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 want to introduce the two main branches of analysis tools, while setting the focus on alignment-free methods.

While the alignment-based approach are based on alignments – as seen with Smith-Waterman or BLAST – alignment-free methods, which are the main part of this report, have different approaches. Here I will showcase a selection of statistical and machine learning approaches and test these methods on a selected metagenomic data set.

TODO

Keywords: alignment-free; report; metagenome

Introduction

Metagenomics

A puddle of mud The metagenome is the whole set of genomes coding or noncoding of a population of microorganisms of a sample of a microbiome. The DNA of organisms is isolated form these samples. As such metagenomics is the study and analysis of these metagenomes[1].

A microbiome consists of countless bacteria, archea and viruses; for which >90% are unculturable, using sequencing and metagenomic analysis to study these.

NGS – Next Generation Sequencing The sheer amount of metagenomic data – Kakirde *et al.*[2] 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 handle this problem. 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.

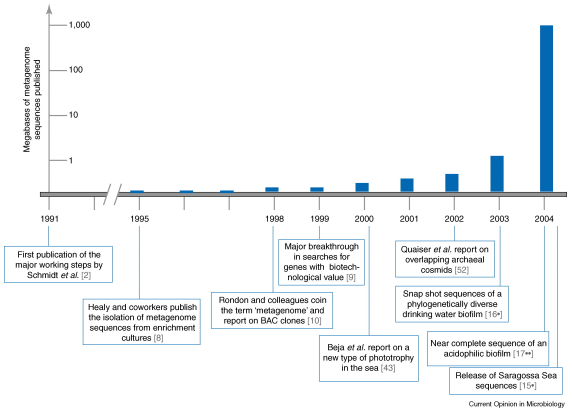

Figure 1 Timescale of metagenomic-derived and published DNA sequences. The timescale ranges from 1991, the initial outline of the major working steps, to the first mapping of archaeal comids in 2002 and the snap shot sequence analysis of the Sargasso Sea published earlier this year.–taken from Streit *et al.* [3] just for comparison of published DNA sequences – single events are not of importance for this

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

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 or example, while not used for metagenomics anymore, its accuracy is well over 80%[4] and similar values are expected 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 – aligns 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 a lightweight tool 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.* [5] and Laczny *et al.* [6] 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

Statistics

The idea behind hods like Smith-Waterman or BLAST are associated with this term; here I will report a methods discussed in Song *et al.*[5] using k-word counts as a statistical basis of analysis.
By counting the occurrences of these k-words – substrings of length k – applying a distance or dissimilarity metric based on the resulting k-word frequencies and clustered according to this metric and comparing the clusters with current biological knowledge.

The D_2 statistic Torney *et al.*[7] introduced the D_2 using k-word 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 w in the corresponding sequence and $\mathcal{A}$ is the alphabet.

Nucleotide bias

Machine learning for alignment-free data analysis

In his work van der Maaten[8] 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.*[6] used reads of metagenomes.

Barnes-Hut and vantage-point trees for faster computation The Barnes-Hut algorithm is often used by astronomers to perform N-body simulations.[8] . 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 nodes of interest resulting in a decrease of runtime.

sequence signatures as objects Observations suggest the existence of species-specific oligonucleotide signature in genomic sequences[6][9]. 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].
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. Using a Kullback-Leibler divergence and the optimizations stated before the points can be optimized.
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 greater separation of the clusters – and 5-mers as oligonucleotide signatures, which produced better congruency compared to transformed and untransformed 4-mers.

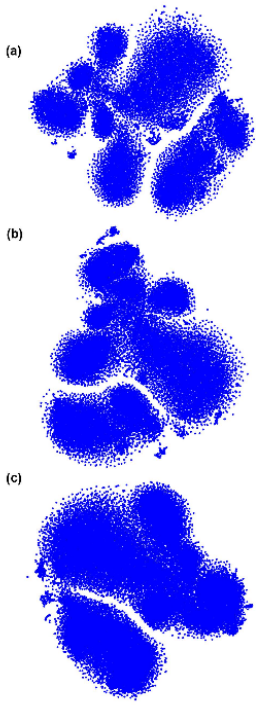

Figure 2 BH-SNE-based visualization of genomic fragment signatures for EqualSet01 (even community, overall reflecting distant taxonomic relatedness) with varying fragment lengths. (a) 800nt. (b) 600nt. (c) 400nt. – from Laczny *et al.*[6]

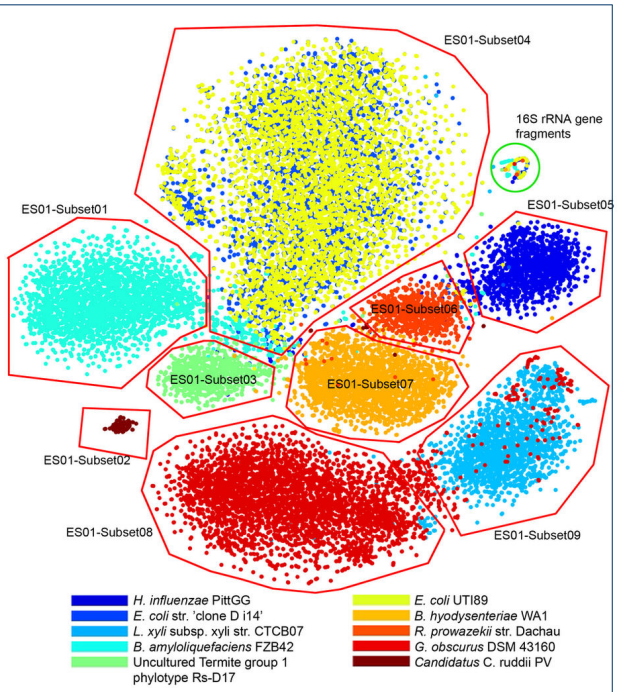

Figure 3 Figure taken from Laczny *et al.* [6] where red polygons mark clusters of congruent sets of interest used for calculation of sensitivity, specificity and accuracy. Colors mark different organisms as seen in the legend. The green polygon marks the cluster of 16s rRNA, forming a distinct group

cluster finding After applying their tool on simulated even (EqualSet01) and logarithmic (LogSet01) distributed data – sequences gathered from the real-world tend to be unevenly distributed, hence the logarithmic data set – and closely related data (EqualSet02), 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 distant relatives.
Overall the runs on simulated data resulted in high sensitivity, specificity and accuracy using human polygonal selection – as seen in figure 3 by the red polygons – and a similar output by fitting (semi-)automated Gaussian Mixture model to the data. A selection of values can be taken from table 1.

application on real-world data sets With great results on simulated data, Laczny *et al.* also performed testing on real-world data taken from ground water[10], the human gut [11] and the deep sea[12]. They reported similar clustering 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%.

Analysis of ground water data had to be carried out differently, since they reported a lack of independently characterized reference genomes. Instead they used what they called "essential genes" which can indicate the completeness of a genome and reported for four out of eight with over 80% percent completeness, indicating a positive result for their tool.
As for the marine sample, the clusters identified by the tool were linked to uncharacterized data. The analysis of this data, while should be carried out, was beyond the scope of their work, so neither sensitivity, specificity nor accuracy were computed.

conclusions of Laczny et al. The work of Laczny *et al.* introduces a new method for alignment-free data analysis which showed, compared to the state of the art ESOM-based approach, better clustering of metagenomic-data while also significantly reducing runtime from around 3.8-fold to 50.4-fold[6] 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 proximal to each other, 16 s rRNA sequences form an own distinct cluster, which could be due to the high conservation of these re-

Table 1 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*

gions in the genome or missing species specific oligonucleotide signatures

As a downside sequences of 1000 nt (or more) were needed to achieve good clustering, which are yet hard to gather through raw reads. Advances 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. 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. http://www.worldscientific.com/doi/abs/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. /oup/backfile/content_public/journal/bib/15/3/10.1093/bib/bbt067/2/bbt067.pdf
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. van der Maaten, L.: Barnes-hut-sne. CoRR **abs/1301.3342** (2013). 1301.3342
9. 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
10. 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
11. 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
12. 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