REPORT

Alignment-free tools for metagenomics-data analysis

Robert Deibel

Abstract

Metagenomics; as the study and analysis of microorganisms of biotopes, like the human gut, is a field of vast research where researchers have to deal with the giant sets of data gathered through NGS-methods. Since the amount of data results in stress on computation and time resources, the development of fast and light analysis tools is appreciated. In this report I introduce the two main branches of analysis tools, while setting the focus on alignment-free methods.

While the alignment-based approach has its foundation in the alignment of 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 and test selected methods on metagenomic data.

$D2z$, Hao and $N2$ are statistical approaches based on k-tupel count frequencies. Here I selected two tools and ran an analysis using the ASARI data set ....hier eine referenz.........

Laczny *et al.* used k-mers as vectors in high-dimensional space and the BH-SNE of van der Maaten[1] visualizing related data in two dimensional scatter plots, resulting in a tool with high accuracy for simulated as well as real-world metagenomes.

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 genomes, coding or noncoding, of a population of microorganisms in a microbiome sample. The DNA of organisms is isolated form these samples. As such metagenomics is the study and analysis of these metagenomes[2].
A microbiome consists of countless bacteria, archea and viruses; for which >90% are uncultureable, using sequencing and metagenomic analysis as a way to study these.

NGS – Next Generation Sequencing The sheer amount of metagenomic data – Kakirde *et al.*[3] states 10 Tb of DNA in a soil sample – resulted in advances of sequencing.
Nowadays new high throughput methods – also Next Generation Sequencing or NGS for short – are used to compute 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.

What do we want to achieve? Metagenomics is used in the design of antibiotics and medicine or to analyze the metabolism of microorganisms and its hosts; making it a rapidly developing field of research.
Here, I'm going to summarize two approaches to data analysis and showcase one of those in more detail.

The "classical" approach

Alignment-based methods are proven but don't include all possibilities The best known approach to analyze reads are the various alignment-based methods.
Sequences are aligned against a database of known genomes, the resulting profiles are analyzed based on several factors.
This approach is proven under various conditions and implemented numerous times; BLAST for example, while not used for metagenomics anymore, has an accuracy well over 80%[5] and similar values are expected

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

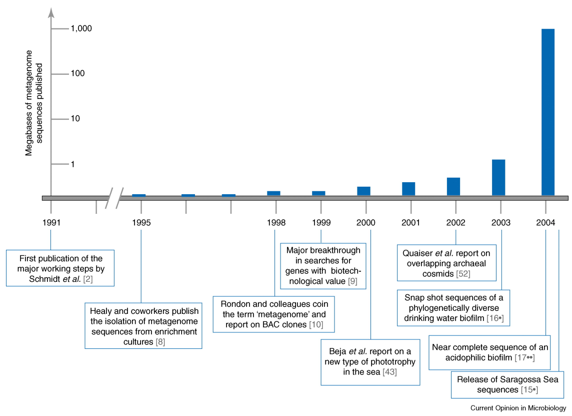

Figure 1 Timescale of metagenomic-derived and published DNA sequences. The timescale ranges from 1991, the initial outline of the major working steps, to the first mapping of archaeal comids in 2002 and the snap shot sequence analysis of the Sargasso Sea published earlier this year.–taken from Streit *et al.* [4] just for comparison of published DNA sequences – single events are not of importance for this

for other alignment-based tools.
However NGS supplies researchers with a lot of data to be analyzed. The analysis of metagenomes is computation heavy. BLAST – and therefor BLAST-like tools – align its queries with the entries in a chosen database – for 10Tb of data one can safely assume this step as time consuming – this results in the pursuit of faster and more effective methods for data analysis.
Research of novel data, not listed in databases, 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 analysis

Apart from alignment of sequences another way is basing the analysis on different factors associated with metagenomic data. For this report I reflect the work of Song *et al.* [6] and Laczny *et al.* [7] both presenting methods for the analysis of metagenomic-data using alignment-free approaches. Their work is based on statistical methods, and visualization and machine learning respectively.

Methods

k-tupels as a measure of similarity

In the work of Song *et al.*[6] different methods based on k-tupel occurrences in sequences are presented. Where a k-tupel is a substring of sequences with length k.
By counting the occurrences of these k-tupels and applying a distance or dissimilarity metric, the tupels are clustered and these clusters analyzed using current biological knowledge. A metric would be based on the resulting counted k-tupel frequencies.
The next section will focus on methods of k-tupel counts as a measure of similarity for sequences.

The D_2 statistic and normalization by $D2z$ Torney *et al.*[8] introduced D_2 using k-tupel matches between sequences to define the similarity of these.

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

where X_w and Y_w are the number of occurrences of string w in the corresponding sequence and $\mathcal{A}$ is the alphabet.
Kantrovitz *et al.*[9] stated that the D_2 statistic depends on the underlying sequence model and performed a normalization to remove the bias. 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 by Markov models for the used sequences.
$D2z$ was compared with five other measures of similarity – see [6][9] for details – through analysis of *cis*-regulatory modules (CRM), outperforming all of them. However $D2z$ requires two parameter. Apart from k, r has to be specified, where r is the order of the sequence Markov chain.

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

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

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.

$$Hao = \frac{1}{2}\left(1 - \frac{\sum_w \left(\frac{X_w - E_w^X}{E_w^X}\right)\left(\frac{Y_w - E_w^Y}{E_w^Y}\right)}{\sqrt{\sum_w \left(\frac{X_w - E_w^X}{E_w^X}\right)^2 \sum_w \left(\frac{Y_w - E_w^Y}{E_w^Y}\right)^2}}\right)$$

Notation is taken from Song *et al.*[6].
Hao calculates the frequencies of appearances of overlapping k-tupels indicated with X_w and subtracts a

random background using the $(k-2)$-th order Markov chain; this is to minimize the influence of random mutation. After computation of correlation – C – a normalization was defined by subtraction from 1 and multiplication with $\frac{1}{2}$.

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

C describing the cosine of the angle between the sequences, were $C = 1 \Leftrightarrow A = B$ 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 on each pair of sequences, neighbor joining then constructs a phylogenetic tree[10].

Nucleotide bias

Machine learning for alignment-free data analysis

In his work van der Maaten[1] introduced a machine learning variant – BH-SNE – based upon the idea that closely related objects have a larger influence upon each other than unrelated ones. While these objects were originally intended to be points in a picture, Laczny *et al.*[7] used reads of metagenomes.
Barnes-Hut-SNE applies the Barnes-Hut algorithm and metric trees to modify the t-SNE method, commonly used in machine learning

Barnes-Hut and vantage-point trees for faster computation The Barnes-Hut algorithm is often used by astronomers to perform N-body simulations.[1] . In this algorithm it is assumed that the force of objects with sufficient distance to one another is infinitesimal and thus can be ignored in further computation. Leading –in the case of BH-SNE – to a cut in objects to include in calculations.
For choosing these objects van der Maaten used vantage-point trees, where similar nodes are saved as the left, dissimilar nodes as the right child. After establishing the data structure one can search the tree and apply the given algorithm to the reduced set of nodes of interest

sequence signatures as objects Observations suggest the existence of species-specific oligonucleotide signature in genomic sequences[7][11]. 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[7].
For construction of these vectors a joint probability is assigned 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 and learned.
Using center log-ratio (CLR)-transformed – a normalization step – 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 1000 nt – they state that 600 nt might be an appropriate length for some applications, but with lower values the separation would drop remarkably as seen in (Figure 2) through lesser separation of the clusters – and 5-mers as oligonucleotide signatures, which produced better congruency compared to transformed and untransformed 4-mers.
For Laczny *et al.* these 5-mers are the before mentioned objects, used for calculation of similarity in BH-SNE.

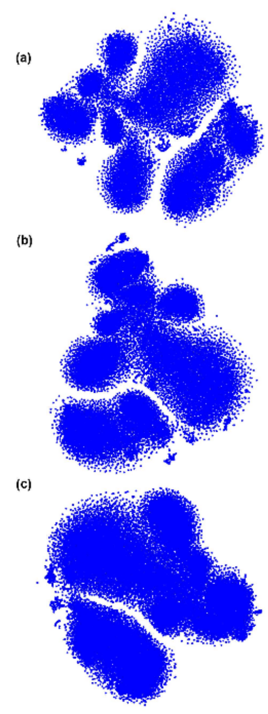

Figure 2 BH-SNE-based visualization of genomic fragment signatures for EqualSet01 (even community, overall reflecting distant taxonomic relatedness) with varying fragment lengths. (a) 800nt. (b) 600nt. (c) 400nt. – from Laczny *et al.*[7]

cluster finding 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.
The equally distributed data was used for reasons of simplicity, in real-world metadenomes, DNA is 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 seen 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. A selection of values is outlined in Table 1. The calculation was performed by enclosing clusters with polygons as seen in Figure 3. Points inside represent the positives, points outside the negatives. A similar output was achieved by fitting a (semi-)automated Gaussian Mixture model to calculate these values.

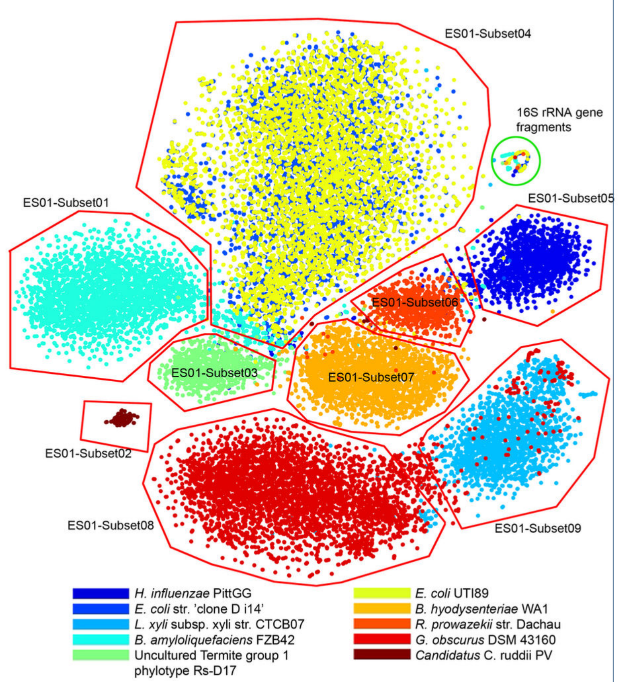

Figure 3 Figure taken from Laczny *et al.* [7] where red polygons mark clusters of congruent sets of interest used for calculation of sensitivity, specificity and accuracy. Colors mark different organisms as seen in the legend. The green polygon marks the cluster of 16s rRNA, forming a distinct group

application on real-world data sets With great results on simulated data, Laczny *et al.* also performed testing on real-world metagenomes of ground water[12], the human gut [13] and the deep sea[14]. They reported similar clustering (Figure 4) as seen in the simulated data with sensitivity, specificity and accuracy well above 90% for all subsets of the human gut data except for one, where accuracy was slightly below 80%. The values were calculated using polygons to mark clusters and verifying these by comparison with the ncbi non-redundant nucleotide database.

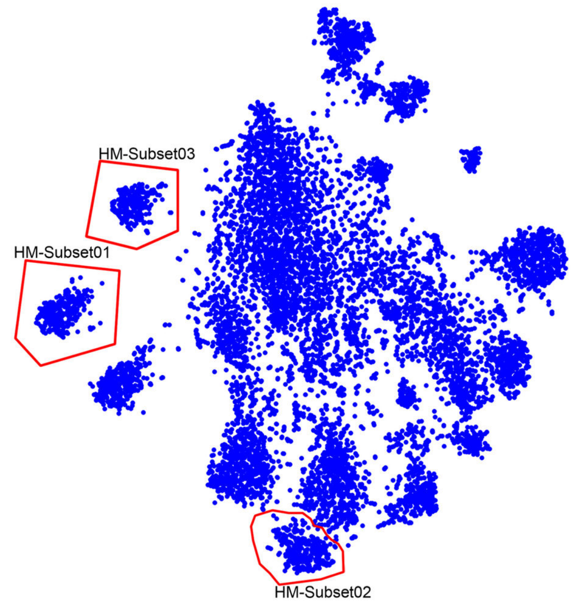

Figure 4 Clustering of human gut metagenome. Polygons represent clusters of great interest – from Laczny *et al.*[7]

The ground water metagenome also produced distinct clusters, as seen in Figure 5. Calculation of sensitivity, specificity and accuracy could not be done 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.

Compared to an ESOM-based approach, Laczny *et al.* reported better clustering of metagenomic-data while also significantly reducing runtime from around 3.8-fold to 50.4-fold[7] for their used sets and good visualization capabilities tested on simulated and real-world data.
Clustering seems robust in the applied data sets, while similar data tends to be near to each other in the visualization, 16S rRNA sequences form a distinct cluster,

Table 1 Sensitivity, specificity and accuracy of EqualSet01 – excerpt from Laczny *et al.*[7]

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*

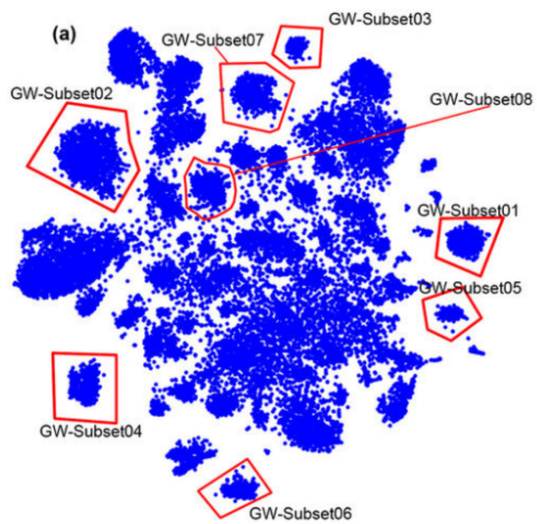

Figure 5 Clustering of groud water metagenome. Polygons represent clusters of great interest – from Laczny *et al.*[7]

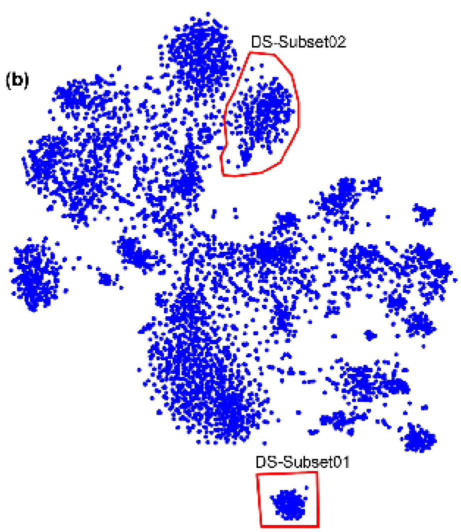

Figure 6 Clustering of deep sea metagenome. Polygons represent clusters of great interest – from Laczny *et al.*[7]

due to the high conservation of these regions in the genome.

As a downside sequences of 1000 nt were 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.

Results

Application of tools on data set

hier kommt was hin

References

1. van der Maaten, L.: Barnes-hut-sne. CoRR **abs/1301.3342** (2013). 1301.3342
2. Handelsman, J.: Metagenomics: application of genomics to uncultured microorganisms. Microbiology and molecular biology reviews **68**(4), 669–685 (2004)
3. 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
4. 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
5. 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. http://www.worldscientific.com/doi/abs/10.1142/9789814295291_0003
6. 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. /oup/backfile/content_public/journal/bib/15/3/10.1093/bib/bbt067/2/bbt067.pdf
7. 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
8. 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.
9. Kantorovitz, M.R., Robinson, G.E., Sinha, S.: A statistical method for alignment-free comparison of regulatory sequences. Bioinformatics **23**(13), 249–255 (2007)
10. 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)
11. 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
12. 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

13. 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
14. 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

Figures

Tables

Additional Files