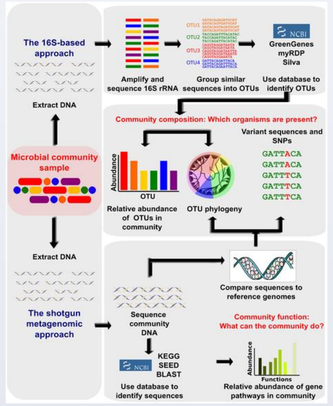
Creating an efficient compression based DNA sequence clustering algorithm for the analysis and comparison of metagenomes

Ashwin Sekar

Under the guidance of Huzefa Rangwala at George Mason University

Introduction

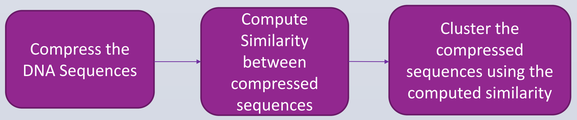
Metagenomics is the sequencing and analysis of genetic samples from the environment. Because these samples contain microorganisms never before studied or cultivated traditional analysis methods provide no insight.



The shotgun approach of using a database to identify sequences has been replaces with the 16S-based approach for this task. This approach uses databases to identify whole clusters rather than individual sequences, resulting in a higher success rate.

Goal

The limiting factor for Metagenomics is the clustering algorithm. Current clustering algorithms do not have the speed and accuracy needed to make Metagenomics a common laboratory practice. My researched focused on creating a 3 step process which would combine common data compression algorithms with existing clustering algorithms to speed up and maintain the accuracy of these algorithms.



The goal of the research was to refine this procedure so that it could be applied to any clustering algorithm and compression algorithm to improve the efficiency and accuracy of that clustering algorithm.

Experimentation

Using my approach I created 6 algorithms to test,

- MC-Lz4DM (Metagenome Clustering using Lz4 based Dissimilarity Measurements)
- MC-SeqDM (Metagenome Clustering using Sequitur based Dissimilarity Measurements)
- MC-Lz4CD (Metagenome Clustering using Lz4 based Compression Distances)
- MC-RLEDM (Metagenome Clustering using RLE based Dissimilarity Measurements)
- MC-SeqCD (Metagenome Clustering using Sequitur based Compression Distances)
- MC-RLECD (Metagenome Clustering using RLE based Compression Distances)

These algorithms were tested alongside the current leading algorithms:

- MC-MinH,
- MC-LSH,
- UCLUST,
- CD-HIT
- ESPRIT
- DOTUR
- Mothur

Preliminary evaluation stage:

Before doing clustering evaluation, each of these algorithms were tested for the accuracy of their similarity measure. The similarity measures were compared to the ground truth similarity between uncompressed sequences measured through direct comparison.

Final Evaluation:

First the algorithms were tested using a simulated data set. The simulated data contains 345,000 short sequences, generated from 43 known 16S rRNA gene fragments using the Roche GS20 system. Since this data set is simulated, an exact OTU number of 43 is expected. However since real life data samples have processing errors, data sets with 3% and 5% error were derived from the simulated data to mimic environmental conditions.

Next the algorithms were tested on an environmental data set containing eight seawater samples taken from a study by Sogin et. Al. Samples contain unequal length sequences with average sequence length of 60 characters.

Performance Metrics:

Number of OTU's – The number of clusters produced

Run Time – The time taken to complete clustering

Weighted Similarity – The average similarity within each cluster

Chao1 index – Overall richness of the species

ACE index – Richness of species, focused on clusters with less than 10 members

Shannon index – Measure of entropy found in the sample as a whole

Results

Preliminary evaluation using 112R data set from 16S environmental set of samples						
Algorithm	MC-SeqCD	MC-Lz4CD	MC-RLECD	MC-SeqDM	MC-Lz4DM	MC-RLEDM
Percent match with LSA similarity algorithm	2.06	91.2	84.9	86.3	86.2	.011

When the similarity measures were tested, MC-SeqCD and MC-RLEDM were deemed to be inaccurate in comparing compressed DNA sequences.

Algorithm	3% error # of clusters	5% error # of clusters
MC-Lz4DM	49	62
MC-Lz4CD	63	65
MC-SeqDM	220	212
MC-RLECD	240	232
MC-MinH	39	37
MC-LSH	47	41
UCLUST	91	53
CD-HIT	108	47
ESPRIT	180	86
DOTUR	210	135
Mothur	214	138

The remaining 4 algorithms were tested alongside the industry algorithms on the simulates data set with 43 OTU's.
MC-SeqDM and MC-RLECD were deemed to be inaccurate in judging the amount of clusters as they reported figures more than five times the actual value.

Algorithm	Metric	53R	55R	112R	115R	137	138
MC-Lz4DM	# Clu	1009	868	1419	1056	849	858
	W.Sim	99.00	98.69	98.72	99.03	99.21	98.90
	Time(s)	34	24.2	23.9	24.1	34.7	27.8
MC-Lz4CD	# Clu	922	777	1302	921	754	789
	W.Sim	98.64	98.26	98.50	98.52	98.66	98.66
	Time(s)	43.7	31.9	43.6	384.4	47.4	37.6
MC-MinH	# Clu	1165	1077	1634	1156	1020	1042
	W.Sim	96.90	92.45	91.18	93.33	95.86	93.10
	Time(s)	2.5	2.1	3.3	3.0	2.7	2.5
MC-LSH	#Clu	1172	1199	1795	1205	1041	1072
	W.Sim	96.90	93.12	91.33	93.50	95.86	93.10
	Time(s)	161.0	183.0	317.0	188.0	172.0	175.0
UCLUST	#Clu	1062	992	1561	1071	900	923
	W.Sim	96.67	91.67	91.02	93.33	93.50	92.82
	Time(s)	2.0	2.0	2.0	2.0	2.0	2.0
CD-HIT	#Clu	824	716	1196	820	712	725
	W.Sim	92.56	90.80	90.61	93.33	91.82	90.16
	Time(s)	3.6	3.1	3.9	3.8	3.2	3.1
ESPRIT	#Clu	940	859	1361	970	818	832
	W.Sim	93.12	91.35	90.88	93.33	91.82	90.16
	Time(s)	283.0	266.0	537.0	348.0	280.0	296.0
DOTUR	#Clu	1241	1258	1854	1279	1096	1121
	W.Sim	96.95	94.06	91.33	93.50	95.86	93.10
	Time(s)	5129.0	3511.0	5567.0	9237.0	6563.0	5618.0
Mothur	#Clu	1238	1256	1853	1278	1094	1119
	W.Sim	96.95	94.06	91.33	93.50	95.86	93.10
	Time(s)	10130.0	5940.0	12303.0	13501.0	12861.0	12310.0

The remaining two algorithms fit the criteria for the 3 step approach, so they were tested on an environmental data set. MC-Lz4DM outperforms all other algorithms in terms of weighted similarity with MC-Lz4CD in a close second. In terms of speed, MC-Lz4DM finishes fourth on average with MC-Lz4CD finishing fifth. Comparable speeds and an increase in efficiency indicates that these two algorithms are viable for clustering based on the standard metrics.

SID	Algorithm	Chao1 Index	Shannon Index	ACE Index
53R	MC-Lz4DM	1949.1	4.3	2140.1
	MC-Lz4CD	1951.1	4.1	2424.1
	MC-MinH	2276.3	4.4	2243.7
55R	MC-Lz4DM	1766.1	4.4	2304.4
	MC-Lz4CD	1619.6	4.1	2096.7
	MC-MinH	2182.8	4.6	2214.1
112R	MC-Lz4DM	3304.4	5.1	4479.5
	MC-Lz4CD	3060.4	5.1	4296.9
	MC-MinH	3931.3	5.3	4202.7
115R	MC-Lz4DM	2087.3	4.3	2822.3
	MC-Lz4CD	1999.9	4.1	2449.6
	MC-MinH	2411.4	4.6	2455.8
137	MC-Lz4DM	1583.0	4.6	2054.3
	MC-Lz4CD	1344.3	4.4	1813.2
	MC-MinH	1992.2	4.8	1800.1
138	MC-Lz4DM	1557.2	4.2	2159.1
	MC-Lz4CD	1305.1	4.1	1942.8
	MC-MinH	1713.8	4.4	1760.3

The diversity metrics of MC-Lz4DM and MC-Lz4CD in comparison to MC-MinH are shown above. On average, both of my algorithms underestimate the richness of the clusters through the Chao1 index, but accurately measure the richness through the ACE index.
Both of my algorithms match MC-MinH in entropy as demonstrated by the Shannon Index. In general, my algorithms create clusters with comparable entropy and richness to the current leading algorithm indicating that both of my algorithms are viable clustering methods in terms of species diversity metrics.

Conclusions

According to the hypothesis, since the algorithms that the compression based approach produced were concluded to be viable clustering algorithms, we have demonstrated the computational efficiency and accuracy of the compression based approach. Therefore by combining any compression algorithm, with appropriate similarity and clustering methods, we can create a viable solution.

The success of this compression based approach has the potential to produce algorithms with serious improvements in terms of efficiency and accuracy.

Future Work

For further testing, more algorithms created by the compression based approach need to be tested on a variety of data sets in order to assess consistency.

- Test with more samples and a greater diversity in DNA sequence length
- Use state of the art compression algorithms
- Use varying clustering algorithms

By reducing the time and increasing the efficiency of clustering results, metagenomic clustering can be implemented in common laboratory practices allowing the easy comparison of microorganisms. Metagenomic clustering can open doors into new biological research involving microorganisms, in fields such as disease diagnostics and treatment.

References

[1] Wooley JC, Godzik A, Friedberg I (2010) A Primer on Metagenomics. PLoS Comput Biol 6(2): e1000667. doi:10.1371/journal.pcbi.1000667

[2] Marco, D, ed. (2011). Metagenomics: Current Innovations and Future Trends. Caister Academic Press. ISBN 978-1-904455-87-5.

[3] Gene W. Tyson and Philip Hugenholtz. Metagenomics. Nature Reviews Microbiology, Sep 2008

[4] Morgan XC, Huttenhower C (2012) Chapter 12: Human Microbiome Analysis. PLoS Comput Biol 8(12): e1002808. doi:10.1371/journal.pcbi.1002808