REPORT

Alignment-free tools for metagenomics-data analysis

Robert Deibel

Abstract

Metagenomics is the analysis of microorganisms of biotopes, like the human gut. The high amount of reads obtained through NGS-methods as well as findings of unsequenced genomes in metagenominc data requires usage of lightweight tools independent of databases and alignment. In this report I introduce the two main branches of analysis tools, while setting the focus on alignment-free methods.

While the alignment-based approaches align a target sequence against a database – as seen with Smith-Waterman or BLAST – alignment-free methods have different approaches. Here I will showcase a selection of statistical and machine learning approaches.

$D2z$, Hao and d_2^* are statistical approaches based on k-tupel count frequencies, similarity is measured by applying a metric and through interpretation with current biological knowledge.

Usage of k-mers as vectors in high-dimensional space and BH-SNE visualizing related data in two dimensional scatter plots leads to an approach with high accuracy for simulated and real-world metagenomes alike. Especially for analysis of novel data, sampled from microbioms, alignment-free applications of metagenomics are essential for understanding the cooperation of microorganisms and for further research in immunology.

Keywords: alignment-free; machine learning; statistic; metagenome; report

Introduction

Metagenomics

A puddle of mud The metagenome is the whole set of transcripts of a population of microorganisms in a microbiome sample. Metagenomics is the study and analysis of these metagenomes [1].

A microbiome consists of countless bacteria, archea and viruses; out of which >90% are uncultureable, using sequencing and metagenomic analysis as a way to study these.

Metagenomics is used in the design of antibiotics and medicine or to analyze the metabolism of microorganisms, making it a rapidly developing field of research.

NGS – Next Generation Sequencing Advances in sequencing made metagenomics as a field possible. Nowadays new high throughput methods, also Next Generation Sequencing or NGS for short, are used to generate comparable data from real-world samples. NGS is a term for methods of rapid parallelized sequencing, producing thousands or millions of reads concurrently.

Data analysis can the be carried out on these reads.

Goals Metagenomics-data, analyzed through alignment-free methods, allows insight into the ways of a broad spectrum of microorganismal life. Differing from the classical approach of alignment, where newly found sequences are often unanalyzed, these alignment-free methods could provide first evidence of origin and function on novel sequences.

In this report, I want to state two branches of metagenomic analysis, while focusing on alignment-free methods. State different approaches, the strengths and weaknesses.

Metagenomics analysis

Alignment-based methods The popular approach to analyze reads are the various alignment-based methods.

Sequences are aligned against a database of known transcripts, the resulting profiles are analyzed based on several factors.

This approach is well established and implemented numerous times. BLAST for example has an accuracy well over 80%[4] and similar values are expected for other alignment-based tools. Still BLAST is not used for metagenomics anymore due to low speed.

However, NGS supplies researchers with a lot of data

Correspondence: robert.deibel@student.uni-tuebingen.de
Eberhard-Karls Universität, Tübingen, DE
Full list of author information is available at the end of the article

to be analyzed. The analysis of metagenomes is computation heavy. BLAST, and therefore BLAST-like tools, aligns queries with the entries in a database. Applied on metagenomic data this method can be time consuming.
Research of novel sequences is a main focus of metagenomics. This data stays unanalyzed following an alignment-based approach, resulting in a high demand for lightweight tools independent of databases.

Here, I will showcase methods with differed approaches to the analysis of such data.

An alternative for alignment-based methods

The work of Song *et al.* [5] and Laczny *et al.* [6] utilizes statistics, and visualization and machine learning for alignment-free approaches of analysis, respectively.
The statistical methods stated by Song *et al.* are on their own inapt to analyze metagenomes, as they examine the similarity of two sequences. After application to every pair of sequences, a similarity or dissimilarity matrix is achieved. This can then be used for clustering or visualization.
Laczny *et al.* utilizes a number of tools to construct a promising approach to alignment-free analysis.

Methods

In this section "power" refers to the statistical term, which is the probability that a test correctly rejects the null hypothesis. With H_0: the sequences are unrelated; and H_1: the sequences are related under the underlying model. If not stated differently.

k-tuples as a measure of similarity

Song *et al.*[5] present different methods based on k-tuple occurrences in sequences. A k-tuple is a substring of length k.
By counting the occurrences of these k-tuples and applying a distance or dissimilarity metric, the k-tuples are clustered and these clusters analyzed. Different metrics are stated as statistics, which are a function of two sequences, computing the similarity of those.

The D_2 statistic Torney *et al.*[7] introduced D_2 using shared k-tuple occurrences between sequences to define similarity.

$$D_2 = \sum_{w \in \mathcal{A}^k} X_w Y_w$$

where X_w and Y_w are the numbers of occurrences of string w in the corresponding sequence A or B, $\mathcal{A}$ is the alphabet and k is the length of w.

Kantrovitz *et al.*[8] stated that the D_2 statistic depends on the underlying sequence model and performed a normalization to remove the bias. A sequence model describes the different background probabilities of the observed sequences A and B, here Markov models are used. The resulting statistic is called $D2z$ and is defined as

$$D2z(A,B) = \frac{D_2(A,B) - E(D_2)}{\sqrt{Var(D_2)}}$$

The expected value and variance are calculated through consideration of background Markov models for the sequences.
$D2z$ was compared to five other measures of similarity – [5, 8] – through analysis of *cis*-regulatory modules (CRM), outperforming all of them. However $D2z$ requires two parameters; k and r, where k is the length of w and r is the order of the sequence Markov chain. The order of a MC states the dependence of the probability on the r previous states. A MC of order zero is just the probability of a state, while a MC of order 1 is the probability of the last state given the previous state $P(w_n|w_{n-1}\dots w_1) = P(w_n|w_{n-1})$

Expected value using Markov models For these calculations the order of the Markov chain is essential.
Generally the expected value $E(D_2)$ is the probability that the sequences are identical for a word of length k. For a MC of r=0 this is equal to the sum of the background probabilities $f_a^A\ f_a^B$ to the power of k, where a is the letter .

$$E(D_2) = P(I(A = B) = 1) = \left(\sum_{a \in \mathcal{A}} f_a^A f_a^B\right)^k$$

where I is the indicator function.
For a MC with r=1, $E(D_2)$ is the sum of all probabilities of words with length k.

$$\sum_{|w|=k} P^A(w) P^B(w)$$

where $P^A(w) = p^A(w_1)p^A(w|w_1)$ (for B analogous). Since $r = 1$ the probability is given by the probability for the first letter multiplied by the probability of the word given the first letter. Similar calculations can be done for MC of higher order.
The calculation of the variance relies on its relationship with covariance, further insight into this is given in [8].

Phylogenetic trees through statistics – CVTree Another approach utilizes the expected count of a k-tupel under the $(k-2)$-th order Markov chain, estimated by

$$E_w^X = \frac{X_w X_{w_2 \dots w_k}}{X_{w_2 \dots w_{k-1}}}$$

where w is a substring of length k, w_i is the letter at index i in w and X_w is the number of occurrences of w in a sequence A.
The correlation coefficient of the relative difference vectors with the expected count is then used to measure similarity of sequences[5].

$$Hao = \frac{1}{2}(1 - C)$$

Hao calculates the frequencies of observations of overlapping k-tupels indicated with X_w in a composition vector and subtracts a random background using the $(k-2)$-th order Markov chain. This step is similar to the normalization in $D2z$ and is to minimize the influence of random mutation. After computation of correlation C a normalization was defined by subtraction of C from 1 and multiplication with .5.

$$C = \frac{\sum_w \left(\frac{X_w - E_w^X}{E_w^X}\right)\left(\frac{Y_w - E_w^Y}{E_w^Y}\right)}{\sqrt{\sum_w \left(\frac{X_w - E_w^X}{E_w^X}\right)^2 \sum_w \left(\frac{Y_w - E_w^Y}{E_w^Y}\right)^2}}$$

C describes the cosine of the angle between the composition vectors of the sequences, where $C = 1 \Leftrightarrow A = B$ and $C = 0 \Leftrightarrow \forall a_i \in A,\, b_i \in B : a_i \neq b_i$.
For CVTree a distance matrix is computed by applying Hao's method to each pair of sequences. Neighbor joining is then used to construct a phylogenetic tree.
CVTree was tested by Qi *et al.* on a set of 139 prokaryotic genomes computing a robust result[9].
The application of CVTree on virus data, provided on the web service, shows the relationship of the provided data (Figure 1).

Similarity through nucleotide frequency A related approach was presented by Karlin and colleagues, where they observed the relative di-nucleotide frequency defined by

$$\rho_{ab}(A) = \frac{f_{ab}}{f_a f_b}$$

where f_w is the frequency of w in a sequence.
It is stated that ρ_w is stable across a genome and differs in different genomes. The extension to tri- and tetranucleotides is achieved by

$$\gamma_{abc} = \frac{f_{abc} f_a f_b f_c}{f_{ab} f_{bc} f_{aNc}}$$

and

$$\tau_{abcd} = \frac{f_{abcd} f_{ab} f_{aNc} f_{aN_1N_2d} f_{bc} f_{bNd} f_{cd}}{f_{abc} f_{abNd} f_{bcd} f_a f_b f_c f_d}$$

l_p norm was applied as a dissimilarity measure as

$$\delta(A,B) = \sum_{j \in A} |\theta_j(A) - \theta_j(B)|$$

where A, B are sequences, $\theta \in \{\rho, \gamma, \tau\}$ and

$$j = \begin{cases} \{a,b\} & \text{if } \theta = \rho \\ \{a,b,c\} & \text{if } \theta = \gamma \\ \{a,b,c,d\} & \text{if } \theta = \tau \end{cases}$$

Evolutionary studies on viruses, bacteria, plasmids, prokaryotes and eukaryotes were performed using this measure[5].

D_2^S, D_2^*, d_2^S and d_2^* The D_2^S statistic is defined by

$$D_2^S = \sum_{w \in \mathcal{A}^k} \frac{\widetilde{X}_w \widetilde{Y}_w}{\sqrt{\widetilde{X}_w^2 + \widetilde{Y}_w^2}}$$

where $\widetilde{X}_w$ and $\widetilde{Y}_w$ are the normalization of X_w and Y_w respectively

$$\widetilde{X}_w = X_w - \bar{n} p_w^X$$

$\bar{n} = n - k$ and p_w^X is the probability of the k-tupel w under the background model of a sequence A. This is based on Shepp[12]. Where it was observed that for two normal random variables with mean zero, $XY/\sqrt{X^2 + Y^2}$ is also normally distributed.
D_2^* defined by

$$D_2^* = \sum_{w \in \mathcal{A}^k} \frac{\widetilde{X}_w \widetilde{Y}_w}{\sqrt{\bar{m}\bar{n} p_w^X p_w^Y}}$$

utilizes the idea that the number of occurrences of w is approximately Poisson distributed and mean and variance are the same.
Through simulations and theoretical studies the null hypothesis H_0 was tested against H_1; the conclusions were:

1 D_2^S and D_2^* have higher power than D_2 increasing with sequence length
2 D_2^* has the highest power when the length of k equals the $motif$ length
3 For short sequences the power of D_2^* is higher while for long sequences D_2^S is generally more powerful.

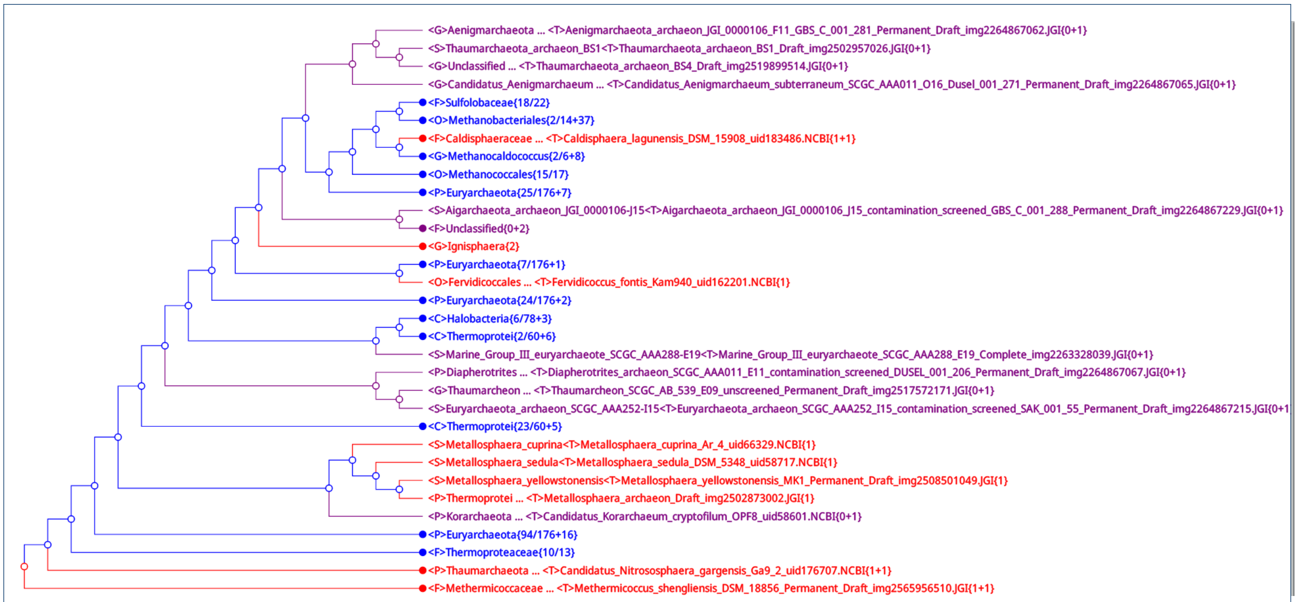
<G>Aenigmarchaeota ... <T>Aenigmarchaeota_archaeon_JGI_0000106_F11_GBS_C_001_281_Permanent_Draft_img2264867062.JGI{0+1}
<S>Thaumarchaeota_archaeon_BS1<T>Thaumarchaeota_archaeon_BS1_Draft_img2502957026.JGI{0+1}
<G>Unclassified ... <T>Thaumarchaeota_archaeon_BS4_Draft_img2519899514.JGI{0+1}
<G>Candidatus_Aenigmarchaeum ... <T>Candidatus_Aenigmarchaeum_subterraneum_SCGC_AAA011_O16_Dusel_001_271_Permanent_Draft_img2264867065.JGI{0+1}
<F>Sulfolobaceae{18/22}
<O>Methanobacteriales{2/14+37}
<F>Caldisphaeraceae ... <T>Caldisphaera_lagunensis_DSM_15908_uid183486.NCBI{1+1}
<G>Methanocaldococcus{2/6+8}
<O>Methanococcales{15/17}
<P>Euryarchaeota{25/176+7}
<S>Aigarchaeota_archaeon_JGI_0000106-J15<T>Aigarchaeota_archaeon_JGI_0000106_J15_contamination_screened_GBS_C_001_288_Permanent_Draft_img2264867229.JGI{0+1}
<F>Unclassified{0+2}
<G>Ignisphaera{2}
<P>Euryarchaeota{7/176+1}
<O>Fervidicoccales ... <T>Fervidicoccus_fontis_Kam940_uid162201.NCBI{1}
<P>Euryarchaeota{24/176+2}
<C>Halobacteria{6/78+3}
<C>Thermoprotei{2/60+6}
<S>Marine_Group_III_euryarchaeote_SCGC_AAA288-E19<T>Marine_Group_III_euryarchaeote_SCGC_AAA288_E19_Complete_img2263328039.JGI{0+1}
<P>Diapherotrites ... <T>Diapherotrites_archaeon_SCGC_AAA011_E11_contamination_screened_DUSEL_001_206_Permanent_Draft_img2264867067.JGI{0+1}
<G>Thaumarcheon ... <T>Thaumarcheon_SCGC_AB_539_E09_unscreened_Permanent_Draft_img2517572171.JGI{0+1}
<S>Euryarchaeota_archaeon_SCGC_AAA252-I15<T>Euryarchaeota_archaeon_SCGC_AAA252_I15_contamination_screened_SAK_001_55_Permanent_Draft_img2264867215.JGI{0+1
<C>Thermoprotei{23/60+5}
<S>Metallosphaera_cuprina<T>Metallosphaera_cuprina_Ar_4_uid66329.NCBI{1}
<S>Metallosphaera_sedula<T>Metallosphaera_sedula_DSM_5348_uid58717.NCBI{1}
<S>Metallosphaera_yellowstonensis<T>Metallosphaera_yellowstonensis_MK1_Permanent_Draft_img2508501049.JGI{1}
<P>Thermoprotei ... <T>Metallosphaera_archaeon_Draft_img2502873002.JGI{1}
<P>Korarchaeota ... <T>Candidatus_Korarchaeum_cryptofilum_OPF8_uid58601.NCBI{0+1}
<P>Euryarchaeota{94/176+16}
<F>Thermoproteaceae{10/13}
<P>Thaumarchaeota ... <T>Candidatus_Nitrososphaera_gargensis_Ga9_2_uid176707.NCBI{1+1}
<F>Methermicoccaceae ... <T>Methermicoccus_shengliensis_DSM_18856_Permanent_Draft_img2565956510.JGI{1+1}

Figure 1 Example run of CVTree with Virus data and $k = 6$ executed with web service at `http://tlife.fudan.edu.cn/archaea/cvtree/cvtree3/` [10, 11]

Here $motifs$ are significantly enriched word patterns [5].

Further normalization of D_2^S and D_2^* removes the property that the magnitude strongly varies depending on different factors. The resulting statistics are d_2^S and d_2^* respectively.

$$d_2^S = \frac{1}{2}\left(1 - \frac{D_2^S}{\sqrt{\sum_{w\in\mathcal{A}^k} \frac{\widetilde{X}_w^2}{\sqrt{\widetilde{X}_w^2+\widetilde{Y}_w^2}}}\sqrt{\sum_{w\in\mathcal{A}^k} \frac{\widetilde{Y}_w^2}{\sqrt{\widetilde{X}_w^2+\widetilde{Y}_w^2}}}}\right)$$

and

$$d_2^* = \frac{1}{2}\left(1 - \frac{D_2^*}{\sqrt{\sum_{w\in\mathcal{A}^k} \frac{\widetilde{X}_w^2}{\bar{n}p_w^X}}\sqrt{\sum_{w\in\mathcal{A}^k} \frac{Y_w^2}{\bar{m}p_w^Y}}}\right)$$

d_2^S and d_2^* now hold the property that they are 0 when the sequences are the same and close to 1 if they are anti-correlated. Through this d_2^S and d_2^* can be used to cluster sequences of interest.

Modification of d_2^S and d_2^* The statistics mentioned above consider exact matches of tupels, since mutations are a fundamental part of evolution and DNA replication in general, mismatches should be considered when applying these methods. For a tupel w its neighborhood can be defined as $\varsigma(w)$ with $w' \in \varsigma(w)$ when w' has up to a certain number of mismatches with w and a weight a is applied, analogously reverse complements can be included in $\varsigma(w)$.
The statistics are modified as $\widetilde{X}_w$ is replaced by $\widetilde{X}_{\varsigma(w)}$ where $\widetilde{X}_{\varsigma(w)} = X_{\varsigma(w)} - EX_{\varsigma(w)}$ and

$$X_{\varsigma(w)} = \sum_{w'\in\varsigma(w)} a_{w'}X_{w'}$$

modifications for $\widetilde{Y}_w$ are analogous. Song *et al.* performed a series of tests to evaluate the effectiveness of the statistics Hao, d_2^S and d_2^* with consideration of mismatches. The neighborhood was defined as

$$\varsigma(w) = \{w', rc(w')|dist_{hamming}(w, w') \le 1\}$$

Testing was based on sequences taken from mouse embryos. The positive set was taken from the forebrain, midbrain, limb and heart tissues, while the negative set was chosen from random samples of the same length with a maximum of 30% repetitive sequences [5].
For 500 samples of each set the dissimilarity was calculated for each pair in the respective set. A threshold for dissimilarity was applied on the resulting values, a score lower than the threshold was predicted as positive, one above indicated negatives. Through comparison with the real data false predictions were identified. Different parameters like tupel size k, Markov chain order and mismatch weight were applied.

Performance under the mismatch model Conclusions of these test were:

1 Hao performed worse than both d_2^S and d_2^*
2 d_2^S and d_2^* performed best with $k = 4$ and mismatch weight of around 0.05. However differences through mismatch weight were negligible. For $k = 5 \vee 6$ a weight close to 1 performed best.

Song *et al.* considered additional statistics for testing, which are not discussed in this report and were therefore not further accounted for.

Additional testing using metagenomes and NGS data was carried out [5], since usually the short reads generated through NGS reduce the power of the discussed statistics. The data consisted of 39 fecal samples of 33 mammalian hosts [13] 56 marine samples [14] and 13 human fecal samples [15] for metagenomes and tree species with unknown complete genomes as NGS data. d_2^S outperformed the other tested measures in terms of consistency and separation as was seen through the tree samples and human feces metagenome respectively [5].

Overall d_2^S produced the best results compared to all mentioned statistics with Hao and d_2^* having similar outcomes.

Machine learning

In his work van der Maaten [16] introduced a machine learning variant (BH-SNE) based on the observation that closely related objects induce a larger force upon each other than unrelated ones. While these objects were originally intended to be points in a picture, Laczny *et al.*[6] used reads from metagenomes.

Barnes-Hut-SNE applies the Barnes-Hut algorithm and metric trees to modify the t-SNE method.

Barnes-Hut and vantage-point trees The Barnes-Hut algorithm is often used by astronomers to perform N-body simulations[16]. In this algorithm it is assumed that the force of objects with sufficient large distance to each another is infinitesimal and thus can be ignored in further computation. Leading –in the case of BH-SNE – to a decrease in objects to include in calculations.

Van der Maaten chose these objects based on vantage-point trees, where similar nodes are saved as the left and dissimilar nodes as the right child. After establishing the data structure one can search the tree and apply a given algorithm to the reduced set of nodes of interest.

Sequence signatures as objects Observations suggest the existence of species-specific oligonucleotide signatures in genomic sequences[6, 17]. These consist of k-mers and can be represented as vectors in high-dimensional Euclidean space. For human interpretation these vectors need to be transformed in a two or three dimensional space [6].

vector construction is achieved through assignment of joint probabilities to the k-mers and a similarity function to the corresponding points in high-dimensional space. Utilizing a Kullback-Leibler divergence and the optimizations stated before the points can be optimized.

Using center log-ratio (CLR)-transformation as a normalization, oligonucleotide signatures and the BH-SNE approach of van der Maaten, Laczny *et al.* constructed a tool for application on metagenomic-data with sequence length of 1kb and 5-mers as oligonucleotide signatures. While these parameters produced the best results Laczny*et al.* stated that 600 nt might be sufficient for some applications, but with lower values the separation of points would drop remarkably, as can be seen in Figure 2 through lesser separation of the clusters. Implementing their approach using 5-mers produced better congruency compared to transformed and untransformed 4-mers.

For Laczny *et al.* these 5-mers are the objects, used for calculation of similarity in BH-SNE.

Clustering The tool was tested on several simulated data sets; EqualSet01,EqualSet02 and LogSet01. The genomes of organisms in these sets were equally and logarithmically distributed; among the equally distributed sets were genomes with small and high similarity respectively.

They consisted of ten organisms depicted in Table 1 for EqualSet01 and LogSet01, and Table 2 for EqualSet02. The equally distributed data set is simplistic. As real-world metagenomes are never evenly distributed. The logarithmic set should simulate this real-world data with varying quantities of different genomes. High and low similarity sets test the discrimination capabilities of the tool.

After applying their tool on the simulated metagenomes their results showed distinct clustering for different species as depicted in Figure 3 for EqualSet01 and LogSet01, respectively. Clustering of EqualSet02 resulted in overlapping of closely related organisms and separation of more distant relatives.

Overall the runs on simulated data resulted in high sensitivity, specificity and accuracy(Table 3). The calculation was performed by enclosing clusters with polygons, by hand, as seen in Figure 3. Points inside represent the positives, points outside the negatives. Similar outputs were achieved by fitting (semi-)automated Gaussian Mixture models to calculate these values.

Table 1 Genomes of EqualSet01 and LogSet01 [6]

Organism	Genome size (nt)	%GC
Leifsonia xyli subsp. xyli str. CTCB07	2,584,158	67.7
Escherichia coli UTI89	5,065,741	50.6
Candidatus Carsonella ruddii PV	159,662	16.6
Haemophilus influenzae PittGG	1,887,192	38.0
Bacillus amyloliquefaciens FZB42	3,918,589	46.5
Brachyspira hyodysenteriae WA1	3,000,694	27.1
Geodermatophilus obscurus DSM 43160	5,322,497	74.0
Rickettsia prowazekii str. Dachau	1,109,051	29.0
Escherichia coli str. 'clone D i14'	5,038,386	50.6
Uncultured Termite group 1 bacterium phylotype Rs-D17	1,125,857	35.2

Table 2 Genomes of EqualSet02 [6]

Organism	Genome size (nt)	%GC
Lactobacillus delbrueckii subsp. bulgaricus ATCC 11842	1,864,998	49.7
Lactobacillus brevis ATCC 367	2,291,220	46.2
Lactobacillus casei ATCC 334	2,895,264	46.6
Lactobacillus gasseri ATCC 33323	1,894,360	35.3
Shewanella amazonensis SB2B	4,306,142	53.6
Shewanella putrefaciens CN-32	4,659,220	44.5
Shewanella baltica OS195	5,347,283	46.3
Shewanella frigidimarina NCIMB 400	4,845,257	41.6
Streptococcus suis A7	2,038,409	41.2
Streptococcus thermophilus CNRZ1066 chromosome	1,796,226	39.1

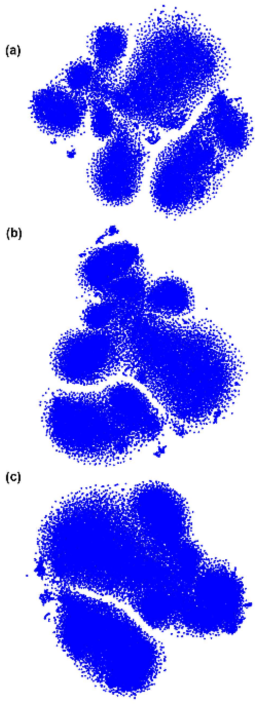

Figure 2 BH-SNE-based visualization of genomic fragment signatures for EqualSet01 (even community, overall reflecting distant taxonomic relatedness) with varying fragment lengths. (a) 800nt. (b) 600nt. (c) 400nt. – from Laczny *et al.*[6]

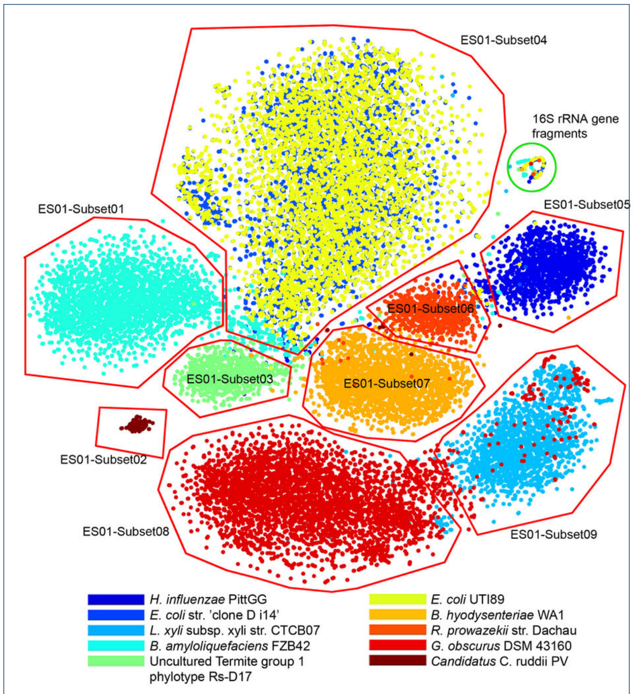

Figure 3 Figure taken from Laczny *et al.* [6] where red polygons mark clusters of congruent sets of interest used for calculation of sensitivity, specificity and accuracy. Colors mark different organisms as seen in the legend. The green polygon marks the cluster of 16s rRNA, forming a distinct group

Table 3 Sensitivity, specificity and accuracy of EqualSet01 – excerpt from Laczny *et al.*[6]

Subset	Sensitivity (%)	Specificity(%)	Accuracy(%)	Organism
01	90.06	99.99	99.94	B. *amyloliquefaciens*
02	91.25	100	100	*Candidatus* C. ruddii
03	95.42	99.90	97.57	Uncultured Termite group1 bacterium
04	98.60	98.23	96.67	E. *coli*

Application on real-world metagenomes Testing on real-world metagenomes of ground water [18], the human gut [19] and the deep sea [20] was also performed. They reported similar clustering (Figure 4) compared to simulated data with sensitivity, specificity and accuracy well above 90% for all subsets of the human gut metagenome with one exception, where accuracy was slightly below 80%.

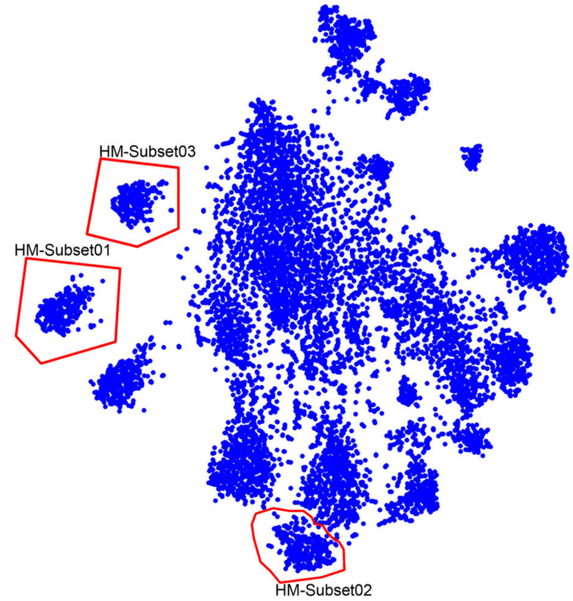

Figure 4 Clustering of human gut metagenome. Polygons represent clusters of great interest – from Laczny *et al.*[6]

The values were calculated using polygons to mark clusters and verifying these by comparison with the NCBI non-redundant nucleotide database.

The ground water metagenome also produced distinct clusters (Figure 5). Calculation of sensitivity, specificity and accuracy could not be carried out since they reported a lack of characterized reference genomes. Instead they used what they called ”essential genes” which can indicate the completeness of a genome. They reported four out of eight of these essential genes as over 80% complete, indicating a positive result for their tool.

As for the marine sample, the clusters, as seen in Figure 6, identified by the tool were linked to yet uncharacterized data.

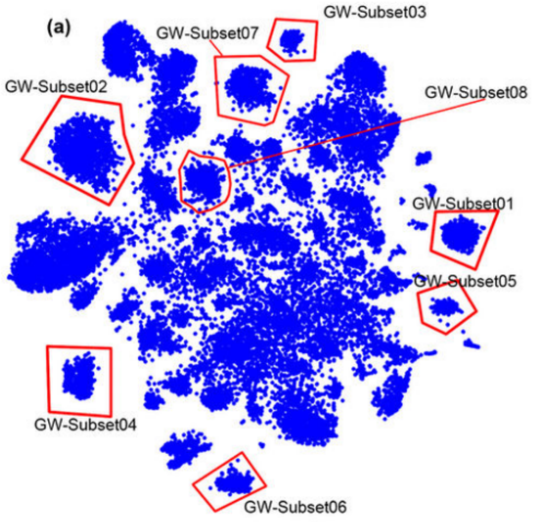

Figure 5 Clustering of groud water metagenome. Polygons represent clusters of great interest – from Laczny *et al.*[6]

Compared to an ESOM-based approach, tested with the same data sets, Laczny *et al.* reported better clustering of metagenomic-data. Additionally also a significant reduction in runtime [6] for their used metagenomes and good visualization capabilities as tested on simulated and real-world data.

Clustering seems robust on the applied metagenomes, while similar data tends to be near to each other in the visualization, 16S rRNA sequences form a distinct cluster, due to the high conservation of these regions in the genome.

As a downside sequences of 1kb are required to achieve good clustering, which are yet hard to gather through raw reads. Advancements in sequencing technologies are needed to fully utilize the capabilities of this tool.

Conclusions

The different methods for k-tuple counts and oligonuclotide signature enable an analysis independent on coding regions and fully sequenced genomes. This is of particular importance for the study of novel transcripts and a better understanding of mirobiotal organisms.

However, the statistics in Song *et al.* have to be applied

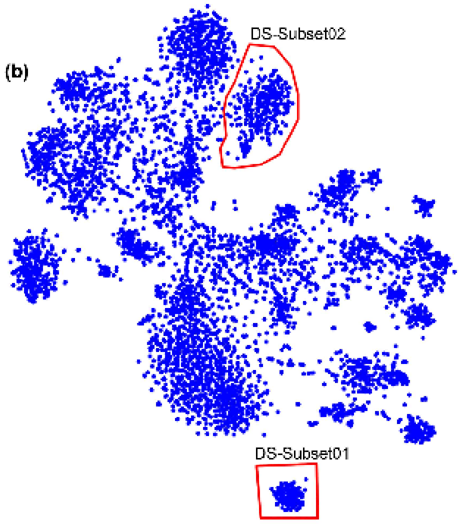

Figure 6 Clustering of deep sea metagenome. Polygons represent clusters of great interest – from Laczny *et al.*[6]

through a dissimilarity matrix to obtain meaningful results for metagenomic data, and rise in complexity as the power increases. This makes for slower computation of statistics like d_2^S compared to $D2z$. For new methods the results have to be normalized, as otherwise they are not comparable in metagenomic context. While Laczny *et al.*'s approach delivers good results for the tested metagenomes it is not clearly defined which clusters are of real significance and which can be ignored in further research. Without additional data, all clusters visible in the plot have to be analyzed manually. However, a general foundation of relationship is still established.

Furthermore I was not able to test any of the tools myself. This was due to different factors. While Laczny *et al.* state the steps leading to the construction of the tool, it is not publicly available so a considerable amount of work is needed to correctly rebuilt it.

The statistics-tools are either out of date or poorly documented. It is not apparent what parameters are required or how the computation is achieved. More often than not a particular file type is required, resulting in conflicts with publicly available metagenomic data. However, CVTree is still in development and a test run with provided data was successful (Figure 1), the upload of custom data was not possible at the time.

References

1. Handelsman, J.: Metagenomics: application of genomics to uncultured microorganisms. Microbiology and molecular biology reviews **68**(4), 669–685 (2004)
2. Kakirde, K.S., Parsley, L.C., Liles, M.R.: Size does matter: Application-driven approaches for soil metagenomics. Soil Biology and Biochemistry **42**(11), 1911–1923 (2010). doi:10.1016/j.soilbio.2010.07.021
3. Streit, W.R., Schmitz, R.A.: Metagenomics – the key to the uncultured microbes. Current Opinion in Microbiology **7**(5), 492–498 (2004). doi:10.1016/j.mib.2004.08.002
4. ESSINGER, S.D., ROSEN, G.L.: BENCHMARKING BLAST ACCURACY OF GENUS/PHYLA CLASSIFICATION OF METAGENOMIC READS, pp. 10–20. WORLD SCIENTIFIC, ??? (2012). doi:10.1142/9789814295291_0003. http://www.worldscientific.com/doi/pdf/10.1142/9789814295291_0003
5. Song, K., Ren, J., Reinert, G., Deng, M., Waterman, M.S., Sun, F.: New developments of alignment-free sequence comparison: measures, statistics and next-generation sequencing. Briefings in Bioinformatics **15**(3), 343–353 (2014). doi:10.1093/bib/bbt067
6. Laczny, C.C., Pinel, N., Vlassis, N., Wilmes, P.: Alignment-free visualization of metagenomic data by nonlinear dimension reduction. Scientific Reports **4**, 4516 (2014). Article
7. Torney, D.C., Burks, C., Davison, D., Sirotkin, K.M.: Computation of d2: a measure of sequence dissimilarity. In: Computers and DNA: the Proceedings of the Interface Between Computation Science and Nucleic Acid Sequencing Workshop, Held December 12 to 16, 1988 in Santa Fe, New Mexico/edited by George I. Bell, Thomas G. Marr (1990). Redwood City, Calif.: Addison-Wesley Pub. Co., 1990.
8. Kantorovitz, M.R., Robinson, G.E., Sinha, S.: A statistical method for alignment-free comparison of regulatory sequences. Bioinformatics **23**(13), 249–255 (2007)
9. Qi, J., Luo, H., Hao, B.: Cvtree: a phylogenetic tree reconstruction tool based on whole genomes. Nucleic acids research **32**(suppl_2), 45–47 (2004)
10. Qi, J., Wang, B., Hao, B.-I.: Whole proteome prokaryote phylogeny without sequence alignment: a k-string composition approach. Journal of molecular evolution **58**(1), 1–11 (2004)
11. Zuo, G., Hao, B.: Cvtree3 web server for whole-genome-based and alignment-free prokaryotic phylogeny and taxonomy. Genomics, proteomics & bioinformatics **13**(5), 321–331 (2015)
12. Shepp, L.: Normal functions of normal random variables. Siam Review **4**(3), 255 (1962)
13. Muegge, B.D., Kuczynski, J., Knights, D., Clemente, J.C., González, A., Fontana, L., Henrissat, B., Knight, R., Gordon, J.I.: Diet drives convergence in gut microbiome functions across mammalian phylogeny and within humans. Science **332**(6032), 970–974 (2011)
14. Rusch, D.B., Halpern, A.L., Sutton, G., Heidelberg, K.B., Williamson, S., Yooseph, S., Wu, D., Eisen, J.A., Hoffman, J.M., Remington, K., *et al.*: The sorcerer ii global ocean sampling expedition: northwest atlantic through eastern tropical pacific. PLoS biology **5**(3), 77 (2007)
15. Kurokawa, K., Itoh, T., Kuwahara, T., Oshima, K., Toh, H., Toyoda, A., Takami, H., Morita, H., Sharma, V.K., Srivastava, T.P., *et al.*: Comparative metagenomics revealed commonly enriched gene sets in human gut microbiomes. Dna Research **14**(4), 169–181 (2007)
16. van der Maaten, L.: Barnes-hut-sne. CoRR **abs/1301.3342** (2013). 1301.3342
17. Cheng, T.-Y., Sueoka, N.: Heterogeneity of dna in density and base composition. Science **141**(3586), 1194–1196 (1963). doi:10.1126/science.141.3586.1194. http://science.sciencemag.org/content/141/3586/1194.full.pdf
18. Wrighton, K.C., Thomas, B.C., Sharon, I., Miller, C.S., Castelle, C.J., VerBerkmoes, N.C., Wilkins, M.J., Hettich, R.L., Lipton, M.S., Williams, K.H., Long, P.E., Banfield, J.F.: Fermentation, hydrogen, and sulfur metabolism in multiple uncultivated bacterial phyla. Science **337**(6102), 1661–1665 (2012). doi:10.1126/science.1224041. http://science.sciencemag.org/content/337/6102/1661.full.pdf
19. Arumugam, M., Raes, J., Pelletier, E., Le Paslier, D., Yamada, T., Mende, D.R., Fernandes, G.R., 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.B., Nielsen, T., Pons, N., Poulain, J., Qin, J., Sicheritz-Ponten, T., Tims, S., Torrents, D., Ugarte, E., Zoetendal, E.G., Wang, J., Guarner, F., Pedersen, O., de Vos, W.M., Brunak, S., Doré, J., members), M.C.a., Weissenbach, J., Ehrlich, S.D., Bork, P.: Enterotypes of the human gut microbiome.

Nature **473**, 174 (2011). Article

20. Konstantinidis, K.T., Braff, J., Karl, D.M., DeLong, E.F.: Comparative metagenomic analysis of a microbial community residing at a depth of 4,000 meters at station aloha in the north pacific subtropical gyre. Applied and Environmental Microbiology **75**(16), 5345–5355 (2009). doi:10.1128/AEM.00473-09. http://aem.asm.org/content/75/16/5345.full.pdf+html