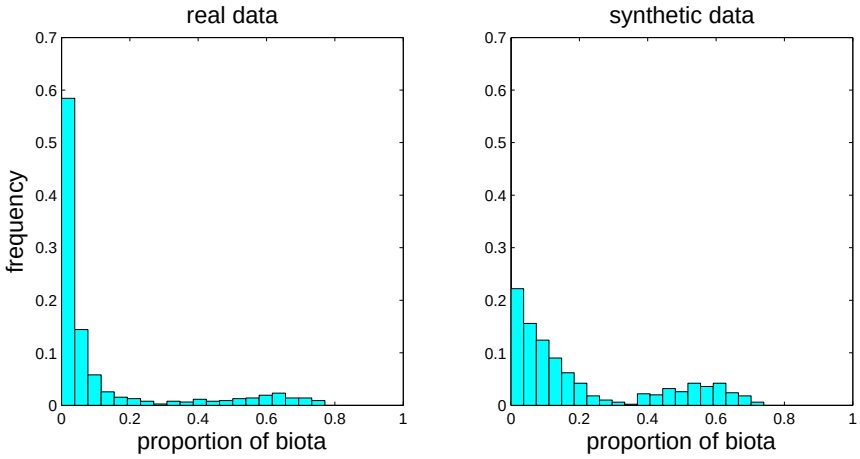

Figure 1.4: Proportion of biota composed by L. crispatus over all profiles

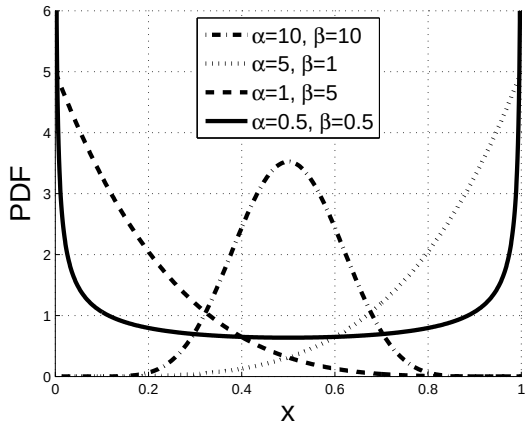

Figure 1.5: Beta probability density functions for several values of input parameters α and β.

Bibliography

Aas, J. A., Paster, B. J., Stokes, L. N., Olsen, I., and Dewhirst, F. E. (2005). Defining the normal bacterial flora of the oral cavity. *J. Clin. Microbiol.*, 43(11):5721–5732.

Abdo, Z., Schuette, U., Bent, S., Williams, C., Forney, L., and Joyce, P. (2006). Statistical methods for characterizing diversity of microbial communities by analysis of terminal restriction fragment length polymorphisms of 16S rRNA genes. *Environ Microbiol.*, 8(5):929–938.

Aggarwal, C. C., Wolf, J. L., Yu, P. S., Procopiuc, C., and Park, J. S. (1999). Fast algorithms for projected clustering. *SIGMOD Rec.*, 28(2):61–72.

Arumugam, M., Raes, J., Pelletier, E., Paslier, D. L., Yamada, T., Mende, D., Fernandes, G., Tap, J., Bruls, T., Batto, J.-M., Bertalan, M., Borruel, N., Casellas, F., Fernandez, L., Gautier, L., Hansen, T., Hattori, M., Hayashi, T., Kleerebezem, M., Kurokawa, K., Leclerc, M., Levenez, F., Manichanh, C., Nielsen, H., Nielsen, T., Pons, N., Poulain, J., Qin, J., Sicheritz-Ponten, T., Tims, S., Torrents, D., Ugarte, E., Zoetendal, E., Wang, J., Guarner, F., Pederson, O., de Vos, W., Brunak, S., Dore, J., Weissenbach, J., Ehrlich, S., and Bork, P. (2011). Enterotypes of the human gut microbiome. *Nature*, 473:174–180.

Bik, E. M., Eckburg, P. B., Gill, S. R., Nelson, K. E., Purdom, E. A., Francois, F., Perez-Perez, G., Blaser, M. J., and Relman, D. A. (2006). Molecular analysis of the bacterial microbiota in the human stomach. *Proc Nat Acad Sci USA*, 103(3):732–737.

Boskey, E., Cone, R., Whaley, K., and Moench, T. (2001). Origins of vaginal acidity: high D/L lactate ratio is consistent with bacteria being the primary source. *Hum. Reprod.*, 16(9):1809–1813.

Calinski, T. and Harabasz, J. (1974). A dendrite method for cluster analysis. *Commun Stat*, 3(1):1–27.

Case, R. J., Boucher, Y., Dahllf, I., Holmstrm, C., Doolittle, W. F., and Kjelleberg, S. (2007). Use of 16S rRNA and rpoB genes as molecular markers for microbial ecology studies. *Appl. Environ. Microbiol.*, 73(1):278–288.

Dethlefsen, L., McFall-Ngai, M., and Relman, D. A. (1987). An ecological and evolutionary perspective on humanmicrobe mutualism and disease. *Nature*, 449:811–818.

Ding, C. and He, X. (2004). K-means clustering via principle component analysis. In *Proc. of Int'l Conf. Machine Learning*, pages 225–232.

Fredricks, D. N., Fiedler, T. L., and Marrazzo, J. M. (2005). Molecular identification of bacteria associated with bacterial vaginosis. *N. Engl. J. Med.*, 353(18):1899–1911.

Guarner, F. and Malagelada, J.-R. (2003). Gut flora in health and disease. *Lancet*, 361:512–519.

Gupta, K., Stapleton, A., Hooton, T., Roberts, P., Fennell, C., and Stamm, W. (1998). Inverse association of h2o2-producing lactobacilli and vaginal Escherichia coli colonization in women with recurrent urinary tract infections. *J Infect Dis*, 178(2):446–450.

Hall, N. (2007). Advanced sequencing technologies and their wider impact in microbiology. *J. Exp. Biol.*, 209:1518–1525.

Hugenholtz, P. (2002). Exploring prokaryotic diversity in the genomic era. *Genome Biol.*, 3(2):reviews0003.1–reviews0003.8.

Hummelen, R., Fernandes, A. D., Macklaim, J. M., Dickson, R. J., Changalucha, J., Gloor, G. B., and Reid, G. (2010). Deep sequencing of the vaginal microbiota of women with HIV. *PLoS ONE*, 5(8):e12078.

Hummelen, R., Macklaim, J. M., Bisanz, J. E., Hammond, J.-A., McMillan, A., Vongsa, R., Koenig, D., Gloor, G. B., and Reid, G. (2011). Vaginal microbiome and epithelial gene array in post-menopausal women with moderate to severe dryness. *PLoS ONE*, 6(11):e26602.

Kaufman, L. and Rousseeuw, P. (1990). *Finding groups in data: an introduction to cluster analysis*. Wiley series in probability and mathematical statistics: Applied probability and statistics. Wiley.

Lai, S. K., Hida, K., Shukair, S., Wang, Y.-Y., Figueiredo, A., Cone, R., Hope, T. J., and Hanesl, J. (2009). Human immunodeciency virus type 1 is trapped by acidic but not by neutralized human cervicovaginal mucus. *J. Virol.*, 83(21):11196–11200.

Liu, W., Marsh, T., Cheng, H., and Forney, L. (1997). Characterization of microbial diversity by determining terminal restriction fragment length polymorphisms of genes encoding 16S rRNA. *Appl. Environ. Microbiol.*, 63(11):4516–4522.

Lloyd, S. P. (1982). Least squares quantization in PCM. *IEEE Trans Inf Theory*, 28(2):129–137.

MacQueen, J. B. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability. University of California Press*, pages 281–297.

Martinez, R. C. R., Franceschini, S. A., Patta, M. C., Quintana, S. M., Gomes, B. C., Martinis, E. C. P. D., and Reid., G. (2009). Improved cure of bacterial vaginosis with single dose of tinidazole (2 g), Lactobacillus rhamnosus gr-1, and Lactobacillus reuteri rc-14: a randomized, double-blind, placebo-controlled trial. *Can J Microbiol*, 55(2):133–138.

Martinez, R. C. R., Franceschini, S. A., Patta, M. C., Quintana, S. M., Nunes, A. C., Moreira, J. L. S., Anukam, K. C., Reid, G., , and Martinis, E. C. P. D. (2008). Analysis of vaginal lactobacilli from healthy and infected Brazilian women. *Appl. Environ. Microbiol.*, 74(14):4539–4542.

Pareek, C. S., Smoczynski, R., and Tretyn, A. (2011). Sequencing technologies and genome sequencing. *J Appl Genetics*, 52:413–435.

Ravel, J., Gajer, P., Abdo, Z., Schneider, G. M., Koenig, S. S. K., McCulle, S. L., Karlebach, S., Gorle, R., Russell, J., Tacket, C. O., Brotman, R. M., Davis, C. C., Ault, K., Peralta, L., and Forney, L. J. (2010). Vaginal microbiome of reproductive-age women. *Proc Nat Acad Sci USA*, 108:4680–4687.

Saitou, N. and Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.*, 4(4):406–425.

Sokal, R. R. and Michener, C. D. (1958). A statistical method for evaluating systematic relationships. *University of Kansas Science Bulletin*, 38:1409–1438.

Sorensen, T. (1948). A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish Commons. *Biologiske Skrifter*, 5:1–34.

Steinbach, M., Ertz, L., and Kumar, V. (2003). The challenges of clustering high-dimensional data. In *In New Vistas in Statistical Physics: Applications in Econophysics, Bioinformatics, and Pattern Recognition*. Springer-Verlag.

Ward, J. H. (1963). Hierarchical grouping to optimize an objective function. *JASA*, 58(301):236–244.

Woese, C. R. (1987). Bacterial evolution. *Microbiol. Mol. Biol. Rev.*, 51(2):221–271.

Zhou, X., Brown, C., Abdo, Z., Davis, C., Hansmann, M., Joyce, P., Foster, J., and Forney, L. (2007). Differences in the composition of vaginal microbial communities found in healthy Caucasian and black women. *ISME J.*, 1:121–133.

Zhou, X., Hansmann, M., Davis, C., Suzuki, H., Brown, C., Schutte, U., Pierson, J., and Forney, L. (2010). The vaginal bacterial communities of Japanese women resemble those of women in other racial groups. *FEMS Immunol Med Microbiol.*, 58:169–181.

Chapter 2

Optimal clustering techniques for vaginal microbiota

2.1 Introduction

In the last decade, interest in the bacterial populations (microbiota) of the human body has been growing. Advances in metagenomic sequencing techniques have made possible the collection of the rich datasets needed to characterize the composition of these populations. For example, recent efforts have been made to characterize microbiota within the human gut (Arumugam et al., 2011); similar efforts have also been directed toward the characterization of microbiota of the stomach (Bik et al., 2006) and oral cavity (Aas et al., 2005). Vaginal microbiota have been a focus of particular recent interest (Ravel et al., 2010; Hummelen et al., 2010; Martinez et al., 2008; Zhou et al., 2007, 2010).

2.1.1 Composition of Microbiota

Modern metagenomic sequencing techniques amplify bacterial genetic elements directly from a sample (Hugenholtz, 2002), such as a faecal sample or vaginal swab. Two common metagenomic techniques that identify the bacterial compositions of biota are high-throughput sequencing, and Terminal Restriction Fragment Length Polymorphism (T-RFLP). High-throughput techniques copy and sequence a highly conserved, universal gene. These are then identified

by comparison to a database of previously sequenced bacterial genomes (Hugenholtz, 2002). In T-RFLP, DNA is cut at common, highly conserved sites and the lengths of the resulting fragments are measured. They are then identified by comparison to a database of previously cut bacterial genomes (Liu et al., 1997). For both techniques, the result is an absolute count of the number of reads for each operational taxonomic unit (OTU) in the sample. The counts are normalized giving an abundance profile indicating the proportion of each OTU in the sample.

An abundance profile can be represented as a point on a simplex whose dimensionality is equal to the total number of unique OTUs detected over all samples. To reduce dimensionality and improve understanding, a clustering algorithm is typically applied to these data. Clustering is used to group points that are close to each other on the simplex so as to identify samples composed of similar bacterial OTUs. The objective is to categorize subjects in ways that are interesting or useful, for example, to identify if a set of core types (clusters) dominate the biota, and how they relate to the health of the subject (Ravel et al., 2010). Similar attempts have been made with gut microbiota, for which one study found 3 core types (Arumugam et al., 2011). Results in clustering vaginal microbiota have been inconsistent, identifying as few as 5 (Ravel et al., 2010), or as many as 12 (Zhou et al., 2007) clusters in the data. However, since an established method for clustering and analysis of abundance profile data has yet to emerge, methodological differences could underlie these inconsistencies.

2.1.2 Clustering

Two broad categories of clustering algorithms have been applied to abundance profile data to date: hierarchical and centroid-based. Hierarchical methods treat each data point as an individual cluster and recursively combine the closest ones until the desired number of clusters is reached. Centroid-based methods place one centroid in the data space for each desired cluster. Each data point is then assigned to the cluster corresponding to the closest centroid. The centroids are recursively moved to minimize some objective function, such as the sum of squared distances between each data point and its cluster centroid. These steps repeat until the

objective function cannot be lowered by another step.

Because finding the optimal solution for a centroid-based method is computationally expensive, a heuristic algorithm is used, making the outcome non-deterministic. The outcome can depend on the initial, random placements of centroids in the dataset, and a single run of the algorithm might find a local minimum of the objective function, rather than the global minimum. To minimize this effect it is common practice to use multiple runs of the algorithm and take the result which minimizes the objective function.

To date, studies of the vaginal microbiome have used hierarchical algorithms such as Ward's method (Zhou et al., 2010), UPGMA (unweighted pair group method with arithmetic mean, also called average linkage clustering) (Zhou et al., 2007), complete linkage clustering (Ravel et al., 2010), and neighbour joining trees (Hummelen et al., 2010), while the gut microbiota study (Arumugam et al., 2011) used Partitioning around Medoids (PAM), a centroid-based algorithm.

In addition to these algorithmic differences, the distance metric, the measure used by the algorithm to define the distance between two data points, also varies widely. Euclidean distance is the familiar metric used in everyday measurement of distance. It has been used to cluster gut microbiota (Arumugam et al., 2011). Angular distance measures the angle between two points from the origin. Correlation distance takes each point in its vector form and measures the correlation between the elements of the two vectors. Two points on a simplex that are not equal will always have a non-zero angular and correlation distance between them. Studies of vaginal microbiota have used Euclidean distance (Ravel et al., 2010; Zhou et al., 2007), angular distance (Hummelen et al., 2010) and correlation distance (Zhou et al., 2010).

A final methodological issue in clustering abundance profile data is to determine the optimal number of clusters. Most clustering algorithms take as input a dataset and an integer number of desired clusters and give as output a partitioning of that dataset into the same number of clusters. The number of clusters must be chosen *a priori* even though the true number of clusters in the dataset is likely unknown or not defined. The general solution to this problem is

to repeat the clustering algorithm for a large range of numbers of clusters and choose the best result. The best result can be chosen subjectively, or through a variety of objective methods which optimize a function, usually by finding a partitioning that best explains the spatial variance of the data with as few clusters as possible. Abdo et al. (2006) have already addressed this issue in some detail, suggesting three algorithms that can be used to determine the optimal number of clusters in abundance profile data.

2.1.3 Our Contribution

The vaginal microbiome plays an important role in women's health. Bacterial vaginosis (BV) is a common condition affecting about 30% of women worldwide and is strongly linked to compositional changes in the subject's vaginal microbiota (Martinez et al., 2009). Microbiota also play a role in the resistance to yeast infections, HIV (Human Immunodeficiency Virus) and other sexually transmitted infections (Ravel et al., 2010). To better understand the rich datasets currently becoming available, establishing a consistent, well-studied methodology for the clustering and analysis of microbiota data is clearly necessary.

Here, we test 24 combinations of clustering algorithms and distance metrics to find which provide the most meaningful analyses of abundance profile datasets. We use data collected in five clinical studies of vaginal microbiota in Tanzania, Brazil and Canada, totalling 777 abundance profiles, each characterized by 260 OTUs. We generate 36,000 synthetic data sets with known clusters based on these clinical data. We use these synthetic data to test which combinations are best at correctly determining the number of clusters in a dataset, and best at assigning abundance profiles to the correct clusters. We then determine which combinations are best at partitioning the clinical data in ways that are predictive of clinical measures.

2.2 Methods

2.2.1 Clustering

We tested two centroid-based clustering algorithms and four hierarchical clustering algorithms. The hierarchical algorithms we tested were UPGMA, neighbour joining, Ward's method and complete linkage clustering. Complete linkage and UPGMA both measure the distance between two clusters to determine which two are closest. For clusters containing multiple points, UPGMA uses the average pairwise distance between all combinations of points in the two clusters (Sokal and Michener, 1958), while complete linkage uses the greatest distance between any two points in the clusters (Sorensen, 1948). Neighbour joining uses a special distance calculation, which incorporates the distance between the two clusters, but also applies a term which favours linking clusters far from the center of the dataset (Saitou and Nei, 1987). Finally, Ward's method calculates the sum of the squared distance from each point to the mean position of the cluster it belongs to and connects the two clusters which minimize the sum of squared distances in the next step (Ward, 1963).

The centroid-based algorithms we tested the K-means and PAM. Both algorithms attempt to minimize the sum of squared distances from the data points to the centroids. K-means recursively moves the centroids through continuous space (Lloyd, 1982), while for PAM only the positions of data points qualify as potential positions for centroids (Kaufman and Rousseeuw, 1990). For both algorithms the centroid movement steps repeat until the sum of squared distances cannot be lowered in the next step.

To mitigate the issue of centroid-based methods finding local minima, we used 10 runs with K-means and 15 with PAM and selected the result with the lowest sum of squared distances. These numbers of runs were selected to consistently produce good clusterings (based on initial observations using up to 50 runs) while minimizing computational load.

We also tested four distance metrics: Euclidean distance, Manhattan distance (also called city block distance), cosine distance and correlation distance. Euclidean and Manhattan dis-

tance are simple distance metrics that use the following formulas to give the distance between two points in n-dimensional space with vector positions v and u,

$$\text{Euclidean distance } (v, u) = \sqrt{(v-u)\cdot(v-u)}, \tag{2.1}$$

$$\text{Manhattan distance } (v, u) = \sum_{i=1}^{n} |u_i - v_i| . \tag{2.2}$$

The cosine distance between two points grows as the angle between them increases, measured from the origin. It is given as

$$\text{Cosine Distance } (v, u) = 1 - \cos(\theta), \tag{2.3}$$

where θ is the angle between v and u. Finally, the correlation distance between two points grows as the correlation between the elements of those points shrinks. It is measured,

$$\text{Correlation Distance } (v, u) = 1 - \text{Correlation } (v, u). \tag{2.4}$$

Each clustering algorithm and distance metric can be combined into a metric/algorithm combination which we will refer to as a *process*. We tested four distance metrics and six algorithms, giving a total of 24 processes.

2.2.2 Synthetic Data

To produce synthetic data, we first grouped subjects in the clinical data into 2 to 12 clusters using K-means clustering with Euclidean distance, and used the technique of Calinski and Harabasz (1974) to determine the optimal number of clusters. This yielded 3 clusters which we used as a basis for generating our synthetic data. Each of the three clusters emphasized a single dominant OTU which composed between 20% and 80% of the microbiota, while other OTUs composed less than 40% of the biota and were often much rarer. Within each cluster, we used maximum likelihood estimates (Hahn and Shapiro, 1994) to fit beta distributions to

the abundance values of all subjects for each OTU. This gave three sets of *Cluster Parameters*, each consisting of 260 beta distributions corresponding to the 260 OTUs. A fourth set of beta distributions was fitted to each OTU in the entire unclustered dataset. This set was called the *Noise Parameters*.

A set of synthetic data contained K clusters and N subjects. To generate a synthetic cluster, we selected one of the three sets of cluster parameters (randomly, with replacement) on which the cluster would be based. Each subject in a synthetic cluster had an abundance profile based on their cluster's respective cluster parameters. The *n*th OTU was always the most common for the *n*th synthetic cluster. We accomplished this by switching the *n*th beta distribution with the beta distribution having the highest mean in the chosen set of cluster parameters. For example, to generate the fourth cluster in a synthetic dataset we would randomly select one set of cluster parameters. The beta distribution for the most common OTU in that set of cluster parameters would be switched with the beta distribution for the fourth OTU, so that the fourth synthetic cluster would have high amounts of OTU four. This was done so that any number of unique synthetic clusters could be drawn from three sets of cluster parameters. To produce a synthetic abundance profile within a cluster, the value for each OTU was drawn from the corresponding beta distribution in the appropriate set of cluster parameters. This gave each of the 260 OTUs a value between 0 and 1. The profile was then normalized to sum to 1. We assigned $\frac{2N}{3}$ subjects to clusters, yielding $\frac{2N}{3K}$ per cluster. Finally, $\frac{N}{3}$ of the subjects were assigned to the *Noise Group*. Abundance profiles for the Noise Group were drawn from the noise parameters. Synthetic datasets were produced with $N = 20, 50, 200$ or 500 subjects, and with $K = 2$ to 9 clusters. We produced 500 replicates for each combination.

For a side-by-side comparison of abundance for several OTUs in real and synthetic data, see section A.2.7.

We used the synthetic datasets to test which processes were best at determining the true number of clusters in a dataset. Each process was used to produce aggregations of 1 to 15 clusters on every synthetic dataset. The optimal number of clusters was determined using the

Pseudo-F method proposed by Calinski and Harabasz (1974), as recommended by Abdo et al. (2006). A trial was considered successful if the optimal number of clusters was equal to the true number of clusters, or the true number of clusters plus one (allowing for the identification of noise data as a separate cluster).

We also used the synthetic datasets to test which processes were best at assigning subjects to the correct clusters. To do this, we used each process to create an aggregation for the true number of clusters. We then determined the conditional entropy of the true clusters given the found clusters, $H(T|F)$ using Shannon Entropy (Shannon, 1948). In order to produce a more intuitive measure of entropy, and allow comparison of results, we converted this value into the proportion of entropy explained by the clustering result, *PE*, which we defined as

$$PE = \frac{H(T) - H(T|F)}{H(T)}, \tag{2.5}$$

where $H(T)$ is the entropy of the true clusters alone. A higher value for PE indicates a better clustering result. A result of $PE = 1$ indicates that every point was assigned to the correct cluster, and a result of $PE \approx 0$ indicates a highly random clustering containing little or no information. Note that in this calculation, the synthetic data points in the Noise Group were omitted, so as not to reward or punish an algorithm for how it classified the noise data.

We repeated the above procedures using a set of *Hard Cluster Parameters* designed to produce clusters with more overlap to determine which processes were best for less easily clustered datasets (see section A.1.1). We also produced an alternate set of cluster parameters based on clusters found with complete linkage clustering rather than K-means. Using these parameters we generated new synthetic datasets and tested each clustering algorithm with Euclidean distance. This was used as a control to ensure that the algorithm used to generate the cluster parameters would not bias the results. In total we generated and tested 36,000 synthetic datasets.

We compared the PE results from the clustering algorithms we used to a control PE result

obtained by clustering the same data randomly. The random clustering algorithm assigned data points independently to each of the N clusters with probability $\frac{1}{N}$.

2.2.3 Clinical Data

We used data from five clinical studies. This included data from women with HIV from Tanzania (Hummelen et al., 2010) and women with or without BV and with or without yeast infections from Brazil (Martinez et al., 2009), as well as post-menopausal women (Hummelen et al., 2011), pregnant women (unpublished data) and women suffering from toxin shock (unpublished data) in Canada. Abundance profile data as well as a clinical diagnosis for BV was available for all women in these studies, and a measure of vaginal pH was available in some studies. Vaginal pH is relevant to women's health and a higher pH is associated with BV (Zhou et al., 2007). We evaluated the conditional entropy of the pH (increments of 0.5 from 3.5 to 8.5) given the clusters found. We also evaluated the conditional entropy of each subject's BV status (normal, intermediate, and BV) given the clusters found. Sufficient data concerning pH was recorded for 344 of the abundance profiles and sufficient data concerning BV was recorded for 668 of the abundance profiles.

We tested each process for $K = 2$ to 9 clusters. We determined the conditional entropy of the vaginal pH and BV status given the found clusters, $H(BV|F)$ and $H(pH|F)$. BV status was determined using the Nugent Criteria (Nugent et al., 1990). Similar to our treatment of entropy with synthetic data, we converted this value into the proportion of entropy explained by the clustering result, which is defined as

$$PE_{BV} = \frac{(H(BV) - H(BV|F))}{H(BV)} \text{ for BV data and,} \tag{2.6}$$

$$PE_{pH} = \frac{(H(pH) - H(pH|F))}{H(pH)} \text{ for pH data,} \tag{2.7}$$

where $H(pH)$ and $H(BV)$ are the entropies of the BV and pH labels respectively. Again, the values for PE_{BV} and PE_{pH} are between 0 and 1, where a higher value indicates a better

explanation of the clinical criteria through clustering. We tested the combined dataset from all 5 studies. We also tested the five studies with sufficient data on BV status and the two studies with sufficient data on pH status individually. We used bootstrapping to find a 95% confidence interval for our PE values (see section A.1.2).

2.2.4 Software

All of the algorithms we tested are available in version 7.12.0 of Matlab (The MathWorks, Inc.). To perform K-means and neighbour joining clustering, we used `kmeans.m` and `seqneighjoin.m` respectively. To perform UPGMA, complete linkage and Ward's method we used `linkage.m`. To perform PAM clustering we used `kmedoids.m` available on the Matlab file exchange (`http://www.mathworks.com/matlabcentral/fileexchange/28860-kmedioids/content/kmedioids.m`).
Optimization of clustered results was carried out using the Cluster Validity Analysis Platform available on the Matlab file exchange (`http://www.mathworks.com/matlabcentral/fileexchange/14620`).

2.3 Results

2.3.1 Synthetic Data

Figure 2.1 shows how well each process assigned synthetic data points to the correct cluster, indicating the PE for each process. Euclidean distance worked best with UPGMA and PAM, and Manhattan distance was best with K-means, Ward's method and neighbour joining. No single distance metric was clearly best for complete linkage clustering, though correlation distance did well. The results did not conflict for smaller sample sizes (Figures A.4 to A.6). All processes performed better than random clusters used as a control, which yielded a PE of 0.05 or less for synthetic datasets with 500 abundance profiles and 2 to 9 clusters (results not shown).

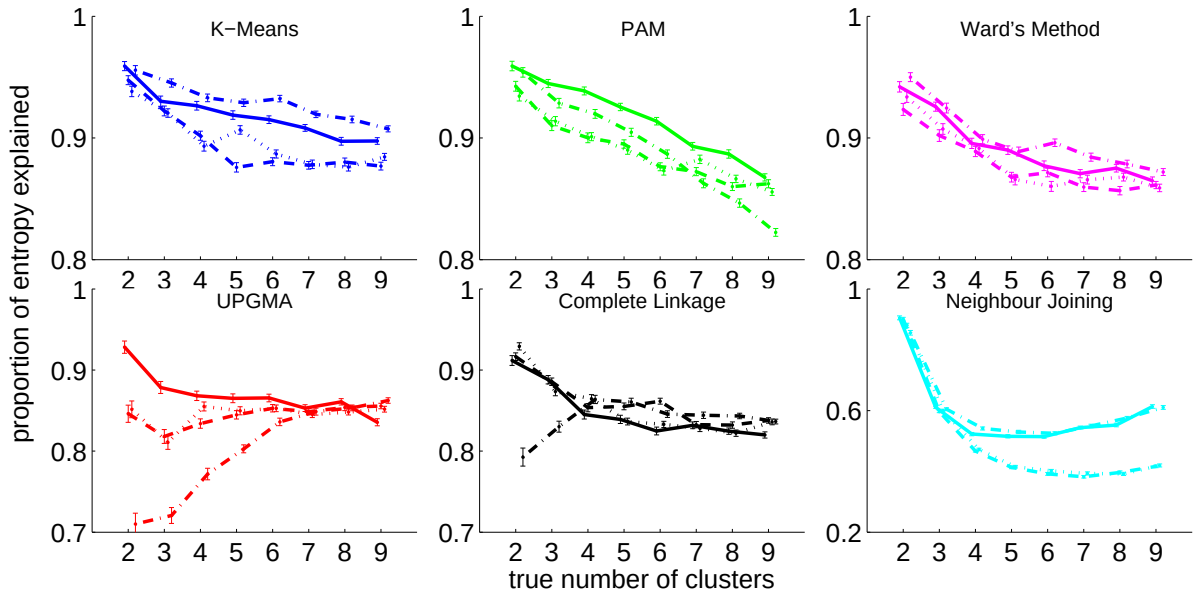

Figure 2.1: Performance of each process on synthetic datasets of 500 abundance profiles with 2 to 9 true clusters. The PE is plotted against the number of synthetic clusters in the dataset. Lines are solid for Euclidean distance, dotted for cosine distance, dashed for correlation distance, and dot-dashed for Manhattan distance. Lines are blue for K-means, green for PAM, purple for Ward's method, red for UPGMA, black for complete linkage, and cyan for neighbour joining. Error bars show one standard error of the mean and are staggered on the x-axis for visibility.

Figure 2.2 compares each clustering algorithm using Euclidean distance. K-means and PAM outperformed the other clustering algorithms with few exceptions for between 2 and 9 clusters. Neither of K-means and PAM consistently outperformed the other. Ward's, UPGMA and complete linkage clustering performed moderately well. Neighbour joining performed very poorly.

When synthetic datasets with closer, less contrasted clusters were drawn from our hard cluster parameters, the results did not contradict the above findings. The same distance metrics were optimal for each clustering algorithm (Figures A.7 to A.10) and K-means with Manhattan distances was the best performing process; it was similar to PAM and Ward's for few clusters and superior with 6 or more clusters (Figure A.11). There was also almost no difference in the results for synthetic data based on cluster parameters found using complete linkage clustering instead of K-means (Figure A.12).

K-means and PAM were also best at determining the number of true clusters Figure 2.3.

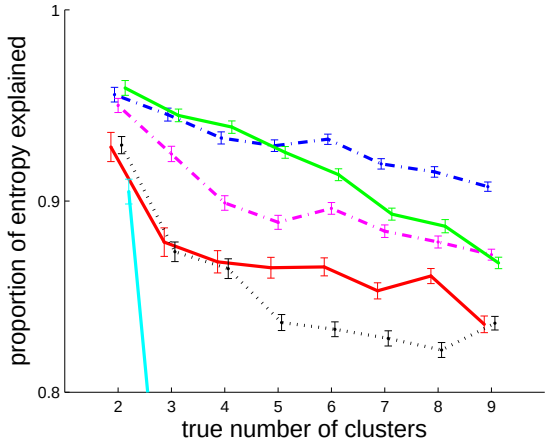

Figure 2.2: Performance of each clustering algorithm using its best distance metric on synthetic datasets of 500 compositional profiles with 2 to 9 true clusters. The *PE* is plotted against the number of synthetic clusters in the data set. PE Values for neighbour joining at 3 to 9 clusters are below 0.6. Lines are solid for Euclidean distance, dotted for cosine distance, dashed for correlation distance, and dot-dashed for Manhattan distance. Lines are blue for K-means, green for PAM, purple for Ward's method, red for UPGMA, black for complete linkage, and cyan for neighbour joining.

These results were consistent for between 3 and 7 true clusters. UPGMA, Ward's method and complete linkage performed similarly over any number of clusters; neighbour joining performed the most poorly. We compare Euclidean distances in Figure 2.3 below because it performs well and consistently between clustering algorithms. Results for all distance metrics can be seen in Figure A.13. No single distance metric outperformed all others for any algorithm, although cosine distance was consistently a top performer.

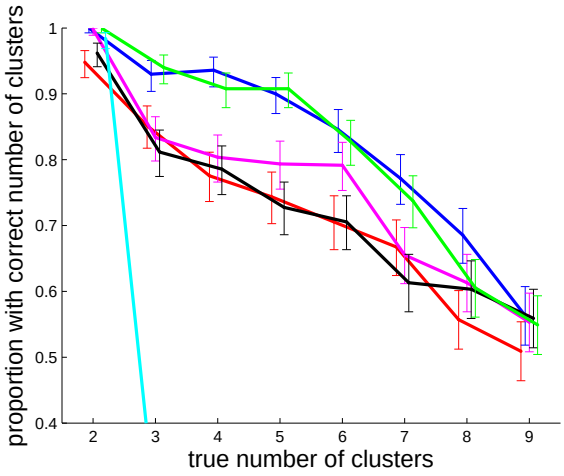

Figure 2.3: Performance of each clustering algorithm with Euclidean distance on synthetic datasets of 500 abundance profiles with 2 to 9 true clusters. Each clustering algorithm gave from 2 to 12 clusters for each dataset. The optimal number of clusters was found determined using the technique of Calinski and Harabasz (1974). The proportion of trials identifying the correct number of clusters is plotted against the number of synthetic clusters in the dataset. PE Values for neighbour joining at 3 to 9 clusters are below 0.2. Lines are blue for K-means, green for PAM, purple for Ward's method, red for UPGMA, black for complete linkage, and cyan for neighbour joining.

2.3.2 Clinical Data

Figure 2.4 shows that K-means, PAM, Ward's method and complete linkage performed similarly for explaining the BV status entropy in the pooled dataset of 668 abundance profiles from 5 studies for which information on subjects' BV statuses were available. Neighbour joining was inferior to these methods, and UPGMA clustering was inferior to neighbour joining. Neighbour joining also performed poorly on the Brazil BV and Canadian toxin shock datasets alone. For simplicity we have shown only the results for Euclidean distance as it was consistently one of the top performing distance metrics. The results for all distance metrics are given on Figures A.14 to A.19, although they did not differ by much. Over these datasets UPGMA and neighbour joining performed very poorly at least once for each distance metric.

Figure 2.5 shows that UPGMA performed poorly for explaining the pH value entropy in the pooled dataset of 344 abundance profiles from 2 studies for which information on subject's pH values was available. PAM, K-means, Ward's method and complete linkage performed

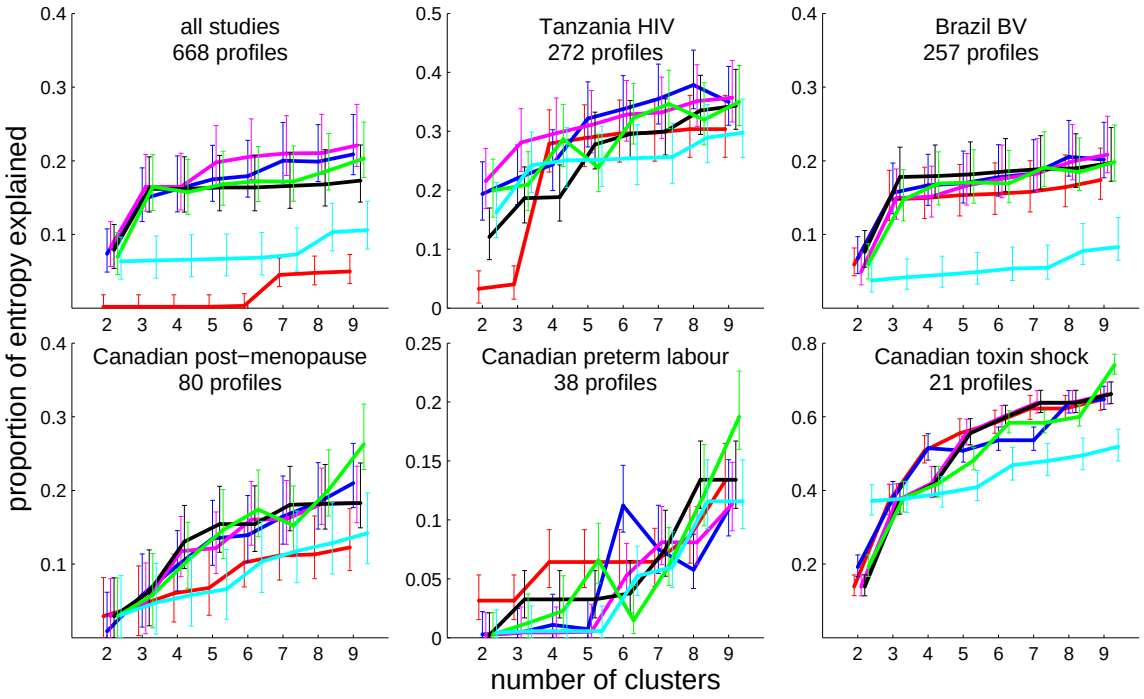

Figure 2.4: Performance of each clustering algorithm with Euclidean distance on clinical data with 2 to 9 clusters. The PE of the subject's BV status is plotted against the number of *a priori* clusters the algorithm was asked to find. Error bars show 95% confidence intervals obtained by bootstrapping over 10,000 replicates, and are staggered on the x-axis for visibility. Lines are blue for K-means, green for PAM, purple for Ward's method, red for UPGMA, black for complete linkage, and cyan for neighbour joining.

similarly well to each other. Neighbour joining was consistently but not significantly worse than these four algorithms. When data from the studies were not pooled, there was insufficient power to distinguish the performance of the algorithms. As with the BV data above we have shown only the results for Euclidean distance here, and the results for all distance metrics are given on Figures A.20 to A.22. The results did not differ much by distance metric. Again, UPGMA and neighbour joining performed very poorly at least once for each distance metric.

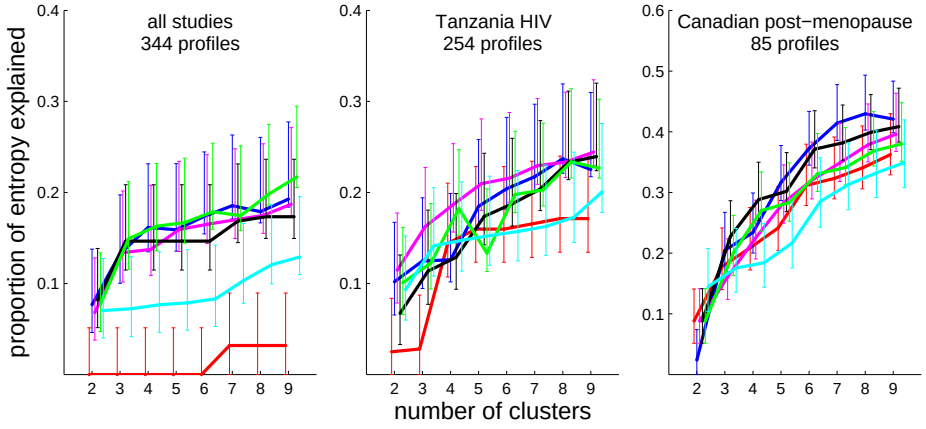

Figure 2.5: Performance of each clustering algorithm with Euclidean distance on clinical data with 2 to 9 clusters. The PE of the pH values is plotted against the number of *a priori* clusters the algorithm was asked to find. Error bars show a 95% confidence interval obtained by bootstrapping over 10,000 replicates, and are staggered on the x-axis for visibility. Lines are blue for K-means, green for PAM, purple for Ward's method, red for UPGMA, black for complete linkage, and cyan for neighbour joining.

2.4 Discussion

K-means with Manhattan distance and PAM with Euclidean distance were the most effective algorithms we tested for clustering synthetic data, as shown in Figure 2.2, and were among the most successful algorithms for clustering clinical data, as shown in Figure 2.4 and Figure 2.5. The spatial distribution of our data provides some insight as to why certain clustering algorithms excelled where others failed. Data were dispersed on a 260-dimensional simplex. Dense clusters of points were found towards the corners of the first three dimensions of the simplex representing the three most abundant OTUs: *L. iners*, *L. crispatus* and *G. vaginalis*. Together these OTUs accounted for 57% of the biota. A large amount of dense noise lay between these three clusters. See Figure A.1 for a visual representation of the composition of the 777 abundance profiles by the six most abundant OTUs.

For hierarchical clustering methods such as UPGMA and neighbour joining, each step consists of combining the two clusters which are closest. In densely packed datasets a cluster becomes closer to other points as it grows, and can continue to grow in a chain (Jain et al., 1999). This chaining can cause the data to be partitioned into one large cluster as well as

several small clusters, which may explain the poor results for these algorithms in synthetic and clinical data. In some cases meaningful clustering can be obtained if small clusters are discarded, but this requires subjective post-hoc input from the researcher and results in the loss of potentially valuable data. The use of UPGMA in the literature has resulted in very small clusters with abundance profile data, some of which were discarded (Zhou et al., 2007). Likewise, the use of neighbour joining in the literature has resulted in the production of many small clusters (Hummelen et al., 2010). K-means and PAM do not suffer from this weakness and are effective at detecting clusters in dense noise (Jain et al., 1999).

Ward's method and complete linkage clustering are hierarchical algorithms that operate similarly to UPGMA and neighbour joining, but use a different function to determine which two clusters are closest. Complete linkage measures the farthest distance between any pair of points in the clusters, meaning that as clusters grow they can never move closer to external points. For this reason it does not produce the chaining results described above (Jain et al., 1999). Ward's method combines the two clusters that will best explain the spatial variance in the data, pursuing the same goal as K-means and PAM. This helps explain why these algorithms did not fail for some data sets but UPGMA and neighbour joining did. Regardless, complete linkage and Ward's method did not perform as well as the centroid-based methods on synthetic data.

There are several known weaknesses of the centroid-based algorithms we studied. K-means and PAM are poorly suited for detecting close or overlapping clusters (Jain et al., 1999). The cluster centroids detected in our clinical data tended towards the corners of the simplex on which the data were distributed, as can be seen in Figure A.1, suggesting non-overlapping clusters.

Another weakness of K-means and PAM is that their results can vary between runs. These algorithms lower an objective function with each iteration, but can become caught in local minima and report suboptimal solutions (Jain et al., 1999). By using 10-15 runs of these algorithms, each run with different random starting positions for the centroids, we were able

to produce clustering results superior to those produced by hierarchical methods. For a given dataset, this took only a few seconds of computational time.

Outliers specifically have been shown to produce clustering results with these local minima, and this problem is more prevalent for K-means than it is for PAM (Hodge and Austin, 2004). Outliers are not an issue for abundance profile data because abundance profiles are bounded and extreme values cannot occur. This fact helps explain why we did not encounter severe problems with local minima with our centroid-based clustering techniques and why PAM did not consistently outperform K-means for synthetic data even though the former is often considered an improvement over the latter (Hodge and Austin, 2004).

A final concern of the K-means and PAM algorithms is their tendency to produce clusters of the same volume (Xiong et al., 2009). The algorithms assign each point to the closest cluster centroid, and data points near the threshold of differently sized clusters can be improperly assigned to the wrong cluster. This is especially problematic for clusters that are close together or overlapping (Xiong et al., 2009). As mentioned above, the distance between clusters tended to be large in our data. Naive observation of the data in Figure A.1 seems to indicate that the majority of the data on the borders between clusters is noise. It is worth noting that clusters found by K-means and PAM in our clinical data differed in population by up to a factor of 3. This shows that these centroid-based algorithms are capable of discerning reasonably different sized clusters in abundance profile data.

Two of the distance metrics tested have properties of special interest. Angular distance metrics are invariant under multiplication by a constant, and correlation distance is invariant under any linear transformation. On a simplex, all elements of a vector sum to one, so such transformations on data points generally cannot occur. These properties may provide an advantage in some spaces, but not on a simplex. Overall, cosine distance (an angular metric) and correlation distance performed poorly and we do not recommend their use in clustering abundance profile data.

The Manhattan distance is greater than the Euclidean distance for small deviations in mul-

tiple dimensions. For example, the distance from $\langle 1, 0, 0\rangle$ to $\langle 0.8, 0.1, 0.1\rangle$ is 0.245 under the Euclidean metric and 0.4 under the Manhattan metric. Such differences are common in abundance profiles, which contain a variety of rare OTUs. The tendency of the Manhattan distance to emphasize these differences may explain why it was superior to Euclidean distance when clustering synthetic data with K-means and Ward's method (see Figure 2.1).

Alternatively, some rare OTUs are the result of sequencing errors (Quince et al 2009) and in the clinical data we tested, the relative standard deviation of rare OTUs was higher than that for the most common OTUs (see Figure A.2). These points suggest that clustering results which rely heavily on these rare sequences are prone to error. In the synthetic clusters we produced, the underlying distributions for these rare strains were different between clusters and identical within clusters. This does not necessarily reflect the structure of the clinical data and it may enable the Manhattan distance with some clustering algorithms to artificially discern between the synthetic clusters we produced.

Abundance profile data generally contains a sizable portion of unidentified sequences. In the clinical data we used it composed on average 16.9% of each abundance profile (standard deviation 0.04%, $N = 777$). It is not clear how this amount should be treated when clustering. In our research we chose to include this unidentified fraction, treating it as a separate OTU (dimension). Figure A.3 shows that the clustering result for a dataset does change depending on whether this unidentified fraction is included. The change was highest for neighbour joining and generally least with Euclidean distance.

In the clinical data we studied, clusters found by the optimal process differed in population by a factor of up to 3. Our synthetic data used clusters of equal sizes. Approximately $\frac{1}{3}$ of the abundance profiles we generated in each dataset were noise. One possible avenue for future study would be to test algorithms using synthetic data with varying cluster sizes and proportions of noise. Specifically, we are interested in which algorithms are best at identifying small true clusters in the presence of noise and large clusters, and which algorithms are most likely to categorize noise as clusters.

We would also like to develop synthetic data which better represents the structure of the clinical data. For example, anecdotal observation of the data suggests a higher density of points along the edges of the simplex. These high density 'lines' between the clusters in the clinical data may be responsible for chaining or other phenomena that can confound UPGMA, which could help explain why it failed for the clinical data but not the synthetic data.

Finally, we are interested in how the results of this study apply to studies in the literature which study human vaginal microbiota by clustering sets of abundance profiles (Ravel et al., 2010; Hummelen et al., 2010; Martinez et al., 2009, 2008; Zhou et al., 2007, 2010). While the individual papers have produced a variety of insightful results, they tend to approach clustering using different algorithms, distance metrics, and optimization techniques. The resulting number of clusters differs for each paper. A unified approach, applying the same method to each dataset, might produce results than can be more easily compared and assist in the search for the core types of human vaginal microbiota.

2.5 Conclusions

We recommend using K-means clustering with Manhattan distance when handling abundance profile data of vaginal microbiota. This was the best tested processes for accurately clustering synthetic data and for identifying the correct number of clusters in synthetic data. It was among the best processes for predicting the BV status and vaginal pH of subjects based on their abundance profile data. K-means and PAM with Euclidean distance performed similarly well, and we can recommend the use of either wherever convenient.

Use multiple runs to avoid local minima. We were successful with as few as 10 runs, but recommend 50 runs as a conservative precaution. Clustering should be carried out for a wide range of numbers of clusters, and the optimal number of clusters should be determined using the method of Calinski and Harabasz (1974) as recommended by Abdo et al. (2006). This can be done easily with many software packages and is computationally fast for large datasets

(around 1000 abundance profiles).

We recommend against using UPGMA and neighbour joining. Both algorithms failed unequivocally on some clinical datasets. Neighbour joining failed on all tests using synthetic data while UPGMA performed only moderately well with synthetic data.

Bibliography

Aas, J. A., Paster, B. J., Stokes, L. N., Olsen, I., and Dewhirst, F. E. (2005). Defining the normal bacterial flora of the oral cavity. *J. Clin. Microbiol.*, 43(11):5721–5732.

Abdo, Z., Schuette, U., Bent, S., Williams, C., Forney, L., and Joyce, P. (2006). Statistical methods for characterizing diversity of microbial communities by analysis of terminal restriction fragment length polymorphisms of 16S rRNA genes. *Environ Microbiol.*, 8(5):929–938.

Arumugam, M., Raes, J., Pelletier, E., Paslier, D. L., Yamada, T., Mende, D., Fernandes, G., Tap, J., Bruls, T., Batto, J.-M., Bertalan, M., Borruel, N., Casellas, F., Fernandez, L., Gautier, L., Hansen, T., Hattori, M., Hayashi, T., Kleerebezem, M., Kurokawa, K., Leclerc, M., Levenez, F., Manichanh, C., Nielsen, H., Nielsen, T., Pons, N., Poulain, J., Qin, J., Sicheritz-Ponten, T., Tims, S., Torrents, D., Ugarte, E., Zoetendal, E., Wang, J., Guarner, F., Pederson, O., de Vos, W., Brunak, S., Dore, J., Weissenbach, J., Ehrlich, S., and Bork, P. (2011). Enterotypes of the human gut microbiome. *Nature*, 473:174–180.

Bik, E. M., Eckburg, P. B., Gill, S. R., Nelson, K. E., Purdom, E. A., Francois, F., Perez-Perez, G., Blaser, M. J., and Relman, D. A. (2006). Molecular analysis of the bacterial microbiota in the human stomach. *Proc Nat Acad Sci USA*, 103(3):732–737.

Calinski, T. and Harabasz, J. (1974). A dendrite method for cluster analysis. *Commun Stat*, 3(1):1–27.

Hahn, G. J. and Shapiro, S. S. (1994). *Statistical Models in Engineering*. Wiley.

Hodge, V. J. and Austin, J. (2004). A survey of outlier detection methodologies. *Artificial Intelligence Review*, 22:2004.

Hugenholtz, P. (2002). Exploring prokaryotic diversity in the genomic era. *Genome Biol.*, 3(2):reviews0003.1–reviews0003.8.

Hummelen, R., Fernandes, A. D., Macklaim, J. M., Dickson, R. J., Changalucha, J., Gloor, G. B., and Reid, G. (2010). Deep sequencing of the vaginal microbiota of women with HIV. *PLoS ONE*, 5(8):e12078.

Hummelen, R., Macklaim, J. M., Bisanz, J. E., Hammond, J.-A., McMillan, A., Vongsa, R., Koenig, D., Gloor, G. B., and Reid, G. (2011). Vaginal microbiome and epithelial gene array in post-menopausal women with moderate to severe dryness. *PLoS ONE*, 6(11):e26602.

Jain, A. K., Murty, M. N., and Flynn, P. J. (1999). Data clustering: A review.

Kaufman, L. and Rousseeuw, P. (1990). *Finding groups in data: an introduction to cluster analysis*. Wiley series in probability and mathematical statistics: Applied probability and statistics. Wiley.

Liu, W., Marsh, T., Cheng, H., and Forney, L. (1997). Characterization of microbial diversity by determining terminal restriction fragment length polymorphisms of genes encoding 16S rRNA. *Appl. Environ. Microbiol.*, 63(11):4516–4522.

Lloyd, S. P. (1982). Least squares quantization in PCM. *IEEE Trans Inf Theory*, 28(2):129–137.

Martinez, R. C. R., Franceschini, S. A., Patta, M. C., Quintana, S. M., Gomes, B. C., Martinis, E. C. P. D., and Reid., G. (2009). Improved cure of bacterial vaginosis with single dose of tinidazole (2 g), Lactobacillus rhamnosus gr-1, and Lactobacillus reuteri rc-14: a randomized, double-blind, placebo-controlled trial. *Can J Microbiol*, 55(2):133–138.

Martinez, R. C. R., Franceschini, S. A., Patta, M. C., Quintana, S. M., Nunes, A. C., Moreira, J. L. S., Anukam, K. C., Reid, G., , and Martinis, E. C. P. D. (2008). Analysis of vaginal lactobacilli from healthy and infected Brazilian women. *Appl. Environ. Microbiol.*, 74(14):4539–4542.

Nugent, R., Krohn, M., and Hillier, S. (1990). Reliability of diagnosing bacterial vaginosis is improved by a standardized method of gram stain interpretation. *J Clin Microbiol*, 29(2):297–301.

Ravel, J., Gajer, P., Abdo, Z., Schneider, G. M., Koenig, S. S. K., McCulle, S. L., Karlebach, S., Gorle, R., Russell, J., Tacket, C. O., Brotman, R. M., Davis, C. C., Ault, K., Peralta, L., and Forney, L. J. (2010). Vaginal microbiome of reproductive-age women. *Proc Nat Acad Sci USA*, 108:4680–4687.

Saitou, N. and Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.*, 4(4):406–425.

Shannon, C. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423.

Sokal, R. R. and Michener, C. D. (1958). A statistical method for evaluating systematic relationships. *University of Kansas Science Bulletin*, 38:1409–1438.

Sorensen, T. (1948). A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish Commons. *Biologiske Skrifter*, 5:1–34.

Ward, J. H. (1963). Hierarchical grouping to optimize an objective function. *JASA*, 58(301):236–244.

Xiong, H., Wu, J., and Chen, J. (2009). K-means clustering versus validation measures: A data-distribution perspective. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, 39(2):318 –331.

Zhou, X., Brown, C., Abdo, Z., Davis, C., Hansmann, M., Joyce, P., Foster, J., and Forney, L. (2007). Differences in the composition of vaginal microbial communities found in healthy Caucasian and black women. *ISME J.*, 1:121–133.

Zhou, X., Hansmann, M., Davis, C., Suzuki, H., Brown, C., Schutte, U., Pierson, J., and Forney, L. (2010). The vaginal bacterial communities of Japanese women resemble those of women in other racial groups. *FEMS Immunol Med Microbiol.*, 58:169–181.

Chapter 3

Summary and Future Work

We have recommended a single data analysis technique for the treatment of abundance profile data for vaginal microbiota. Such data have been gathered in several studies (Zhou et al., 2007, 2010; Hummelen et al., 2010; Ravel et al., 2010) which have used them to explore different characteristics of the biota using a variety of methodologies. An obvious avenue for future research would be to apply our recommended technique to the data generated in these and other already published studies. Consistent data treatment improves the credibility of comparisons made between the findings of studies. Combining the data from multiple, geographically diverse studies could be a key step in establishing the core vaginal microbiota.

We tested our data analysis techniques using an extensive set of real and synthetic datasets. The K-means clustering algorithm was a top performer for artificial datasets composed of $\frac{1}{3}$ noise data, and real datasets wherein clusters differed in size (population) by a ratio of three. Nevertheless, generalizations of our methods could help determine under which circumstances different data analysis techniques are preferable. Testing synthetic datasets with clusters of varying sizes could be useful in determining which techniques are best for detecting small but biologically meaningful clusters.

It would be useful to improve the method used to generate noise in our synthetic data. While synthetic clusters and noise match the approximate distribution of strains in the real data, the high density 'lines' of noise observed on the edge of the simplex in our real data are not well emulated in our synthetic data. This is illustrated in Figure 3.1, which shows three examples of

synthetic data compared to real data. Some hierarchical clustering methods such as UPGMA are sensitive to high density, which may explain why this algorithm failed for real datasets when it performed moderately well on synthetic datasets. Future studies should identify a method of emulating this characteristic of real abundance profile data or be mindful of the fact that synthetic data produced in this way may not robustly test density sensitive clustering methods.

An important next step in improving our ability to analyse the human microbiome would to be expand the research in this thesis to microbiota elsewhere in the human body. Metagenomic techniques have been used to produce abundance profiles of microbiota in the stomach (Bik et al., 2006) and oral cavity (Aas et al., 2005). Analysis of abundance profile datasets for these biota could determine if the same techniques are effective across the human microbiome. Synthetic data based on these real data could be used to test data analysis techniques using methods similar to those in this thesis.

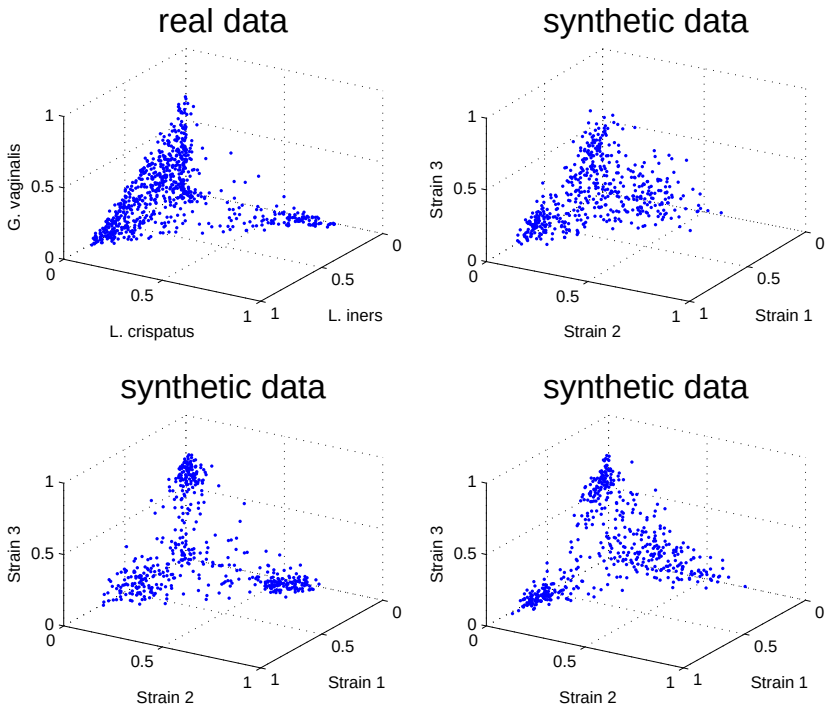

Figure 3.1: Visual comparison of abundance of three most common strains in one real and three synthetic datasets. Major differences in synthetic dataset appearance are due to emulation of different random strains in each instance. Note that the synthetic datasets fail to reproduce the high-density lines along the simplex edges seen in the real data.

Bibliography

Aas, J. A., Paster, B. J., Stokes, L. N., Olsen, I., and Dewhirst, F. E. (2005). Defining the normal bacterial flora of the oral cavity. *J. Clin. Microbiol.*, 43(11):5721–5732.

Bik, E. M., Eckburg, P. B., Gill, S. R., Nelson, K. E., Purdom, E. A., Francois, F., Perez-Perez, G., Blaser, M. J., and Relman, D. A. (2006). Molecular analysis of the bacterial microbiota in the human stomach. *Proc Nat Acad Sci USA*, 103(3):732–737.

Hummelen, R., Fernandes, A. D., Macklaim, J. M., Dickson, R. J., Changalucha, J., Gloor, G. B., and Reid, G. (2010). Deep sequencing of the vaginal microbiota of women with HIV. *PLoS ONE*, 5(8):e12078.

Ravel, J., Gajer, P., Abdo, Z., Schneider, G. M., Koenig, S. S. K., McCulle, S. L., Karlebach, S., Gorle, R., Russell, J., Tacket, C. O., Brotman, R. M., Davis, C. C., Ault, K., Peralta, L., and Forney, L. J. (2010). Vaginal microbiome of reproductive-age women. *Proc Nat Acad Sci USA*, 108:4680–4687.

Zhou, X., Brown, C., Abdo, Z., Davis, C., Hansmann, M., Joyce, P., Foster, J., and Forney, L. (2007). Differences in the composition of vaginal microbial communities found in healthy Caucasian and black women. *ISME J.*, 1:121–133.

Zhou, X., Hansmann, M., Davis, C., Suzuki, H., Brown, C., Schutte, U., Pierson, J., and Forney, L. (2010). The vaginal bacterial communities of Japanese women resemble those of women in other racial groups. *FEMS Immunol Med Microbiol.*, 58:169–181.

Appendix A

Supplementary Information

A.1 Supplementary Methods

A.1.1 Hard Cluster Parameters

We produced a set of *Hard Cluster Parameters* designed to generate clusters with more overlap to determine which processes were best for less easily clustered datasets. This was achieved by reducing the amount of the most common OTU in each set of Cluster Parameters. Each cluster had a single primary OTU associated with it, which was more common than any other OTU in that cluster. The beta distribution for this primary OTU is defined by a pair of beta parameters, α and β and the distribution has a mean value of $\frac{\alpha}{\alpha+\beta}$. We reduced the α parameter of the distribution for the primary OTU to $\frac{2}{3}$ of its original value, thus reducing the mean of the distribution. This reduced the proportion of biota composed by the primary OTU for synthetic data points drawn from the Hard Cluster Parameters, which causes clusters to be closer together and thus harder for clustering algorithms to discern.

A.1.2 Bootstrapping

We used bootstrapping to find a 95% confidence interval for our *PE* values. The result of clustering a clinical data set was a set of BV status or pH value labels each paired to a cluster label. We resampled from this set of paired labels, with replacement. The size of the resampled data set was equal to the number of data points with relevant pH or BV data in the clustered clinical data set. Next, the *PE* was calculated for the resampled data set. The resampling and calculation was repeated 10,000 times to estimate the 95% confidence interval.

A.1.3 Distribution of Clinical Data

Our clinical data contained 777 abundance profiles each describing the proportion of a sample composed by 260 OTUs. Three of these OTUs were highly abundant and together composed over 57% of the biota. Figure A.1 below shows the proportion of the 777 abundance profiles composed by the six most abundant OTUs.

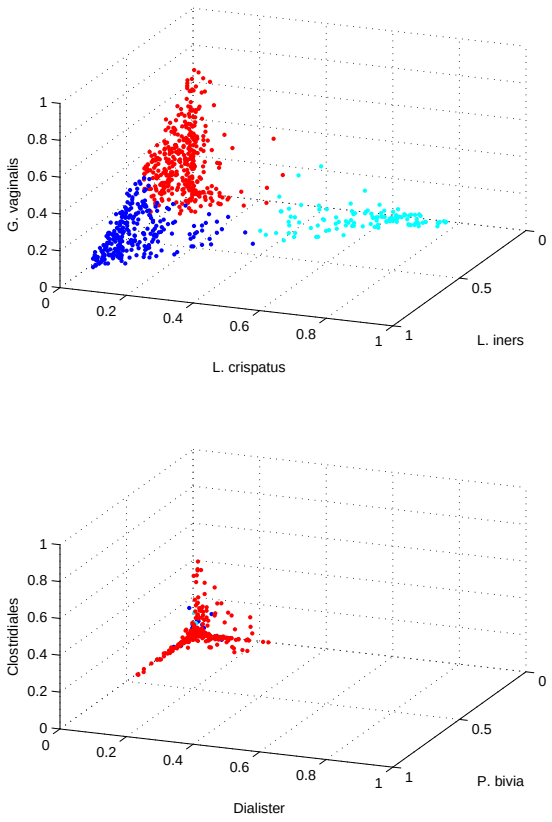

Figure A.1: Proportion of biota composed of the six most common OTUs for 777 abundance profiles. Note that the first three OTUs are dominant while other OTUs compose a much smaller proportion of the biota. Data points have been separated into three differently colored clusters dominated by the three most common OTUs for visualization using the K-means technique.

A.1.4 Relative Standard Deviation for Rare OTUs

We calculated the relative standard deviation (RSD) for each OTU in our clinical data. The RSD is a measure of the variation in the data relative to the mean, and is calculated as the standard deviation divided by the mean. This measure indicates how widely the composition by a given OTU differs between profiles relative to its own size. Figure A.2 below shows that rare OTUs had a higher RSD than the three most common OTUs.

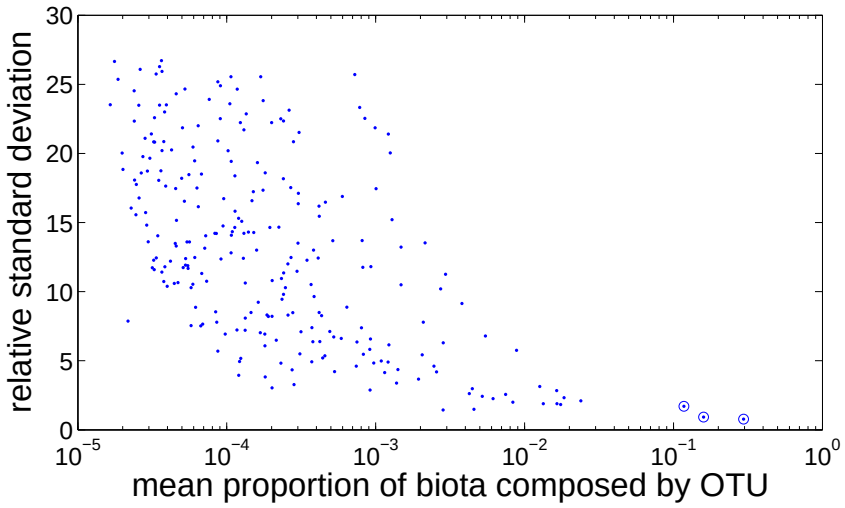

Figure A.2: Relative standard deviation (RSD) of each OTU over the 777 abundance profiles. Note that the RSD for the three most common OTUs, indicated with circles, is below 2. The RSD for many rare strains is much higher, indicating that they vary between abundance profiles by a large amount proportional to their own average magnitude. Not displayed are three outliers of the RSD values for three rare OTUs.

A.1.5 Removal of unidentified sequence reads

Unidentified sequence reads can be removed from the data or treated as their own OTU (dimension) before applying clustering. The result can differ as shown below on Figure A.3.

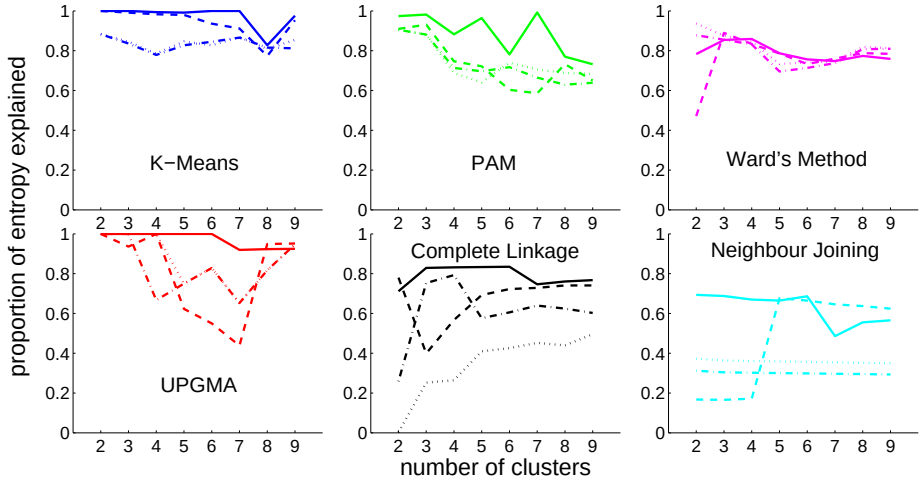

Figure A.3: The proportion of entropy in clustering results with unidentified sequence reads explained by those without. Each process was applied to the pooled clinical data with and without unidentified sequence data and produced from 2 to 9 clusters for each case. A PE of 1 indicates that the clustering results were identical for each case. A lower PE indicates a greater difference in the two clustering results. Lines are solid for Euclidean distance, dotted for cosine distance, dashed for correlation distance, and dot-dashed for Manhattan distance. Lines are blue for K-means, green for PAM, purple for Ward's method, red for UPGMA, black for complete linkage, and cyan for neighbour joining.

A.2 Supplementary Figures

A.2.1 *PE* by distance metric and algorithm for synthetic datasets of 200, 50 and 20 profiles

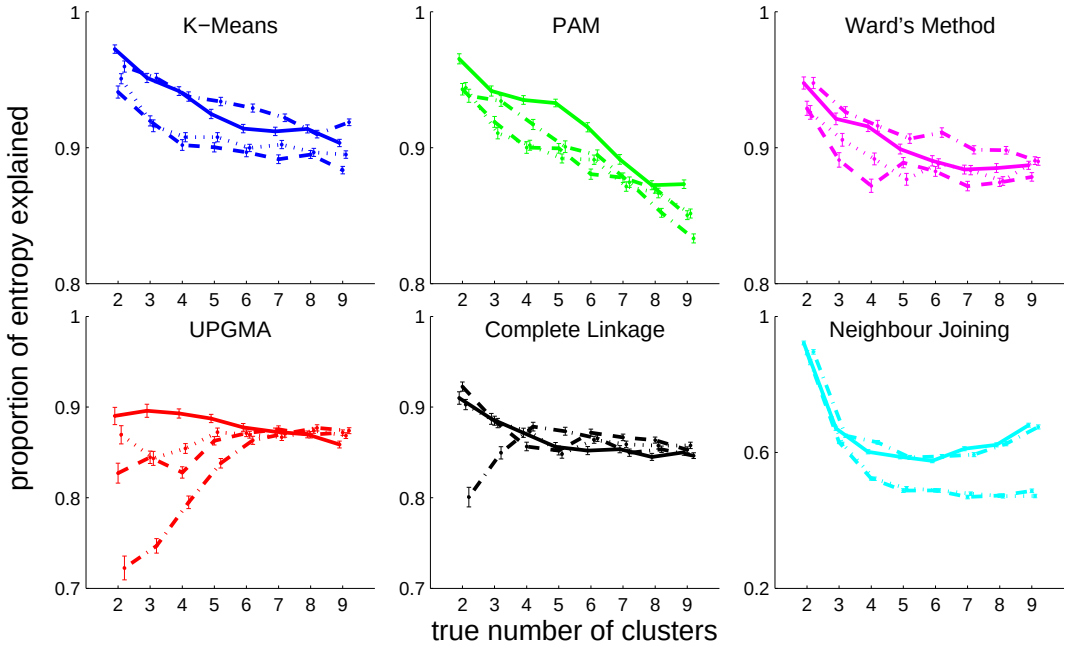

Figure A.4: Performance of each process on synthetic datasets of 200 compositional profiles with 2 to 9 true clusters. The *PE* is plotted against the number of synthetic clusters in the data set. Lines are solid for Euclidean distance, dotted for cosine distance, dashed for correlation distance, and dot-dashed for Manhattan distance. Lines are blue for K-means, green for PAM, purple for Ward's method, red for UPGMA, black for complete linkage, and cyan for neighbour joining. Error bars show one standard error of the mean and are staggered on the x-axis for visibility. Error bars for neighbour joining are indiscernibly small.

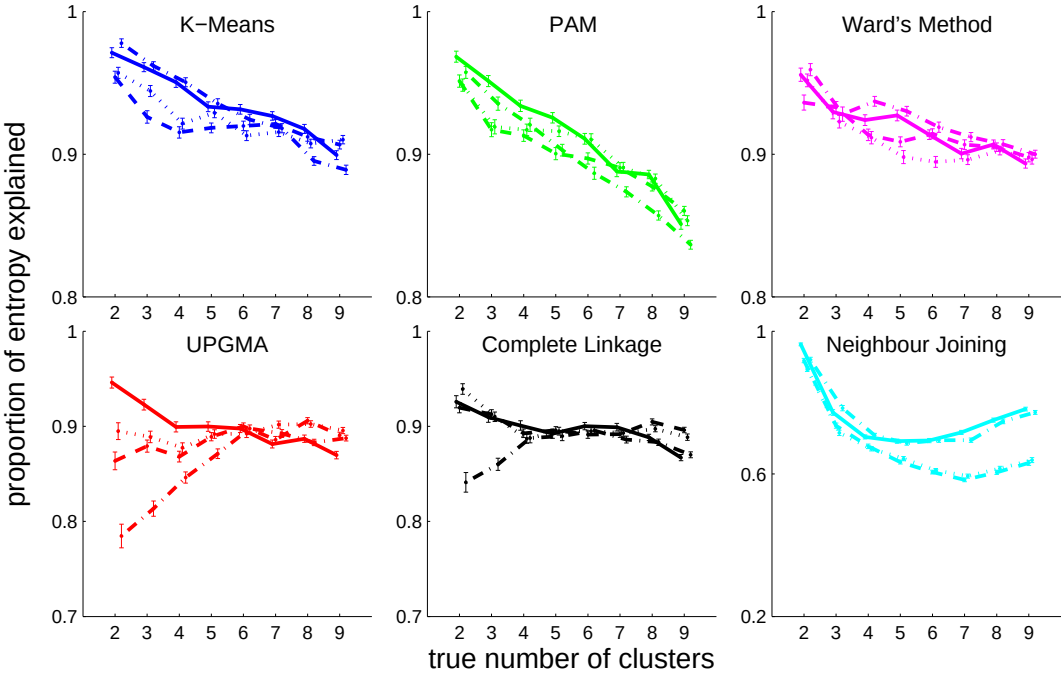

Figure A.5: As Figure A.4, but for synthetic datasets of 50 compositional profiles.

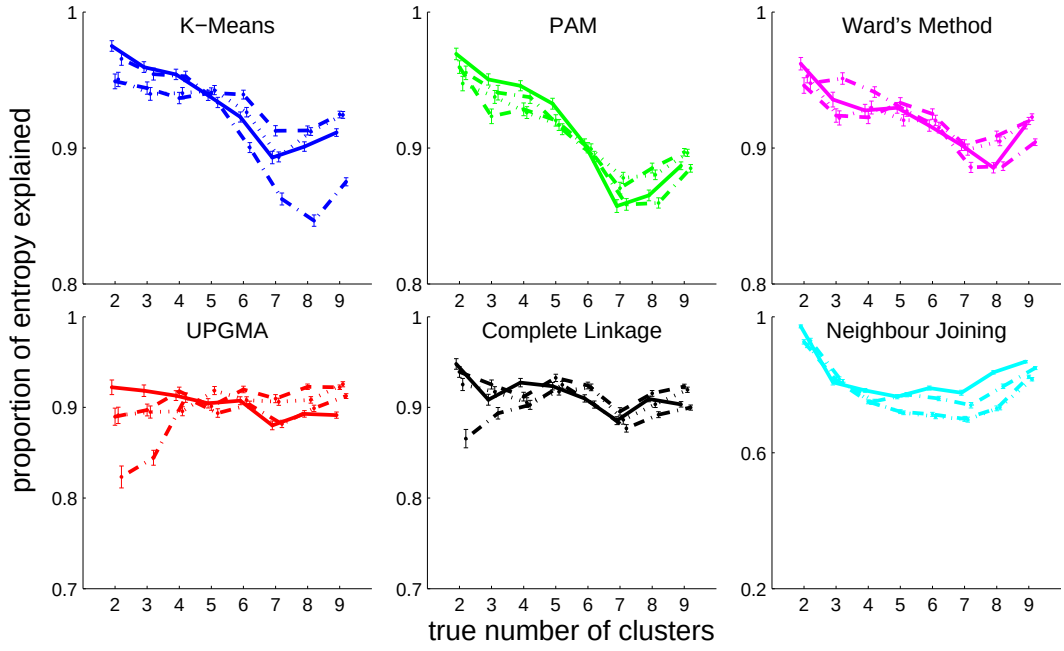

Figure A.6: As Figure A.4, but for synthetic datasets of 20 compositional profiles.

A.2.2 *PE* by distance metric and algorithm for synthetic datasets of 500, 200, 50 and 20 profiles created with Hard Cluster Parameters

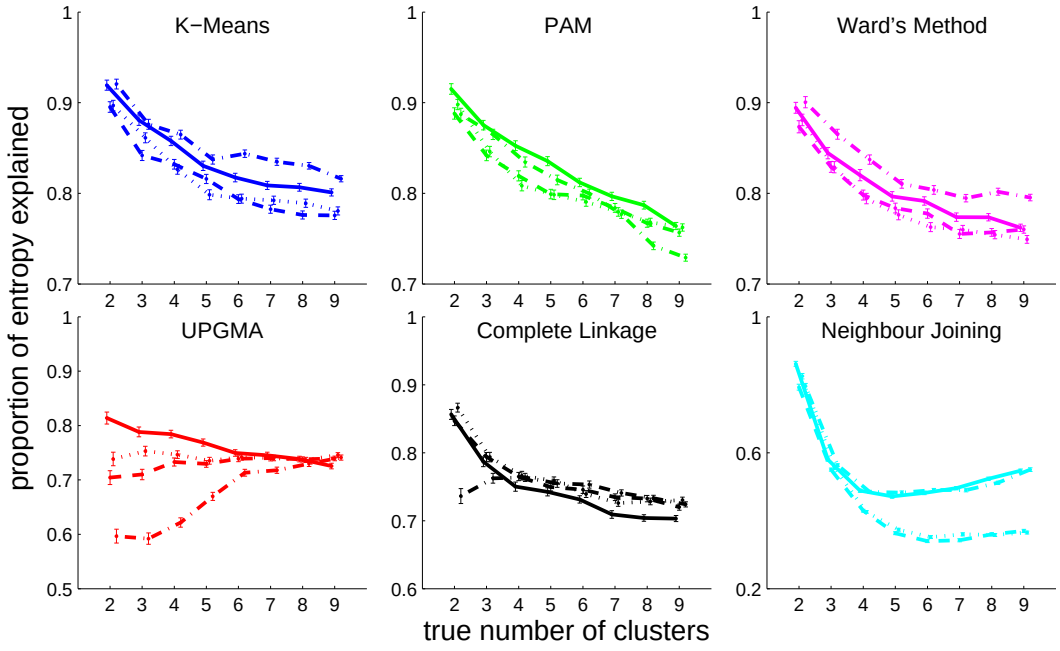

Figure A.7: Performance of each process on synthetic datasets of 500 compositional profiles with 2 to 9 true clusters which were produced using Hard Cluster Parameters. The *PE* is plotted against the number of synthetic clusters in the data set. Error bars for neighbour joining are indiscernibly small. Line styles and colors are as in Figure A.4.

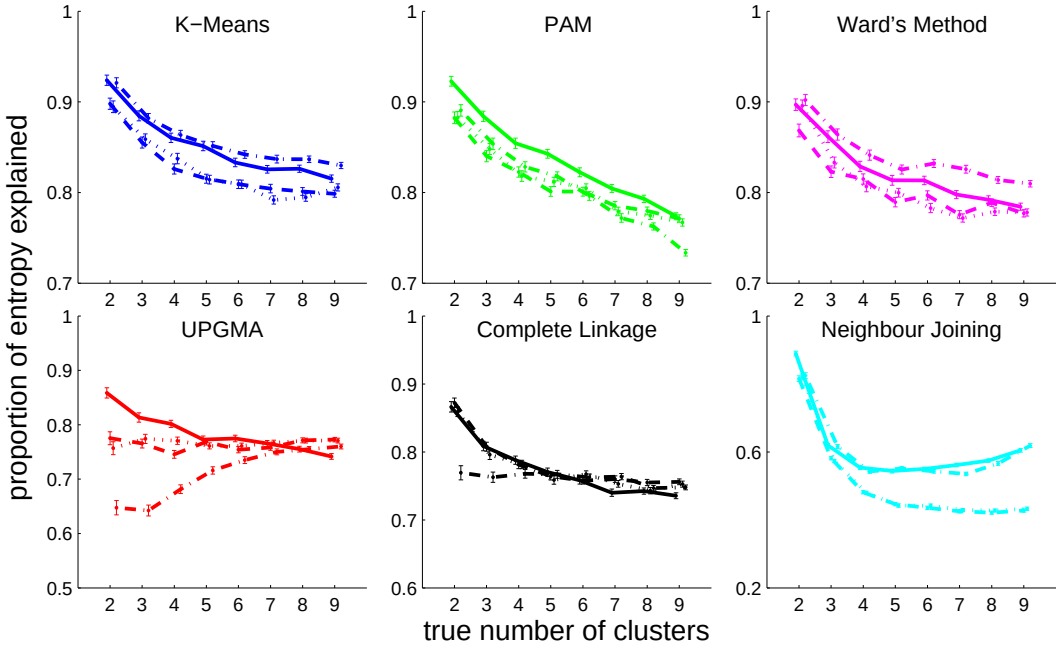

Figure A.8: As Figure A.7, but for synthetic datasets of 200 compositional profiles.

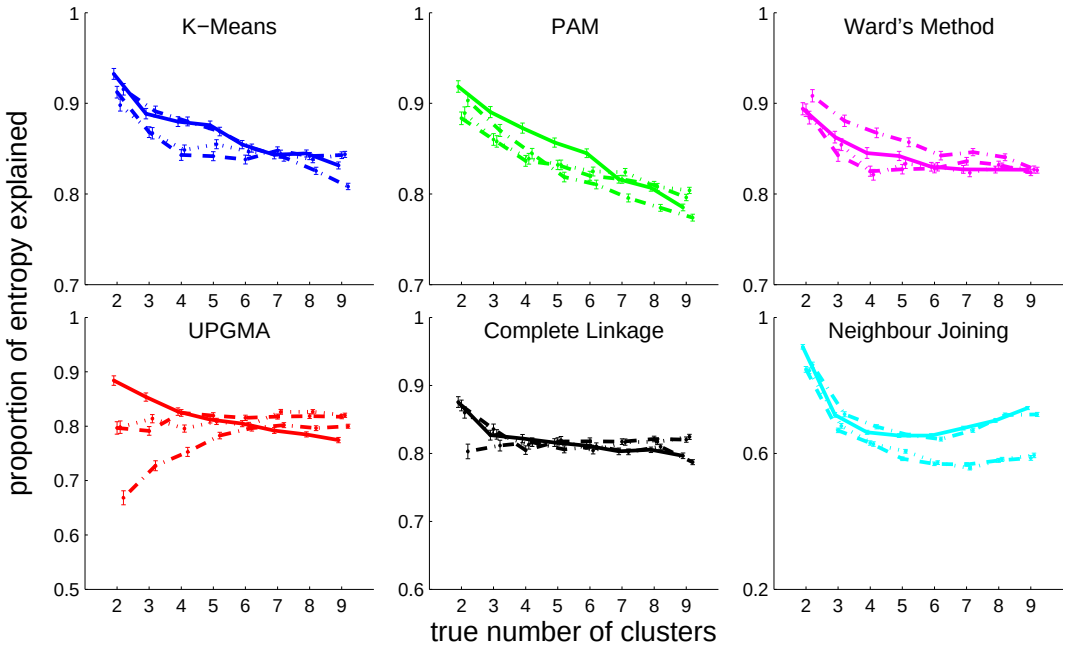

Figure A.9: As Figure A.7, but for synthetic datasets of 50 compositional profiles.

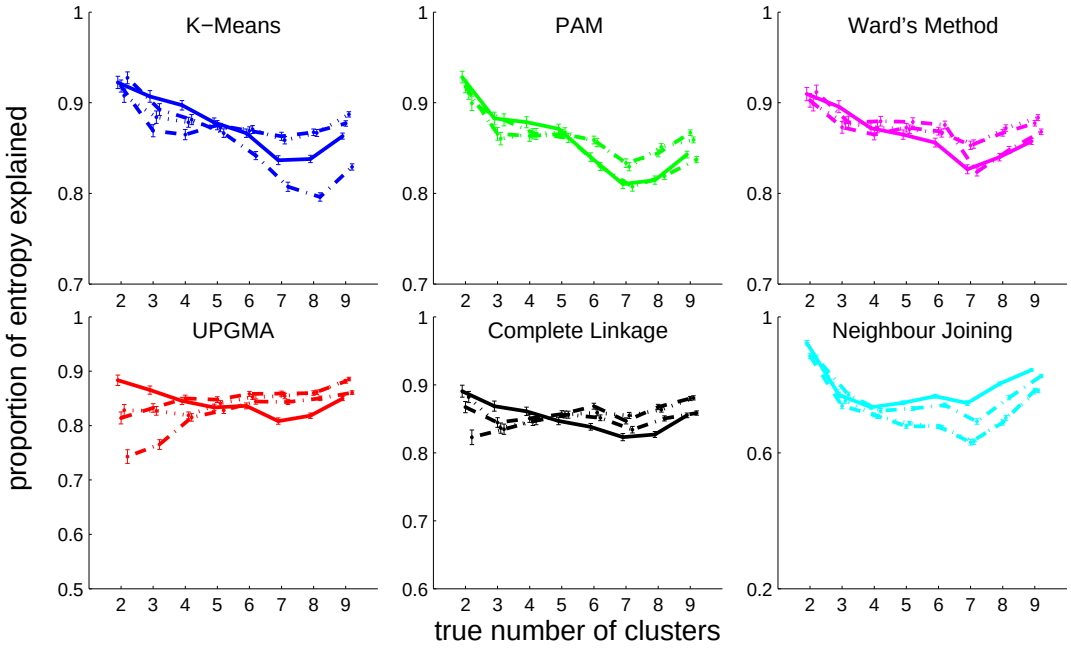

Figure A.10: As Figure A.7, but for synthetic datasets of 20 compositional profiles.

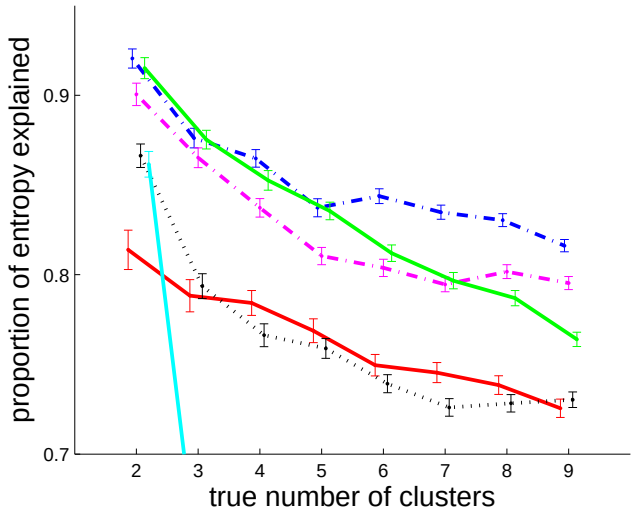

Figure A.11: Performance of each clustering algorithm using its best distance metric on synthetic datasets of 500 compositional profiles with 2 to 9 true clusters which were produced using Hard Cluster Parameters. The *PE* is plotted against the number of synthetic clusters in the data set. *PE* Values for neighbour joining at 3 to 9 clusters are below 0.6. Line styles and colors are as in Figure A.4.

A.2.3 *PE* by distance metric and algorithm for synthetic datasets of 500 profiles created with Cluster Parameters based on complete linkage clustering.

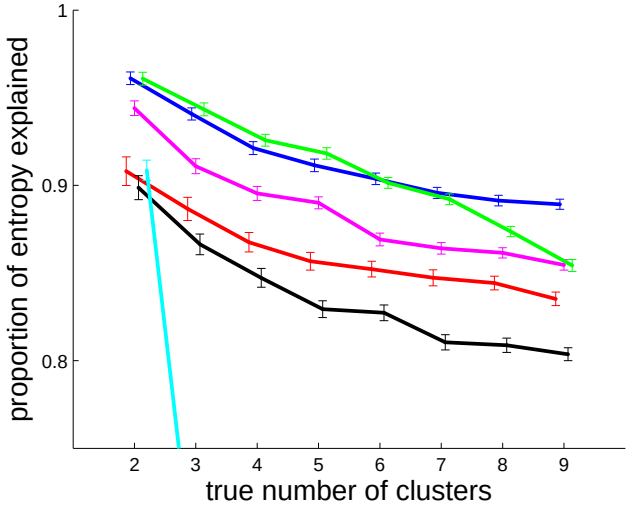

Figure A.12: Performance of each clustering algorithm using Euclidean distance on synthetic datasets of 500 compositional profiles with 2 to 9 true clusters which were produced Cluster Parameters based on complete linkage clustering. The *PE* is plotted against the number of synthetic clusters in the data set. *PE* Values for neighbour joining at 3 to 9 clusters are below 0.7. Line styles and colors are as in Figure A.4.

A.2.4 Proportion finding correct number of clusters by distance metric and algorithm for synthetic datasets of 500 profiles.

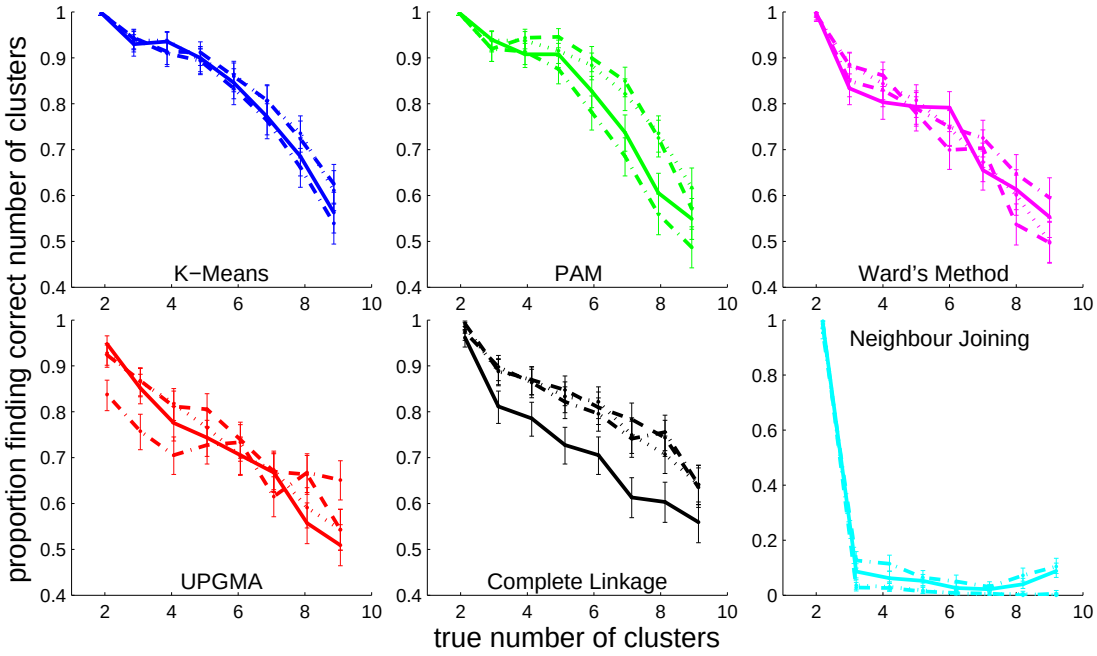

Figure A.13: Performance of each clustering algorithm by distance metric on synthetic datasets of 500 compositional profiles with 2 to 9 true clusters. Each clustering algorithm was applied to 500 synthetic datasets and gave from 2 to 12 clusters. The optimal number of clusters was found for each trial using the technique of Calinski and Harabasz. The proportion of trials identifying the correct number of clusters is plotted against the number of synthetic clusters in the data set. Line styles and colors are as in Figure A.4.

A.2.5 *PE* by distance metric and algorithm for BV status from pooled and unpooled data in 5 clinical trials.

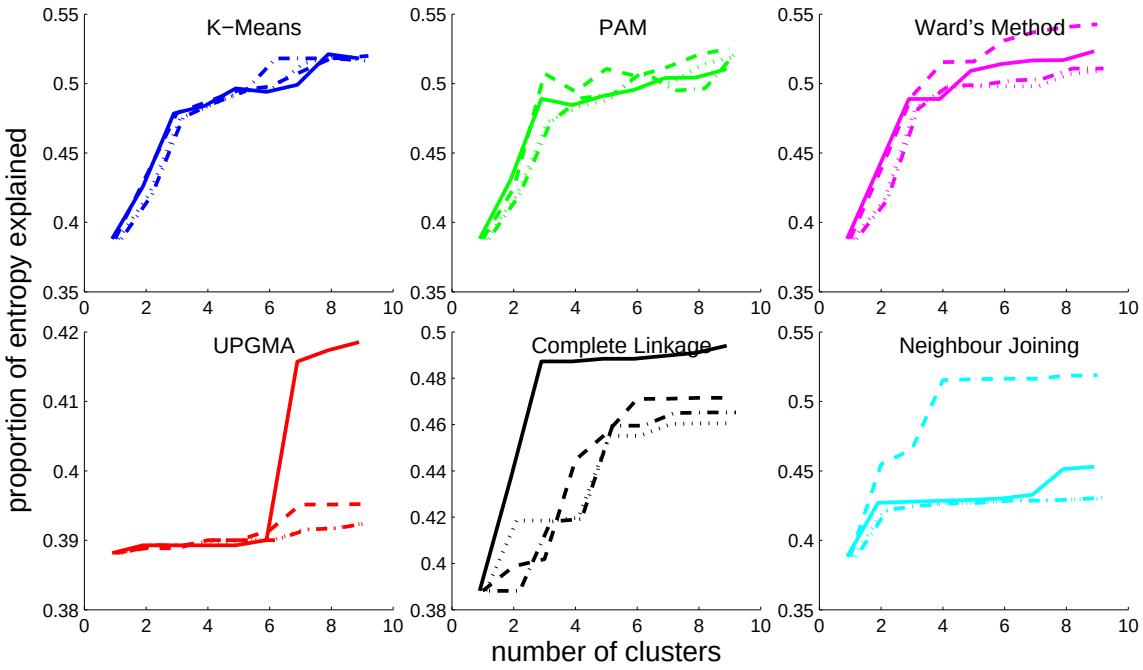

Figure A.14: Performance of each clustering algorithm by distance metric on pooled data from five clinical studies with 2 to 9 clusters. The PE of the subject's BV status is plotted against the number of *a priori* clusters the algorithm was asked to find. Line styles and colors are as in Figure A.4.

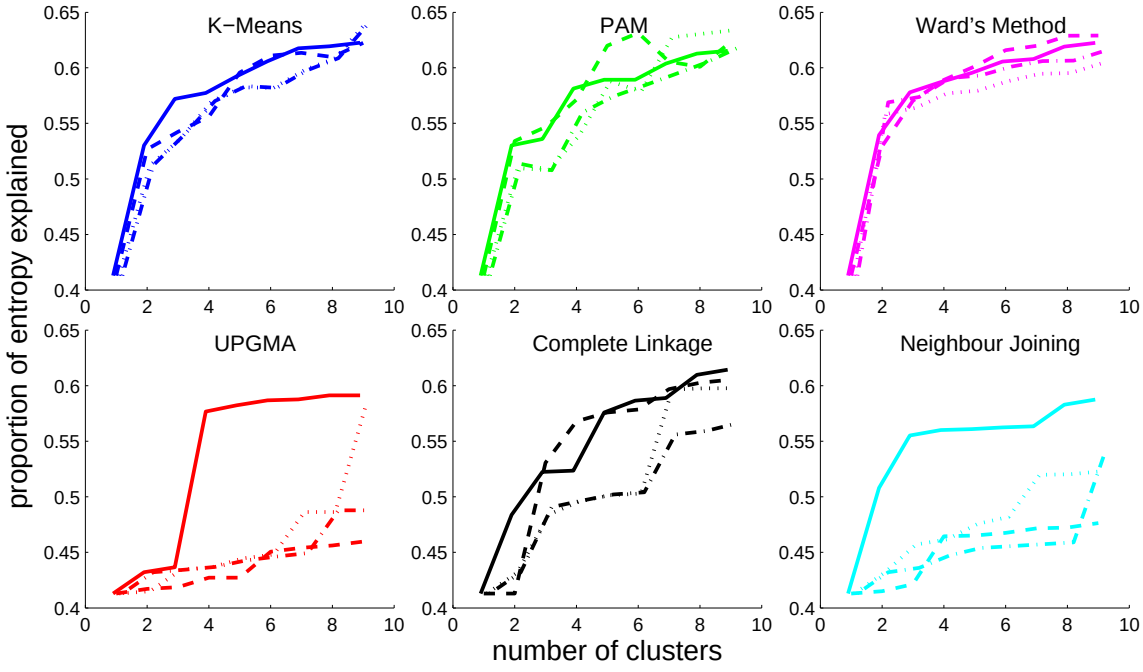

Figure A.15: As Figure A.14, but for clinical data from the Tanzania HIV study.

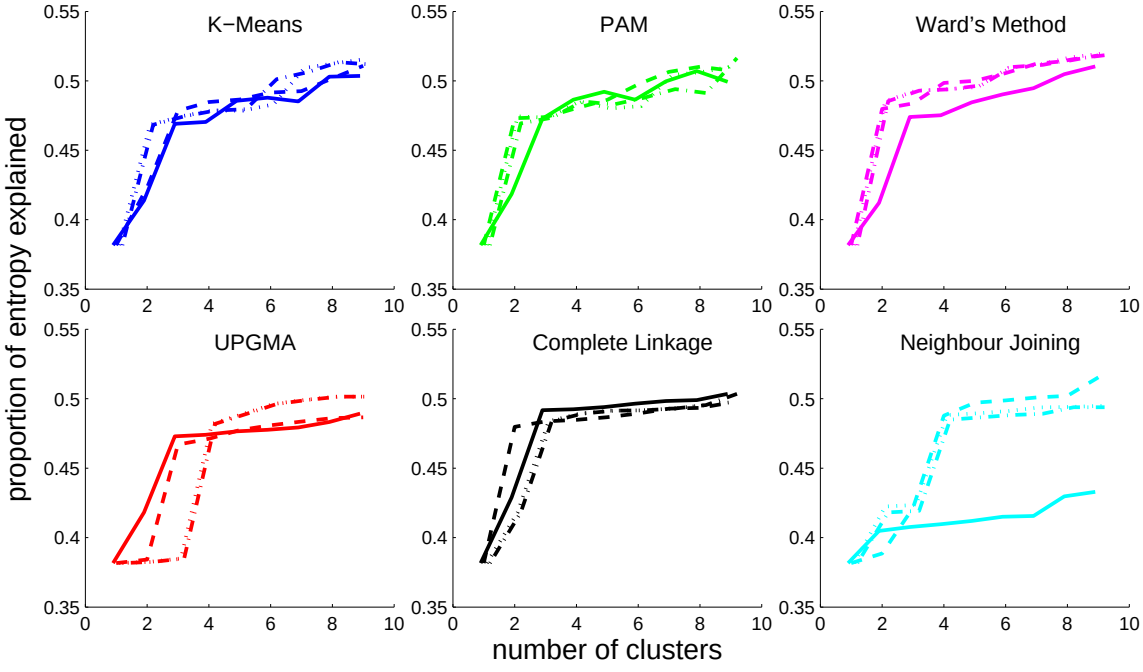

Figure A.16: As Figure A.14, but for clinical data from the Brazil BV study.

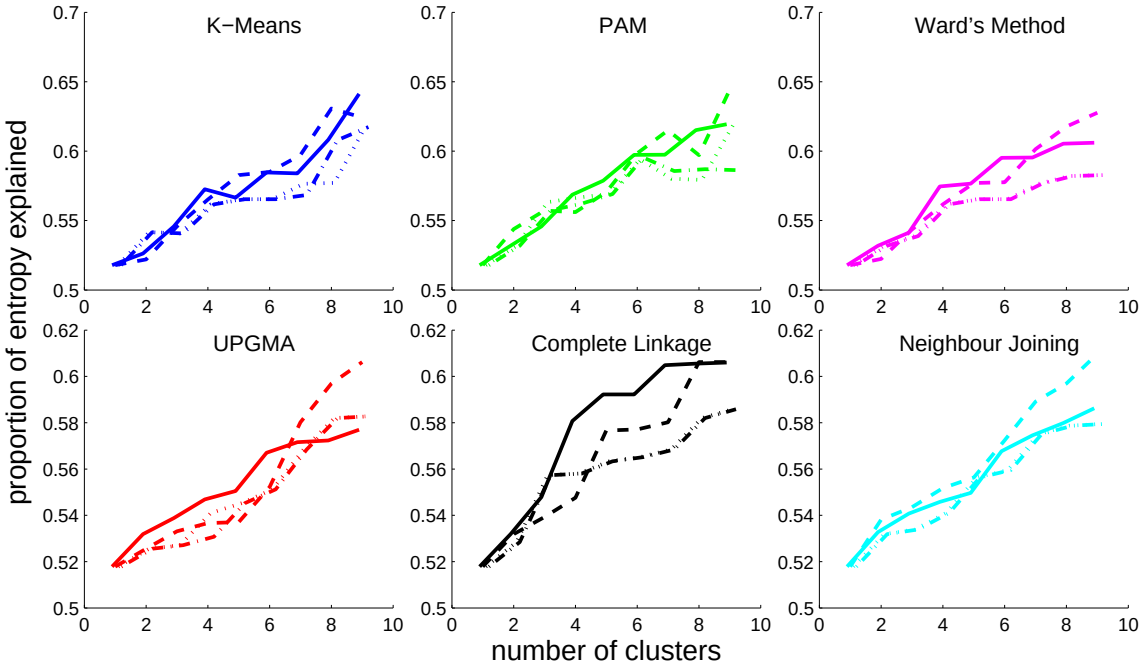

Figure A.17: As Figure A.14, but for clinical data from the Canadian post-menopause study.

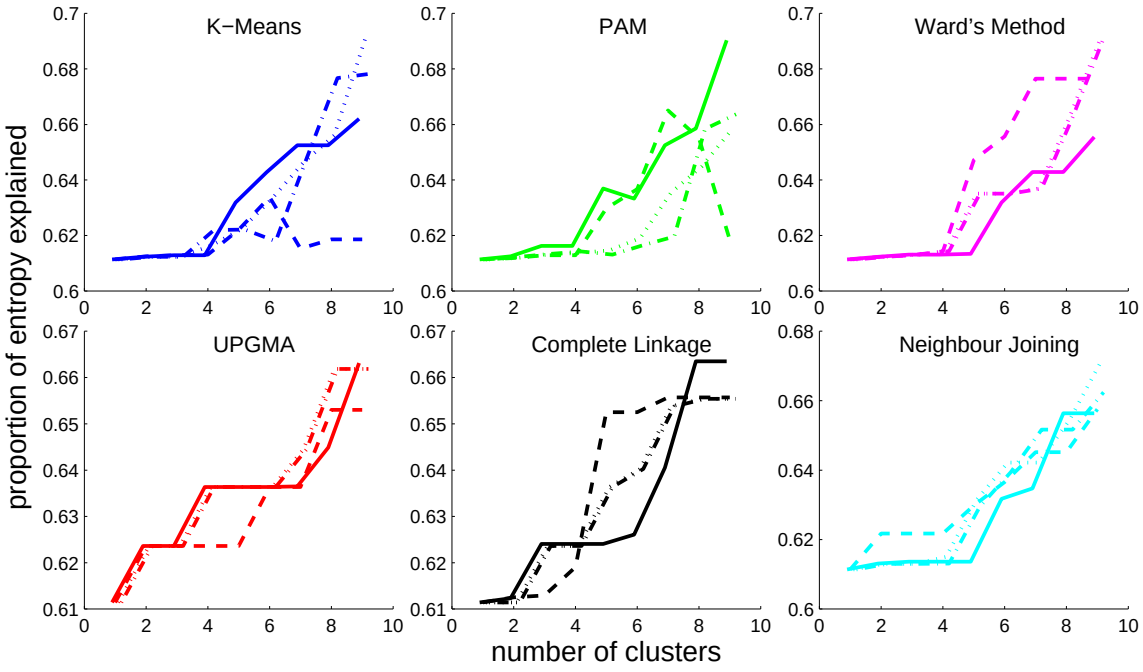

Figure A.18: As Figure A.14, but for clinical data from the Canadian preterm labour study.

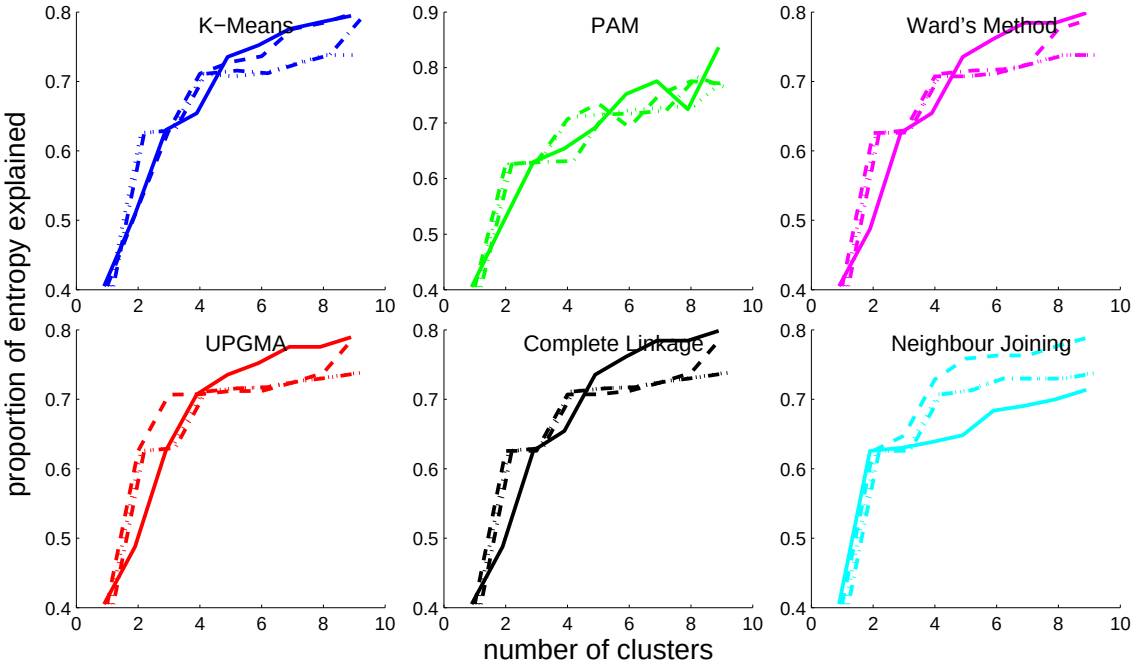

Figure A.19: As Figure A.14, but for clinical data from the Canadian toxin shock study.

A.2.6 *PE* by distance metric and algorithm for vaginal pH value from pooled and unpooled data in 2 clinical trials.

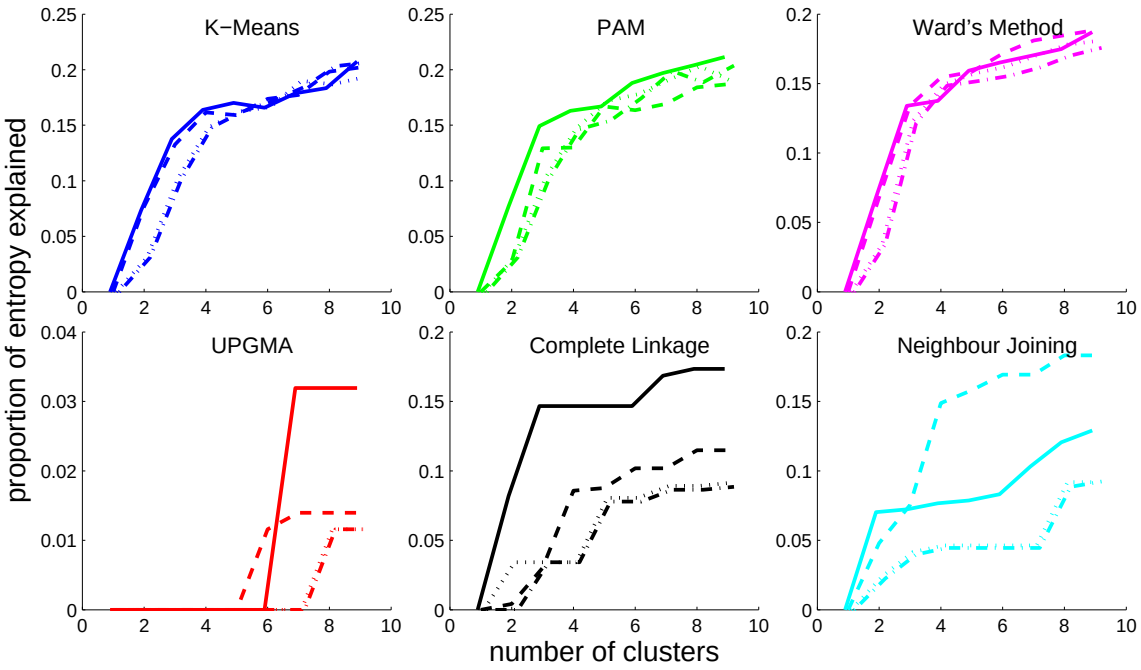

Figure A.20: Performance of each clustering algorithm with by distance metric on pooled data from two clinical studies with 2 to 9 clusters. The PE of the subject's vaginal pH level is plotted against the number of *a priori* clusters the algorithm was asked to find. Line styles and colors are as in Figure A.4.

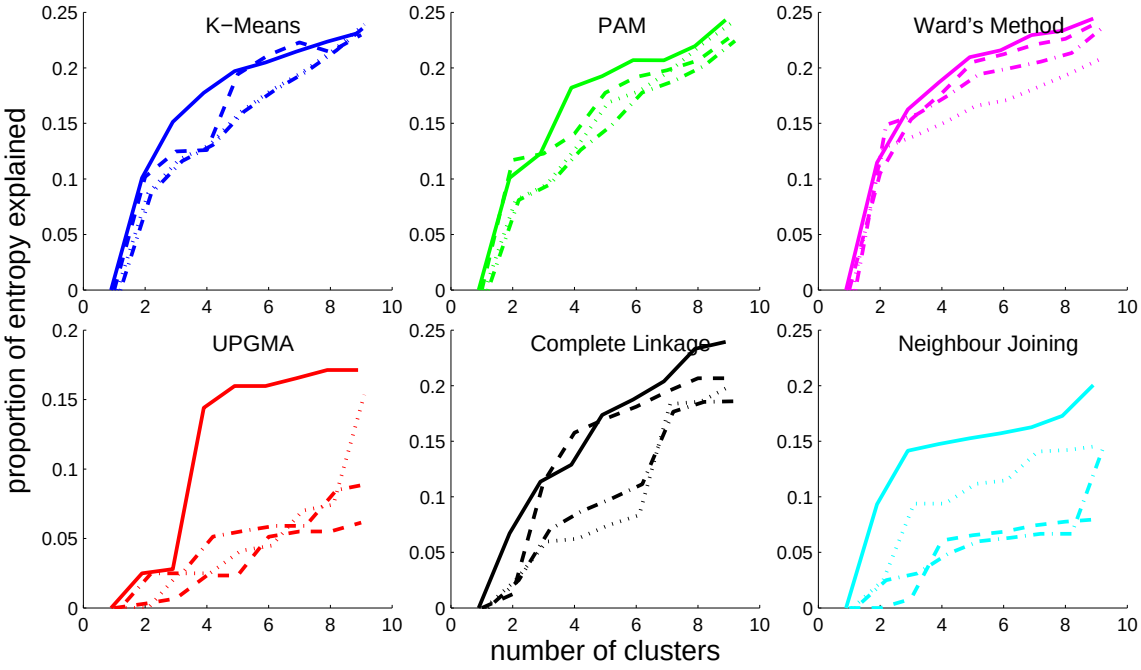

Figure A.21: As Figure A.20, but for clinical data from the Tanzania HIV study.

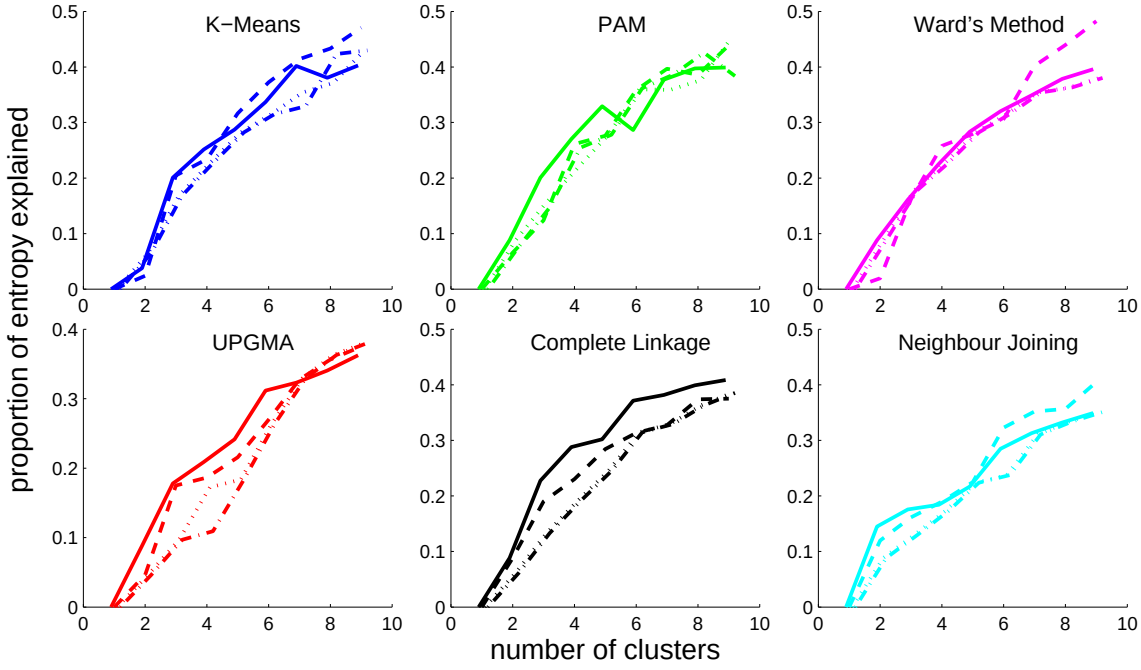

Figure A.22: As Figure A.20, but for clinical data from the Canadian post-menopause study.

A.2.7 Synthetic Data

We found that the single peaked beta distributions which had a Gaussian-like shape emulated common OTUs well, as was the case in Figure 1.3. They emulated rare OTUs like that in Figure 1.2 well when they took on an exponential-like shape with a maximum at 0. For OTUs which had frequency distributions with two peaks, a double peaked beta distribution was effective as shown in Figure 1.4.

Curriculum Vitae

Name:	Erik Cameron
Post-Secondary Education and Degrees:	Trent University Peterborough, Ontario 2006 - 2010 BSc. University of Western Ontario London, ON 2010 - 2012 MSc.
Honours and Awards:	NSERC CGS M 2010-2011 Bagnani Medal (Trent University) 2010 - High standing on graduation NSERC USRA (Trent University) 2009 - Research award NSERC USRA (Trent University) 2008 - Research award
Related Work Experience:	Teaching Assistant The University of Western Ontario 2010 - 2012