Assignment of sequences to genomes and metagenomes

Contents:

http://gscompare.ehu.eus

Introduction

Oligonucleotide frequencies within genomes or DNA sequences have been often used to determinate relationships among DNA sequences. **The aim** of this study was to determinate the best method to compute frequencies and distances to correctly assign a DNA sequence to its precedence.

To achieve this goal, the assignment of sequences to genomes was searched, and clustering experiments with metagenomic data were performed.

Concerning the **assignment of sequences to genomes,** the basic experimental procedure was as follows: subsequences of completely sequenced prokaryotes were randomly selected, frequencies were computed, the frequencies of the subsequences and genomes were compared, and the percentage of correct assignment of subsequences to their genome of origin was computed. Based on this procedure, we did the following computing:

- As tetranucleotides are the most extensively searched oligonucleotides in the literature, we applied to them different methods to compute distances, and we determined the best method for assigning subsequences to the genome of origin.

- Next, we applied identical methods to oligonucleotides 2 to 7 bases long.

As prokaryotic genomes are not long enough to apply the same procedure to search longer oligonucleotides, we did two additional **clustering experiments with metagenomic** data:

- Capacity of oligonucleotides 4 to 10 bases long to correctly group metagenomic data.

- Effect of size of sample on distances.

Assignment of sequences to genomes (tetranucleotides)

For this study 1,102 completely sequenced prokaryotic genomes were used. All sequences were retrieved from NCBI (Tatusova *et al.*, 2014). In case the sequenced strain contained two or more chromosomes, only the main chromosome was included in the experiment. It must be pointed out that the list of genomes is enriched in prokariotes with medical or industrial importance, and for some generas and species different strains were available.

The complete list of prokaryotic genomes included in the experiment is available here.

FREQUENCIES AND DISTANCES

The frequencies and distances used in this experiment are described in-depth in a separate document, so they are mentioned here only to provide the abbreviation used in the tables and graphs shown in this document.

The following types of frequencies for **tetranucleotides** were computed by searching **both DNA strands**:

- Oligos4: Tetranucleotide frequencies
- Oligos4st: Standardized tetranucleotide frequencies
- ZOM: Zero'th Order Markov chain frequencies
- ZOMst: Standardized ZOM frequencies
- FOM: First Order Markov chain frequencies
- FOMst: Standardized FOM frequencies
- SOM: Second Order Markov chain frequencies
- SOMst: Standardized FOM frequencies
- Zscore: Z-scores values for the tetranucleotides
- Zscorest: Standardized z-scores for the tetranucleotides

NOTE: FOM and SOM frequencies are standardized frequencies for long random DNA sequences, but they are not standardized frequencies when they are computed from prokaryotic genomes (even though they are very close to standardized frequencies). And they are not at all standardized frequencies for short DNA fragments. Standardized ZOM frequencies are similar but not equal to Tetranucleotide Usage Deviations (TUD) used by Pride *et al*. (2006). To the best of our knowledge, standardized z-scores have not been used in the literature, but we computed them in order to be able to apply Genomic Signature Distance and Euclidean distance to these kinds of frequencies.

To compute distances among tetranucleotide frequencies the following statistics were used:

- Pd: Standard Pearson's distance.
- wPd: Weighted Pearson's distance
- GS: Genomic Signature distance
- E: Euclidean Distance

METHOD

An schematic representation of the method used in the experiment is shown in figure 1.

From each genome, 100 random positions were randomly selected, and from each position, subsequences of 250, 500, 750, 1000, 1250, 2500, 5000, 10000, 20000 and 40000 bp were obtained.

After computing tetranucleotide frequencies above for all subsequences and all genomes we proceeded as follows:

- The distances among the frequency of the subsequence and the same type of frequency of the 1102 genomes was computed.
- The distances among the subsequence and the genomes were sorted from lowest to highest.
- The position in the list of the genome to which the subsequence under study belonged to was recorded. For example, if the DNA fragment was a subsequence of the *E. coli* K12 genome and this genome was in position 6 within the list of sorted distances, value 6 was recorded.

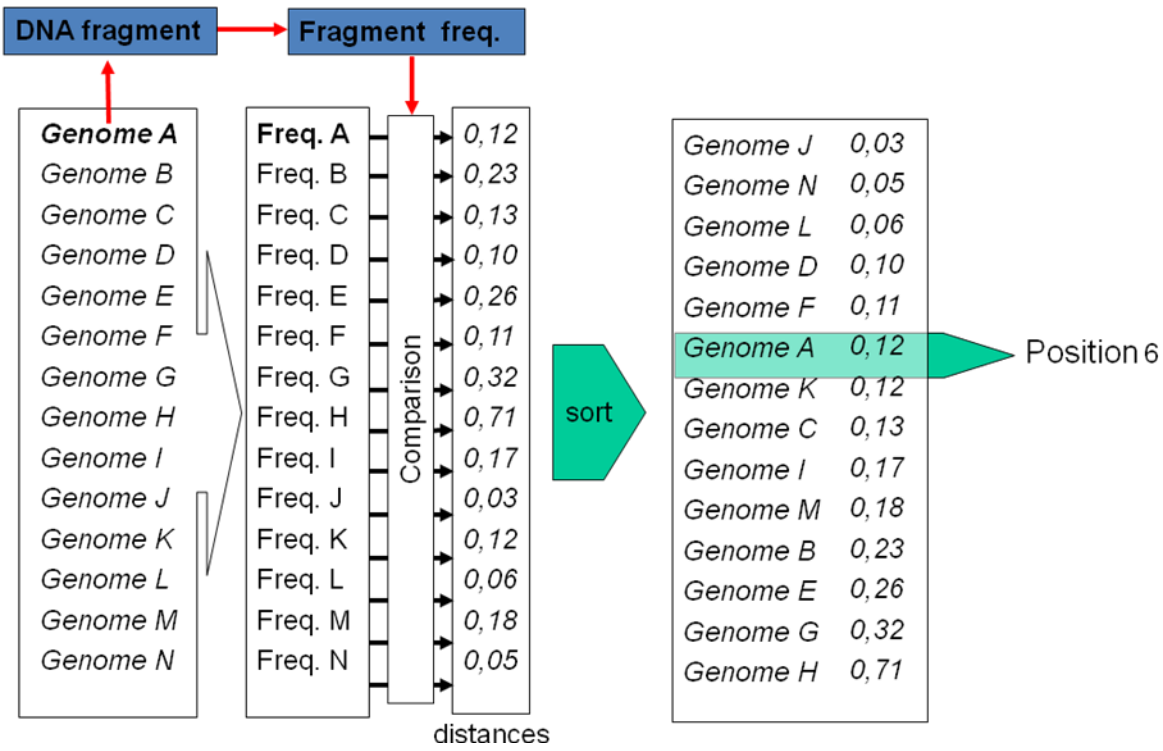


Figure 1: Representation of the method used in this experiment. Genomes A to N represent the 1,102 genomes used in the experiment. In the example, a DNA fragment from Genome A is extracted, and oligonucleotide frequencies of the fragment and of all genomes are computed. Then, oligonucleotide frequencies of the DNA fragment and all frequencies of genomes are compared and distances are obtained. Finally, genomes are sorted based on distances, and the position of genome A in the list is recorded.

- This procedure was applied to all randomly selected subsequences of a given length (10 different lengths) from all searched genomes (1102 genomes).
- For each DNA fragment length the percentage of times the genome to which the DNA belonged was in the top 5, 10, and 20 positions was recorded. For example, for 40,000 bp fragments (100 fragments x 1102 genomes), the genome from which the fragment was obtained was located in the top 5 positions of the sorted lists of distances 93,484 times (out of 110,200 lists), and 101,317 times and 105,767 times in the top 10 and 20 positions respectively.
- A table with all data generated with this procedure and their graphical representation was generated for the evaluation of the results.

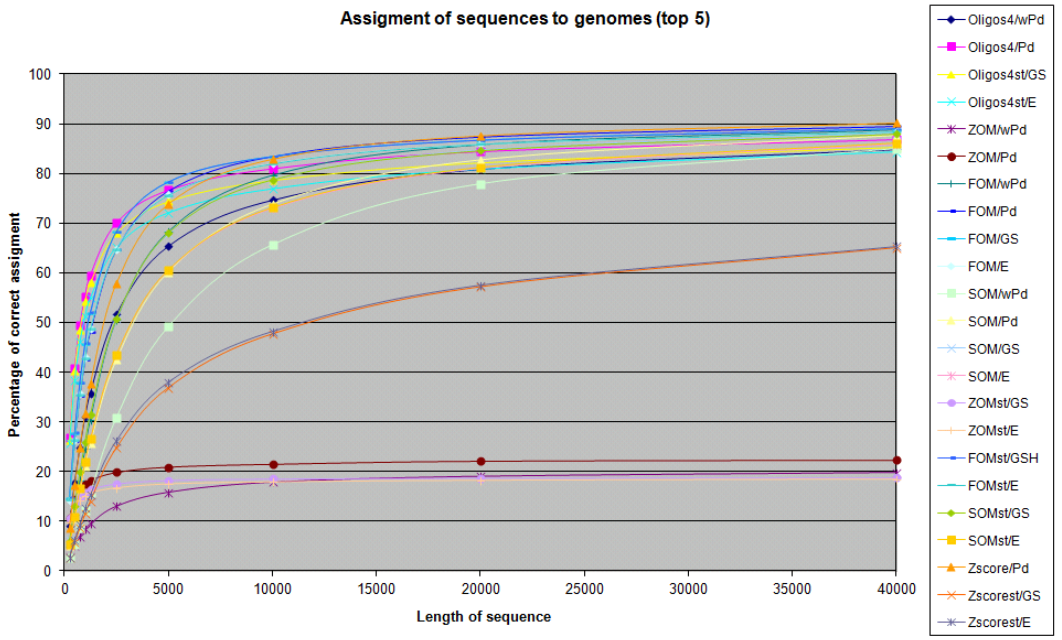
Identical procedure was applied to search the assignment of DNA fragments to genera. In this case the position of the first member of the genera from which the DNA fragment was obtained was recorded.

RESULTS

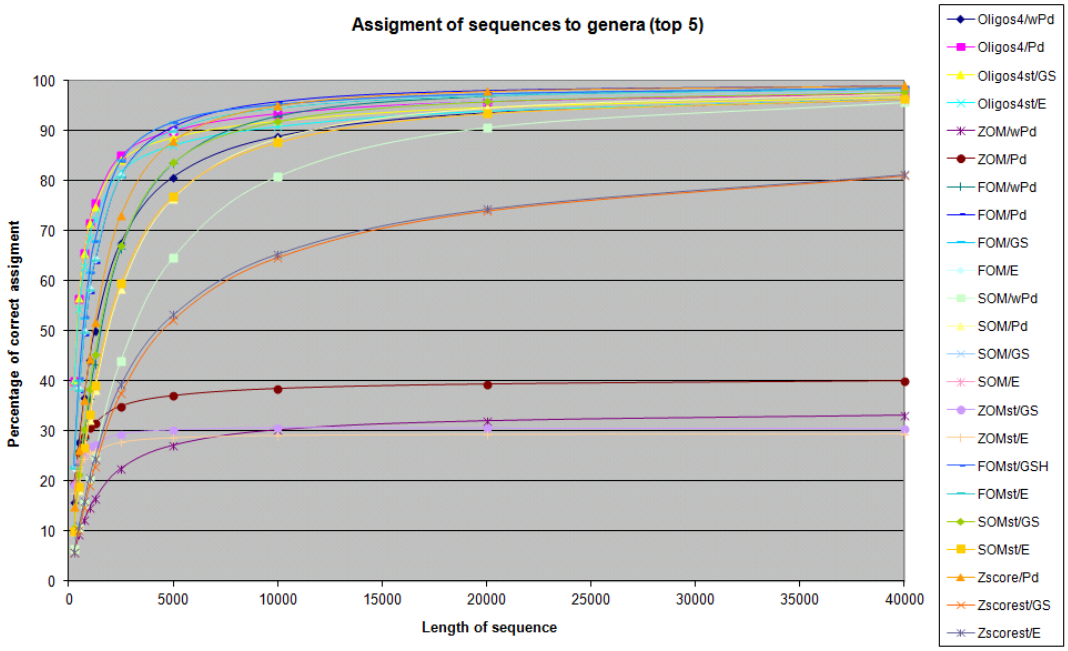
Performance of different methods to compute tetranucleotide frequencies and distances to correctly assign a DNA sequence to its genome (Appendix table 1) or genera (Appendix table 2) of precedence by using the method described above was searched. The tables show the frequency and distance combinations used to perform the experiment, and for each DNA fragment length, the percentage of times the genome or genera to which the DNA belonged was in the top 5, 10, and 20 positions. The following combinations are not shown in tables, although they were computed, because they were totally useless: Zscore/wPd, Zscorest/Pd and Zscorest/wPd.

Graphical representation of data in the tables is shown in Figure 1.

Figure 2: The graphs below show the capacity to assign correctly a DNA sequence to its genome of precedence (1a,1c,1e) or to a member of the same genera (1b,1d,1f) of the different combinations of methods to compute tetraoligonucleotide frequencies and distances.

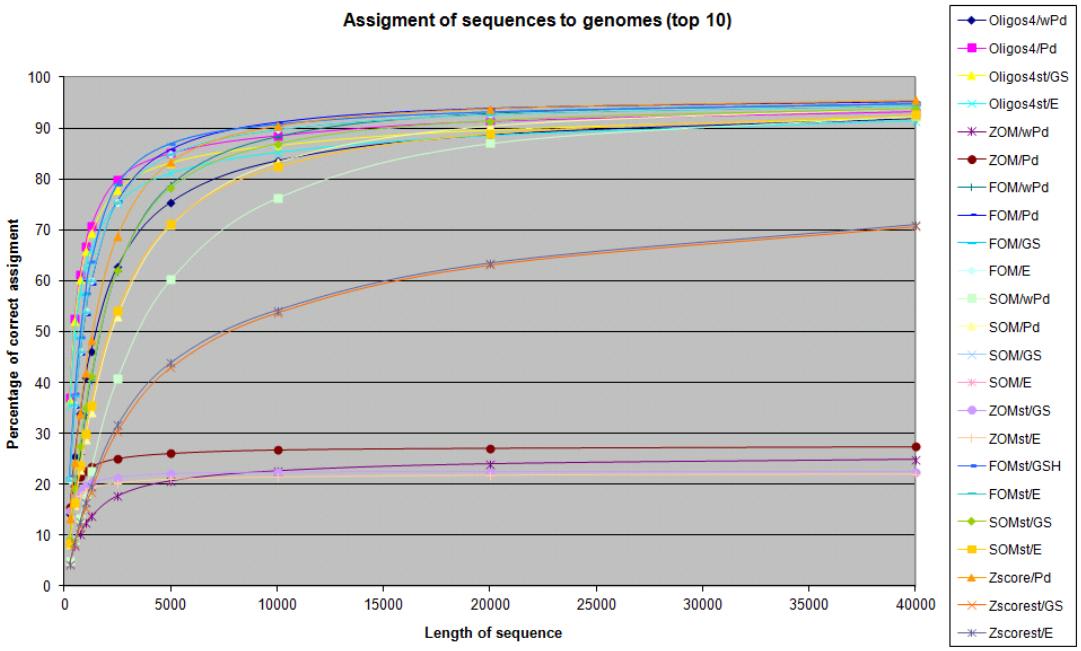


1a

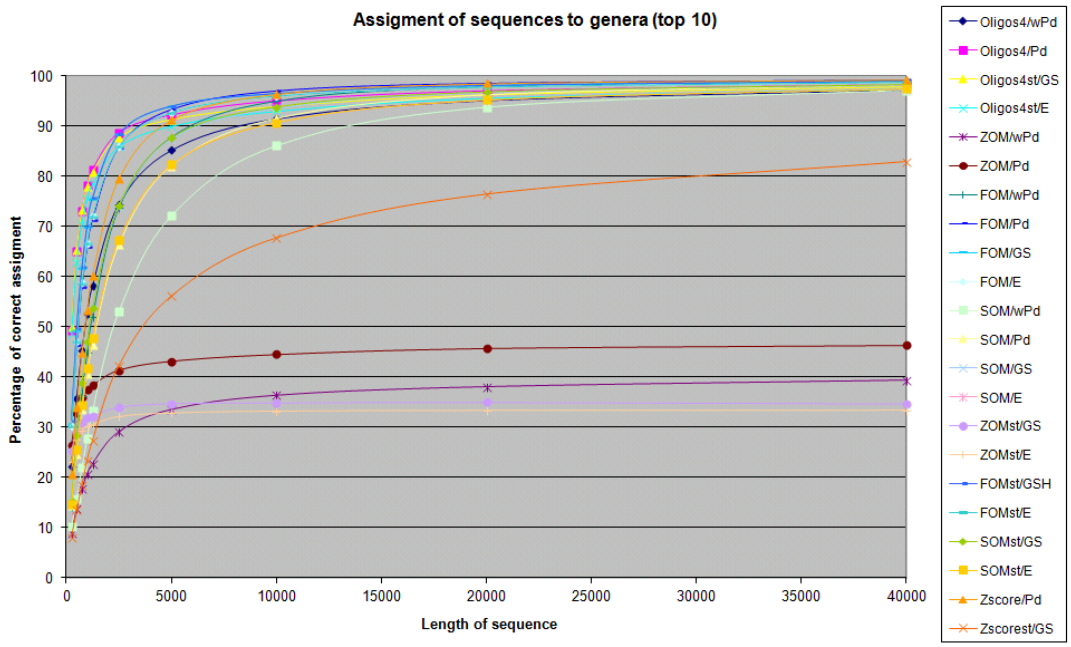


1b

Figure 1: (continuation)

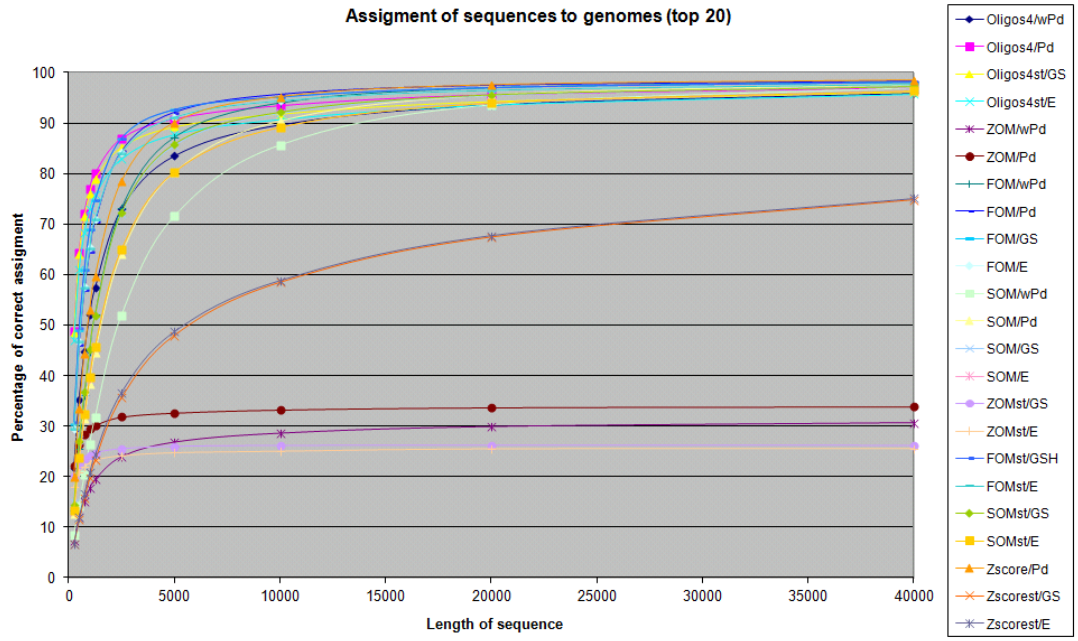


1c

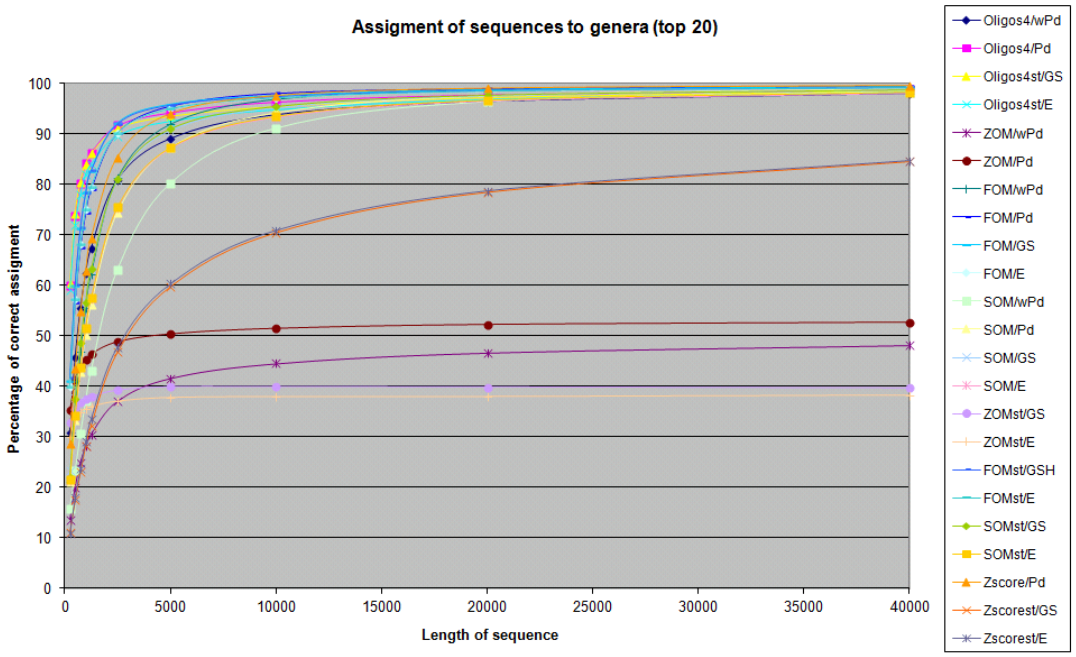


1d

Figure 1: (continuation)



1e



1f

The **best performing combinations of frequencies and distances** for assignment of DNA fragments to its genome (Table 1) or genera (Table 2) of precedence are shown in the tables below.

Table 1: Best performing combinations of frequencies and distances for assignment of DNA fragments to its genome.

Length of fragments (bp)	The best performing combination of frequencies and distances
250 500 750 1,000	Standardized oligonucleotide frequencies (Oligos4st) / Genomic Signature Distance (GS)
1,250 2,500	Oligonucleotide frequencies (Oligos4) / Pearson's distance (Pd)
5,000	First Order Markov chain frequencies (FOM) / Genomic Signature Distance (GS)
10,000 20,000	First Order Markov chain frequencies (FOM) / Pearson's distance (Pd)
40,000	Z-scores values (Zscore) / Pearson's distance (Pd)

Table 2: Best performing combinations of frequencies and distances for assignment of DNA fragments to their genera.

Length of fragments (bp)	The best performing combination of frequencies and distances
250 500 750 1,000 1,250 2,500	Oligonucleotide frequencies (Oligos4) / Pearson's distance (Pd)
5,000	First Order Markov chain frequencies (FOM) / Genomic Signature Distance (GS)
10,000	First Order Markov chain frequencies (FOM) / Pearson's distance (Pd)
20,000 40,000	Z-scores values (Zscore) / Pearson's distance (Pd)

CONCLUSIONS

The performance is better when assignment to genera is searched than for assignment to genome. At this point it is important to mention that for some prokaryotic species included in this experiment more than one genome was completely sequenced. As they were all included in the experiment, it is expected the assignment of a DNA fragment to its genera to be poorer than to the corresponding genera. A simple example may illustrate this happening: the distance between a DNA fragment obtained from *E. coli* K12 genome and the complete genome is similar to the distances among the fragment and the additional 30 *E. coli* genomes included in the experiment. As a consequence, after sorting the distances among the DNA fragment and all searched genomes, it is hard for *E. coli* K12 genome to be in the top 5, 10 or 20 positions, while it is more probable for genomes which are the only representatives of their genera.

The following conclusions were also obtained:

- Regardless of the type of frequency, the performance of Pearson's distance was always better than the performance of Weighted Pearson's distance.
- The performance of First Order Markov chain frequencies (FOM) and Second Order Markov chain frequencies (SOM) and corresponding standardized frequencies (FOMst and SOMst) provided basically the same performance.
- Usage of Zero'th Order Markov chain frequencies (ZOM) or the standardized ones (ZOMst) yielded very poor results.
- Z-scores values for tetranucleotides must be compared with normal Pearson's distance. The performance of other approximations based in Z-scores values was poor or useless.
- It will not be a surprise to get similar conclusions in studies comparing long DNA sequences regardless of the computing approach due to the similar performance of some computing methods. On the contrary, the selection of the computing method may be critical for shorter sequences due to the poorer performance of all computing methods for those sequences.

Assignment of sequences to genomes (di- to heptanucleotides)

The computation was limited to oligonucloetide frequencies and the standardized oligonucloetide frequencies. Other frequencies used to compute tetranucleotide frequencies were not applicable to all di- to heptanucleotides.

The procedure was similar to the one described above for tetranucletides. For the study 1,124 completely sequenced prokaryotic genomes were used.

FREQUENCIES AND DISTANCES

The di- to heptanucleotide frequencies and corresponding standardized frequencies were computed by searching both DNA strands.

To compute distances among oligonucleotide frequencies the following statistics were used:

- Pd: Standard Pearson's distance (for non-standardized frequencies).
- wPd: Weighted Pearson's distance (for non-standardized frequencies).
- Genomic Signature Distance (GS) (for standardized frequencies).
- E: Euclidean Distance (for standardized frequencies).

METHOD

An identical procedure to the one described above for tetranucleotides was applied. From each genome, 100 positions were randomly selected, and from each position, subsequences of 250, 500, 750, 1000, 1250, 2500, 5000, 10000, 20000 and 40000 bp were obtained and compared to all the genomes in the experiment. Only the percentage of times the genome to which the DNA subsequence belonged was in the top 5 positions was recorded.

RESULTS:

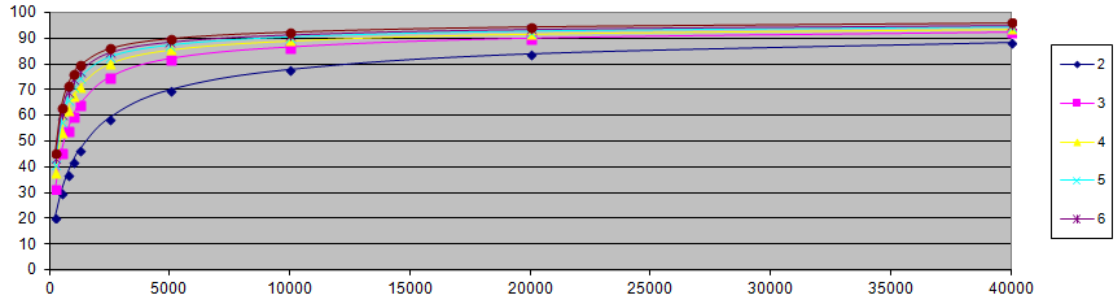
- The combination of oligonucleotide frequencies and Pearson distance yielded the best results in all cases except for very short oligonucleotides.
- The quality of results from Pearson distance, Genomic Signature Distance and Euclidean distance was similar, but the result from Weighted Pearson distance was very poor.

Table 3: Performance of different methods based in di- to heptanucleotides to correctly assign a DNA subsequence to its genome of origin. The table shows the percentage of times the genome to which the DNA belonged was in the top 5 positions, and in yellow, the better performing method to compute distances for each subsequence length.

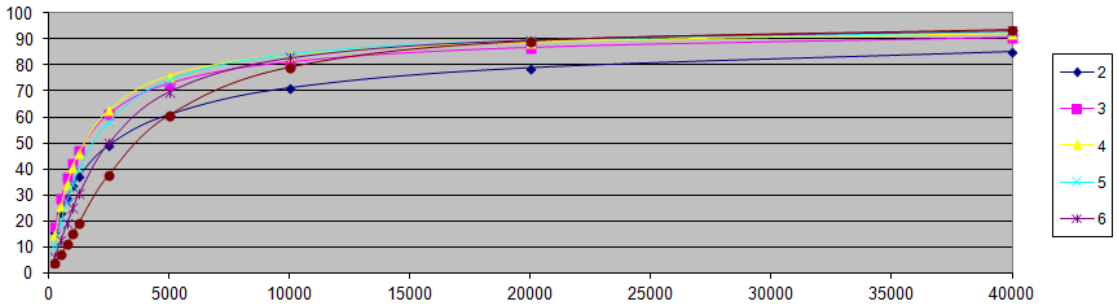
	Length of randomly selected sequence form the genomes									
	250	500	750	1000	1250	2500	5000	10000	20000	40000
Length=2										
Pearson	20,65	30,28	37,32	42,35	46,57	59,04	69,88	77,81	83,88	88,31
Wpearson	15,82	23,36	29,12	33,74	37,4	49,23	60,9	71,14	78,86	85,04
GSignatue	20,76	29,7	36,23	41,01	44,64	55,55	65,38	73,02	79,12	84,32
Euclidean	21,44	30,21	36,52	41,11	44,87	55,48	65,25	72,82	78,97	84,15
Length=3										
Pearson	31,55	45,63	54,18	59,8	64,22	74,78	81,97	86,44	89,88	92,2
Wpearson	18,06	28,81	36,7	42,46	47,23	61,32	72,94	81,13	86,64	90,51
GSignatue	30,62	44,1	52,04	57,45	61,33	71,41	78,64	83,41	87,22	90,47
Euclidean	30,63	43,2	50,81	55,89	59,61	69,61	77,19	82,21	86,38	89,93
Length=4										
Pearson	37,53	53,2	61,82	67,21	71,12	80,02	85,39	88,9	91,58	93,42
Wpearson	14,86	25,97	34,36	41,07	46,65	62,79	75,85	83,98	88,84	92,12
GSignatue	37,23	52,62	60,95	66,16	69,77	78,1	83,51	86,93	89,87	92,26
Euclidean	36,09	50,24	58,01	62,97	66,5	75,26	81,42	85,42	88,82	91,54
Length=5										
Pearson	41	57,38	65,92	71,13	74,69	82,54	87,13	90,29	92,54	94,17
Wpearson	10,24	19,69	27,78	34,31	40,31	58,3	74,37	84,18	89,58	92,71
GSignatue	30,31	55,58	65,39	70,74	74,08	81,53	85,86	88,74	91,34	93,31
Euclidean	39,51	54,3	62,05	66,92	70,25	78,22	83,61	87,14	90,15	92,55
Length=6										
Pearson	43,44	60,22	68,79	73,81	77,1	84,36	88,5	91,33	93,48	95,01
Wpearson	6,33	12,67	18,99	25,01	30,46	50,2	69,74	82,74	89,67	93,26
GSignatue	9,05	22,2	38,34	53,66	64,95	82,18	87,17	89,88	92,38	94,23
Euclidean	41,39	57,06	64,88	69,64	72,84	80,16	85,01	88,21	91,12	93,33
Length=7										
Pearson	45,58	63,02	71,52	76,27	79,47	86,05	89,97	92,55	94,59	96,11
Wpearson	4,14	7,6	11,4	15,57	19,7	38,08	60,79	79,09	89,13	93,49
GSignatue	1,74	3,97	6,72	9,89	13,79	42,42	81,69	91,18	94,01	95,59
Euclidean	39,57	55,88	64,43	69,71	73,17	81,1	86,2	89,51	92,38	94,49

Figure 3: Graphical representation of data in table 3 above.

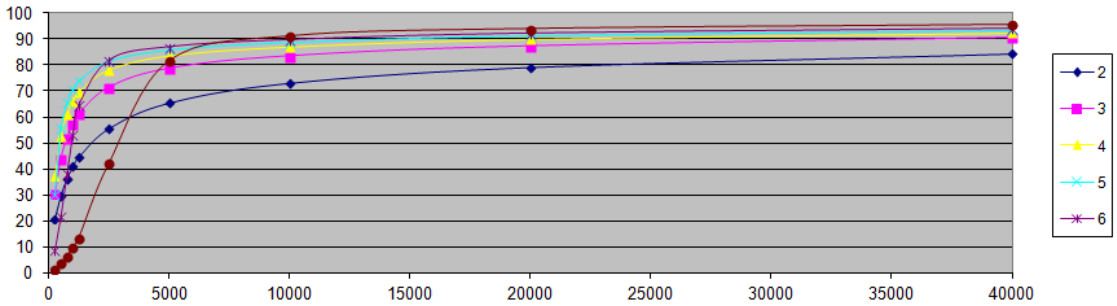
Pearson distance for di- to hepta nucleotides



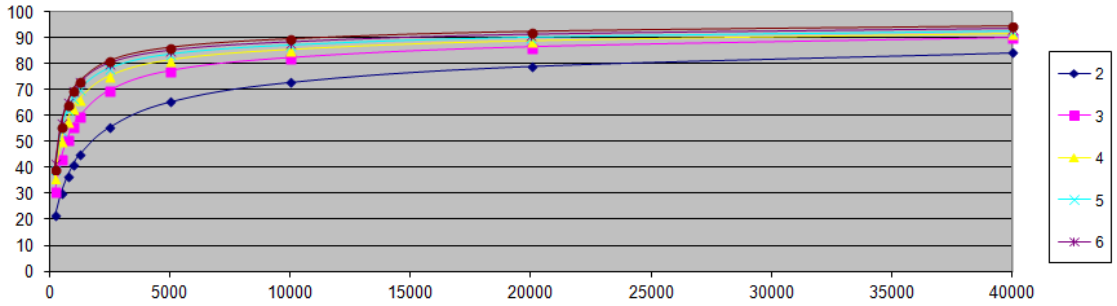
Weighted Pearson distance for di- to hepta nucleotides



Genomic Signature distance for di- to hepta nucleotides



Euclidean distance for di- to hepta nucleotides



Capacity of oligonucleotides 4 to 10 bases long to correctly group metagenomic data

Searching for the presence of longer oligonucleotides (over 8 mer) in genomes becomes a problem due to the increased information that must be manipulated, but also due to the very low number of occurrences of each oligonucleotide in a genome. Just for reference, the average occurrence of a 10 mer oligonucleotide is one in 1,048,576 bp ($1/4^{10}$). When occurrences of each oligonucleotide are very low the statistical procedures mentioned in this document are unable to compute distances correctly.

Even so, we wanted to evaluate whether longer oligonucleotides were useful to discern between samples, so we searched samples with a higher amount of information: metagenomic data.

METAGENOMIC DATA AND HYPOTHESIS

Metagenomic data from Monterey Bay coastal microbial picoplankton (Rich *et al*, 2011) obtained from the Camera database (Sun *et al*, 2011) were used in the experiment. In this work three California costal samples were pyrosequenced: one was obtained previous to a phytoplankton bloom that happens annually in the area (pre-bloom sample), and two after that bloom (post-bloom samples). In the Camera database six sets of reads were available, and according to the authors (personal communication form Edward F. De Long), each pair of reads' sets correspond to the same DNA samples which was sequenced separately but in the same sequencing slide.

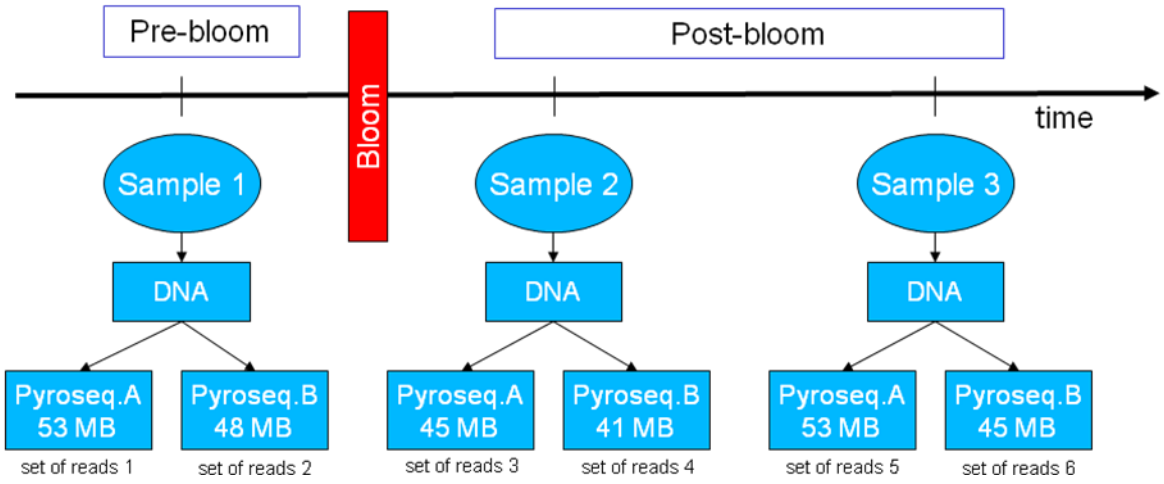


Figure 4: Relationship among three samples from Monterey Bay and the six sets of reads obtained by pyrosequencing by Rich *et al*. (2011)

By searching oligonucleotide composition of the six sets of reads we realised **that longer oligonucleotides allow us to discern between the two sets of reads obtained from the same sample**. To establish that we followed the next procedure (figure 5 below): two subsets of reads from each set of reads were extracted, oligonucleotide frequencies were computed for each subset, distances between subsets were computed and UPGMA clustering was performed. The dendrogram that was generated when

searching longer oligucleotides grouped together the two subsets of reads obtained from each set of reads, and in a second clustering step the two sets of reads from the same sample were grouped. The final clustering steps related sample 1 with sample 3 (as pointed out by Rich *et al*. [2011] as the most closely related samples), and then sample 2 with the previous ones.

We used the following hypothesis: a hypothetically correct dendrogram will cluster the subsets of reads as described in figure 5 below.

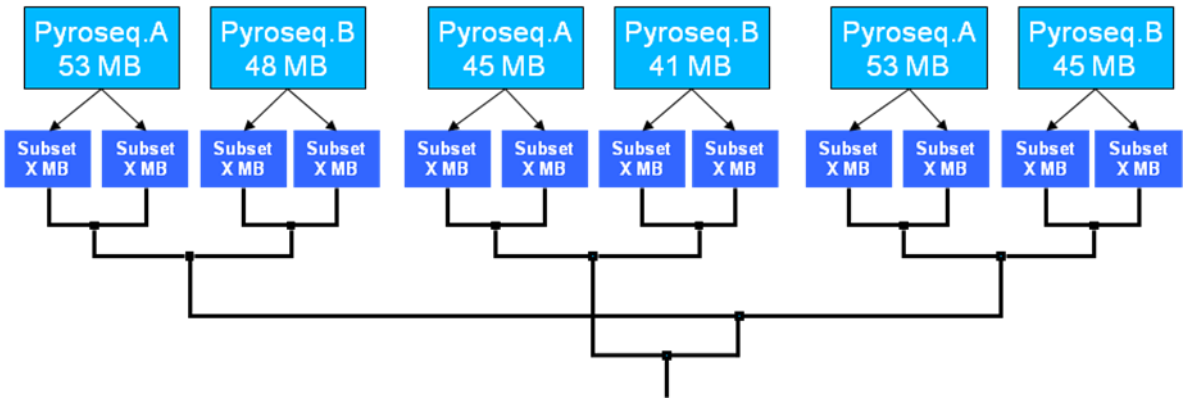


Figure 5: Hypothetically correct dendrogram. See text for more details. This type of grouping was observed in many clustering experiments obtained by searching the clustering capacity of long oligonucleotides based comparison of subsets of reads.

SUBSETS OF SAMPLES, FREQUENCIES AND DISTANCES

From each set of reads from Monterey Bay, a random selection of 10 subsets of reads totalling 1 to 10 MB were extracted and duplicated (10 lengths x 6 sets of reads x 2 duplicates). This random selection of reads was repeated 100 times per set of reads.

Tetra- to decanucleotide frequencies and corresponding standardized frequencies were computed for each of the subsets of reads and distances among oligonucleotide frequencies were computed using the following statistical methods:

- Pd: Standard Pearson's distance (for non-standardized frequencies).
- wPd: Weighted Pearson's distance (for non-standardized frequencies).
- GS: Genomic Signature distance (for standardized frequencies).
- E: Euclidean Distance (for standardized frequencies).

Once the distances were computed, UPGMA clustering was applied and the number of times the hypothetical clustering pattern matched the results was counted. The number of marches is shown in the tables below.

Table 4. For a particular experiment with 12 subsets of reads of a specific size (1 to 10 MB) and a specific oligonucleotide length (4 to 10 bases), the number of times the statistical method used in the experiment matched the hypothetical correct dendrogram shown in figure 5 (see text for details) is shown. A total of 100 experiments were performed for each situation. A value of zero means none of the 100 experiments matched the expected dendrogram, while a value of 100 means all dendrogams showed the expected profile.

Pearson + UPGMA		Size of subset of reads									
		1 MB	2 MB	3 MB	4 MB	5 MB	6 MB	7 MB	8 MB	9 MB	10 MB
Oligonucleotide length	4	1	0	2	2	9	18	19	26	37	47
	5	1	1	6	12	21	49	52	61	70	86
	6	0	7	35	51	80	94	94	100	100	100
	7	3	41	75	97	100	100	100	100	100	100
	8	6	81	100	100	100	100	100	100	100	100
	9	0	1	92	100	100	100	100	100	100	100
	10	0	0	0	0	0	14	80	99	100	100

W. Pearson + UPGMA		Size of subset of reads									
		1 MB	2 MB	3 MB	4 MB	5 MB	6 MB	7 MB	8 MB	9 MB	10 MB
Oligonucleotide length	4	1	0	2	0	5	7	3	8	15	19
	5	0	0	3	2	10	17	15	21	22	37
	6	0	2	6	7	19	32	48	38	49	64
	7	0	1	4	4	12	9	17	19	21	22
	8	0	0	3	2	2	1	0	0	0	1
	9	0	0	1	2	0	1	0	1	1	0
	10	0	0	0	3	10	7	6	6	9	5

GS + UPGMA		Size of subset of reads									
		1 MB	2 MB	3 MB	4 MB	5 MB	6 MB	7 MB	8 MB	9 MB	10 MB
Oligonucleotide length	4	0	0	1	0	5	15	15	18	27	28
	5	0	0	3	9	18	30	36	50	56	75
	6	1	8	25	44	73	86	91	97	100	100
	7	5	52	90	98	99	100	100	100	100	100
	8	54	99	100	100	100	100	100	100	100	100
	9	97	100	100	100	100	100	100	100	100	100
	10	93	100	100	100	100	100	100	100	100	100

Euclidean + UPGMA		Size of subset of reads									
		1 MB	2 MB	3 MB	4 MB	5 MB	6 MB	7 MB	8 MB	9 MB	10 MB
Oligonucleotide length	4	1	1	1	1	8	11	18	15	18	24
	5	1	1	5	8	16	26	30	36	47	64
	6	1	5	26	32	62	70	75	89	94	95
	7	6	28	70	88	94	99	100	100	100	100
	8	16	83	98	95	99	99	98	100	100	100
	9	51	73	85	80	84	85	88	82	93	90
	10	46	54	48	34	38	34	30	26	24	27

RESULTS:

- Genomic Signature distance and Pearson distance yielded the best results. The performance of Genomic Signature distance was better for longer oligonucleotides and for smaller samples. For shorter oligonucleotides and longer samples, Pearson distance was better.
- Results obtained by applying Euclidean distance were poor, and Weighted Pearson distance yielded the poorest results.
- The poorer performance of Pearson distance when comparing smaller sets of reads is probably due to an increased presence of oligonucleotides showing non-occurrences when fewer genomic information is available. In these situations computing Pearson distance (which is based in Pearson`s correlation) is not a good statistical procedure. This is the same reason for the poor performance of Pearson distance for long oligonucleotides.

Effect of size of sample on distance

As pointed out previously in this document, the oligonucleotide-based strategy was able to discern among sets of reads obtained from the same metagenomic sample at the Monterey Bay (for descriptions of the samples check the previous experiment). In this experiment we wanted to determine whether the size of the genomic information affects the distance at which the sets of reads from the same sample are grouped.

SETS OF SAMPLES, FREQUENCIES AND DISTANCES

Six sets of reads from three metagenomic samples (two set of reads per sample) were available in the Monterey Bay experiment. From each set of reads a random selection of 3 subsets of reads totalling 2, 5 and 10 MB were extracted, and standardized octanucleotide frequencies were computed. The same types of frequencies were also computed for the six complete sets of reads, so in total 6 x 4 standardized frequencies were obtained.

Then the frequencies were compared in two ways:

- Oligonucleotide frequencies from samples of the same size (frequencies of the complete set of reads or of 2, 5 and 10 MB long subsets of reads) were compared to each other (Genomic Signature distances) and clustered (UPGMA). The dendrograms generated in this experiment are shown in figure 6.
- Frequencies from all samples were compared (Genomic Signature distances) and clustered (UPGMA). The dendrograms generated in this experiment are shown in figure 7.

RESULTS

In the first comparison experiment (Figure 6), as expected, the two sets of reads per sample were grouped together, and higher level clustering was also as expected (see previous experiment for details). In the figure the distance at which the two sets of reads per sample were grouped is marked with a red rectangular area. Those distances were identical when samples of reads of the same size were searched, but an increase of the distance at which the reads were grouped was observed when the sample size was smaller.

In the second comparison experiment (Figure 7), all sets of reads obtained from pre-bloom samples were clustered together regardless of the size of the set of reads, and this also happened for both post-bloom samples.

The clustering behaviour of samples in those experiments is not a surprise. The smaller the size of the sets of reads, the less representative the computed frequencies are for long oligonucletides. As a consequence, the computed distances are also higher. This influenced the clustering in both experiments.

Although only results obtained with Genomic Signature distances are shown in this document, this behaviour was identical when other types of distances were computed.

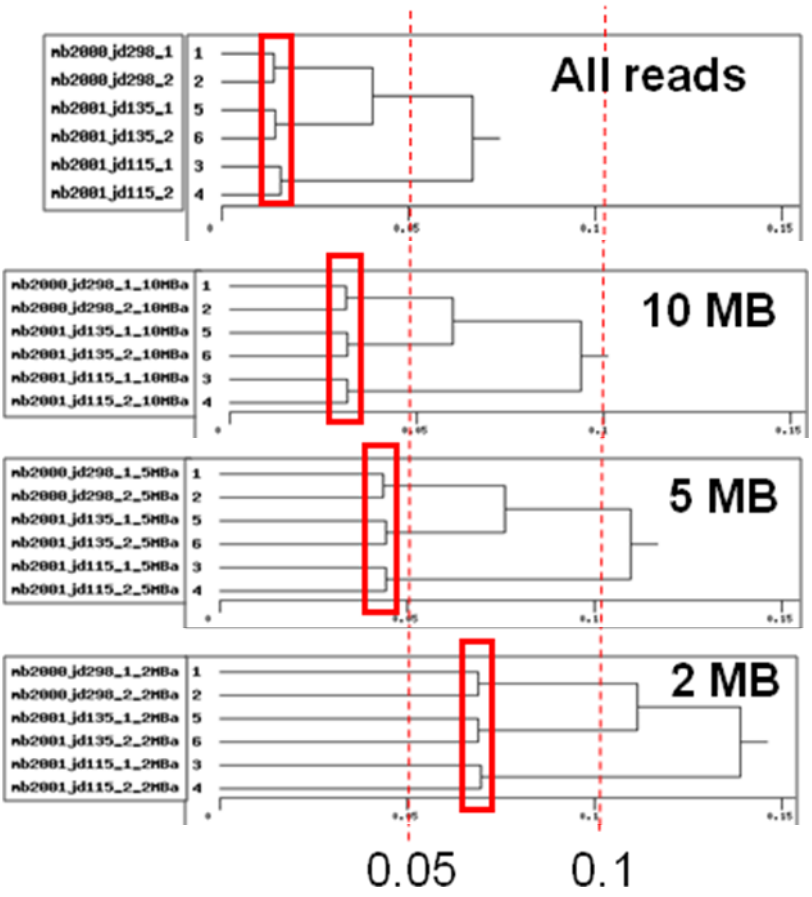


Figure 6: Clustering profiles of groups of reads from metagenomic data from Monterey Bay experiment. One pre-bloom sample (Mb2000 jd298) and two post-bloom samples (Mb2000 jd115 and Mb2000 jd135) were searched. For each sample two set of reads were obtained by pyrosequencing (labelled as 1 and 2). The complete sets of reads ("All reads") or subsets of reads of 10, 5 and 2 MB were compared to each other (Genomic Signature distances for octanucleotides) and clustered (UPGMA). The red rectangular areas identify the distances at which sets and subsets of reads obtained from the same metagenomic DNA are grouped. An increase of the distance at which the reads are grouped was observed when the sample size was smaller.

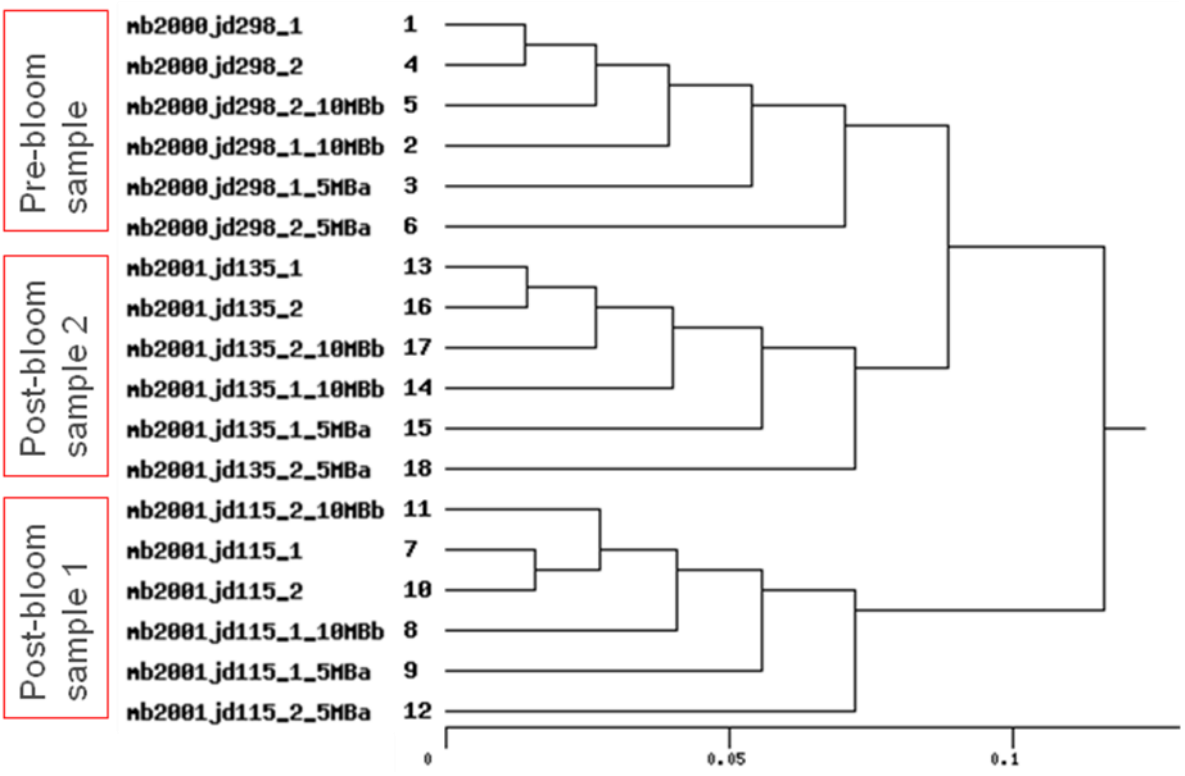


Figure 7: Clustering profiles of set and subsets of reads from metagenomic data from Monterey Bay experiment. One pre-bloom sample (Mb2000 jd298) and two post-bloom samples (Mb2000 jd115 and Mb2000 jd135) were searched. For each sample two sets of reads were obtained by pyrosequencing (labelled as 1 and 2). The complete sets of reads ("All reads") and the subsets of reads of 10 and 5 MB were compared to each other (Genomic Signature distances for octanucleotides) and clustered (UPGMA). The figure shows that sets of reads generated from the same metagenomic sample are grouped together, but when the size of the set of reads is higher, the distance at which the grouping occurs is smaller.

BIBLIOGRAFY

Pride, D.T., Wassenaar, T.M., Ghose, C., Blaser, M.J.. 2006. **Evidence of host-virus co-evolution in tetranucleotide usage patterns of bacteriophages and eukaryotic viruses**. BMC Genomics 2006, 7:8.

Rich VI, Pham VD, Eppley J, Shi Y, DeLong EF. **Time-series analyses of Monterey Bay coastal microbial picoplankton using a 'genome proxy' microarray**. Environ Microbiol 2011;13:116-34. doi: 10.1111/j.1462-2920.2010.02314.x.

Sun S, Chen J, Li W, Altinatas I, Lin A, Peltier S, Stocks K, Allen EE, Ellisman M, Grethe J,Wooley J. **Community cyberinfrastructure for Advanced Microbial Ecology Research and Analysis: the CAMERA resource**. Nucleic Acids Res. 2011;39(Database issue):D546-51. doi: 10.1093/nar/gkq1102.

Tatusova T, Ciufo S, Fedorov B, O'Neill K, Tolstoy I. RefSeq microbial genomes database: new representation and annotation strategy. Nucleic Acids Res 2014 1;42:D553-9.

APPENDIX

Appendix table 1: Performance of different methods to correctly assign a DNA sequence to its genome. The performance is shown as the proportion of times (0 to 1) the genome from which a subsequence has been obtained is located in the top 5, 10 or 20 positions of the list of **genomes** after the comparison described in the methods. The better performing combinations of frequencies and distances are shown in bold.

Frequency /distance	Length of fragments (bp)										Top positions
	250	500	750	1000	1250	2500	5000	10000	20000	40000	
Oligos4 /Pd	**0.2683** **0.3693** **0.4872**	**0.4081** **0.5259** **0.6438**	**0.4949** **0.6122** **0.7215**	**0.5524** **0.6677** **0.7696**	**0.5951** **0.7075** **0.8002**	**0.7004** **0.7984** **0.8687**	0.7666 0.8504 0.9073	0.8099 0.8846 0.9337	0.8445 0.913 0.9553	0.8688 0.932 0.9693	top5 top10 top20
Oligos4 /wPd	0.0915 0.1426 0.2164	0.1776 0.254 0.3531	0.2489 0.3384 0.4486	0.3101 0.4097 0.5205	0.3579 0.4609 0.5734	0.5161 0.6277 0.73	0.6535 0.7546 0.8354	0.7462 0.8361 0.8969	0.8084 0.8873 0.9367	0.8483 0.9194 0.9598	top5 top10 top20
Oligos4st /GS	0.2638 0.3652 0.4865	0.403 0.5201 0.6404	0.4849 0.6036 0.7151	0.5428 0.6584 0.7599	0.5802 0.6934 0.7894	0.6802 0.7789 0.8524	0.744 0.8318 0.8929	0.7864 0.866 0.9187	0.8221 0.8963 0.9434	0.8534 0.9213 0.9622	top5 top10 top20
Oligos4st /E	0.257 0.3538 0.47	0.3827 0.4945 0.6106	0.4603 0.5725 0.6831	0.5129 0.6246 0.7286	0.5497 0.661 0.7576	0.6517 0.7516 0.8297	0.7209 0.811 0.8778	0.7691 0.8519 0.9077	0.8085 0.8855 0.9363	0.8434 0.9145 0.9577	top5 top10 top20
ZOM /Pd	0.1072 0.1547 0.2211	0.1451 0.1964 0.2646	0.1638 0.2156 0.284	0.1746 0.227 0.2954	0.1827 0.2353 0.302	0.1986 0.251 0.3184	0.2095 0.2611 0.3266	0.2154 0.268 0.3327	0.2216 0.2714 0.3373	0.2236 0.2743 0.3392	top5 top10 top20
ZOM /wPd	0.0281 0.0463 0.0731	0.051 0.079 0.1191	0.0684 0.1027 0.1518	0.0842 0.1231 0.1768	0.0953 0.1369 0.1947	0.1304 0.1773 0.2394	0.1584 0.2066 0.2682	0.1786 0.2262 0.2859	0.1904 0.24 0.2989	0.1981 0.2487 0.3066	top5 top10 top20
FOM /Pd	0.1418 0.207 0.2934	0.2607 0.3535 0.4611	0.3523 0.4572 0.5688	0.4251 0.5336 0.6463	0.4816 0.595 0.7042	0.6471 0.7555 0.8439	0.7649 0.8573 0.9209	**0.8345** **0.9105** **0.9566**	0.8738 0.938 0.9745	0.8952 0.9514 0.9828	top5 top10 top20
FOM /wPd	0.0492 0.0817 0.132	0.1132 0.1722 0.2553	0.1823 0.2594 0.3607	0.2453 0.3374 0.4482	0.3036 0.4048 0.5202	0.505 0.6188 0.7314	0.683 0.7881 0.8721	0.796 0.8824 0.9402	0.8578 0.9281 0.9704	0.8879 0.9471 0.9815	top5 top10 top20
FOM /GS	0.1463 0.2124 0.3035	0.2775 0.3758 0.4902	0.3784 0.4887 0.6087	0.4576 0.5747 0.6887	0.5199 0.6371 0.7462	0.683 0.7894 0.8696	**0.7828** **0.8692** **0.926**	0.834 0.9068 0.951	0.8668 0.9312 0.9697	0.8899 0.9472 0.9805	top5 top10 top20
FOM /E	0.143 0.2083 0.2951	0.2639 0.3576 0.4676	0.3564 0.4622 0.5755	0.4292 0.5401 0.6531	0.4876 0.6012 0.7099	0.6484 0.7544 0.8418	0.7574 0.8487 0.9115	0.8205 0.8972 0.9443	0.8597 0.9268 0.9653	0.8847 0.9437 0.9772	top5 top10 top20
SOM /Pd	0.0502 0.0813 0.1271	0.1058 0.1586 0.229	0.1599 0.2283 0.3141	0.2115 0.2881 0.3836	0.2566 0.3421 0.4448	0.4247 0.531 0.6401	0.6013 0.7088 0.8024	0.7375 0.8333 0.904	0.8266 0.9046 0.9553	0.8759 0.9398 0.9765	top5 top10 top20
SOM /wPd	0.0266 0.0478 0.0844	0.0523 0.0883 0.1428	0.0854 0.1335 0.2036	0.1214 0.1825 0.2646	0.1568 0.2259 0.3168	0.3087 0.4075 0.5192	0.4921 0.602 0.7152	0.6565 0.7624 0.8561	0.7791 0.8703 0.936	0.8499 0.9244 0.9698	top5 top10 top20
SOM /GS	0.062 0.0969 0.1478	0.1325 0.1938 0.2723	0.1994 0.2761 0.3684	0.2592 0.3495 0.4513	0.3141 0.4118 0.5183	0.5069 0.6183 0.72	0.6793 0.78 0.8556	0.7863 0.8672 0.9192	0.8468 0.9142 0.955	0.8807 0.9394 0.9724	top5 top10 top20
SOM /E	0.0519 0.0837 0.1301	0.1086 0.1627 0.2351	0.1664 0.2355 0.3227	0.2188 0.2983 0.3954	0.265 0.3542 0.457	0.4352 0.5418 0.6498	0.6051 0.7108 0.8029	0.7298 0.8231 0.8914	0.8105 0.8885 0.9394	0.8592 0.9253 0.9634	top5 top10 top20
ZOMst /GS	0.1075 0.1489 0.202	0.1362 0.1756 0.2241	0.1487 0.1885 0.2343	0.1579 0.1965 0.2421	0.1644 0.2032 0.2471	0.1758 0.213 0.2554	0.1827 0.2207 0.26	0.1858 0.2233 0.2619	0.1873 0.2242 0.2626	0.1882 0.2241 0.262	top5 top10 top20
ZOMst /E	0.0978 0.134 0.1799	0.1294 0.1654 0.2078	0.1423 0.1782 0.2194	0.1518 0.1869 0.227	0.1577 0.1927 0.2329	0.1671 0.2026 0.2413	0.1751 0.2099 0.2473	0.1791 0.2142 0.2504	0.1825 0.2179 0.2552	0.1849 0.2203 0.2557	top5 top10 top20
FOMst /GS	0.1494 0.2171 0.3092	0.2802 0.3795 0.4939	0.3799 0.4908 0.6107	0.4583 0.5756 0.6897	0.5203 0.6378 0.7475	0.682 0.7885 0.8693	0.7824 0.8689 0.9258	0.8331 0.9067 0.951	0.8663 0.931 0.9694	0.8896 0.947 0.9801	top5 top10 top20
FOMst /E	0.1432 0.2084 0.295	0.2632 0.3569 0.4673	0.3557 0.4615 0.5753	0.4282 0.5392 0.6526	0.4869 0.6003 0.7096	0.6474 0.754 0.8413	0.7568 0.8484 0.9114	0.8199 0.8969 0.9443	0.8594 0.9266 0.9653	0.8845 0.9437 0.9772	top5 top10 top20
SOMst /GS	0.0604 0.0935 0.1431	0.1314 0.192 0.2707	0.1991 0.276 0.3676	0.2589 0.3495 0.4522	0.3144 0.4124 0.5196	0.5076 0.6201 0.7226	0.6804 0.7824 0.8587	0.7873 0.8689 0.9217	0.846 0.9146 0.9562	0.8795 0.9389 0.9725	top5 top10 top20
SOMst /E	0.0526 0.0846 0.1326	0.1088 0.1637 0.2375	0.1666 0.2365 0.324	0.2192 0.2987 0.3965	0.2653 0.355 0.4578	0.4354 0.5421 0.6506	0.6058 0.7116 0.8037	0.7311 0.8244 0.8923	0.812 0.8898 0.9405	0.861 0.9267 0.9643	top5 top10 top20
Zscore /Pd	0.0868 0.1331 0.1997	0.172 0.2438 0.3357	0.2491 0.3388 0.4442	0.3172 0.4199 0.5301	0.3767 0.4835 0.5965	0.5785 0.6878 0.7843	0.7387 0.8331 0.9007	0.8291 0.9042 0.9518	**0.8762** **0.9385** **0.9755**	**0.902** **0.9549** **0.9848**	top5 top10 top20
Zscorest /GS	0.0273 0.0431 0.0663	0.0575 0.0822 0.1147	0.0859 0.1184 0.1576	0.1146 0.1537 0.1975	0.1401 0.1847 0.2329	0.2475 0.3038 0.3565	0.3687 0.4292 0.4797	0.4768 0.5371 0.5844	0.5717 0.6312 0.674	0.6503 0.707 0.7477	top5 top10 top20
Zscorest /E	0.0275 0.043 0.0656	0.0615 0.0875 0.1194	0.0938 0.1268 0.1652	0.1259 0.1645 0.2078	0.1526 0.1962 0.2429	0.2622 0.3164 0.3668	0.38 0.4387 0.4871	0.4821 0.5421 0.5882	0.5753 0.6347 0.6773	0.6531 0.7104 0.7507	top5 top10 top20

Appendix table 2: Performance of different methods to correctly assign a DNA sequence to its genome. The performance is shown as the proportion of times (0 to 1) the genome from which a subsequence has been obtained is located in the top 5, 10 or 20 position of the list of **genera** after the comparison described in methods. The better performing combinations of frequencies and distances are shown in bold.

Frequency /distance	Length of fragments (bp)										Top positions
	250	500	750	1000	1250	2500	5000	10000	20000	40000	
Oligos4 /Pd	0.3999 0.4944 0.5992	0.5639 0.6504 0.738	0.6553 0.731 0.8017	0.7161 0.783 0.8414	**0.7562** **0.8131** **0.8633**	**0.8509** **0.8865** **0.9169**	0.8995 0.922 0.9425	0.9338 0.9499 0.9638	0.9584 0.9699 0.9781	0.9738 0.9824 0.9883	top5 top10 top20
Oligos4 /wPd	0.1584 0.2219 0.3097	0.277 0.3582 0.4582	0.3675 0.454 0.555	0.4414 0.527 0.6253	0.5002 0.5829 0.6742	0.6765 0.7447 0.8099	0.8075 0.8525 0.8907	0.8886 0.9153 0.9381	0.9355 0.9518 0.9644	0.9627 0.9729 0.9805	top5 top10 top20
Oligos4st /GS	**0.4014** **0.4991** **0.6049**	**0.5665** **0.6536** **0.7427**	**0.6561** **0.7323** **0.8047**	**0.7166** **0.7811** **0.8405**	0.7496 0.8087 0.8612	0.8412 0.8777 0.9105	0.8872 0.9127 0.9349	0.9203 0.9385 0.9541	0.9477 0.9617 0.9722	0.9686 0.9779 0.9853	top5 top10 top20
Oligos4st /E	0.3922 0.4863 0.5917	0.5438 0.6293 0.7193	0.6291 0.7054 0.7809	0.6876 0.755 0.8195	0.7233 0.7843 0.8403	0.8189 0.859 0.8961	0.8719 0.9002 0.9248	0.9097 0.9302 0.9477	0.9405 0.9562 0.968	0.9644 0.9749 0.9827	top5 top10 top20
ZOM /Pd	0.1948 0.2653 0.3525	0.2585 0.3292 0.4133	0.2896 0.36 0.4407	0.3072 0.3754 0.4543	0.3165 0.3839 0.4645	0.3499 0.4142 0.4883	0.3706 0.4321 0.5038	0.3846 0.4456 0.5148	0.3936 0.4577 0.5225	0.4002 0.4639 0.527	top5 top10 top20
ZOM /wPd	0.0572 0.0893 0.1356	0.0933 0.1385 0.2009	0.1232 0.1775 0.2492	0.146 0.2071 0.2821	0.1654 0.2275 0.3053	0.2255 0.291 0.3712	0.2716 0.3353 0.4163	0.3023 0.3636 0.4457	0.3203 0.3809 0.4671	0.3314 0.394 0.4822	top5 top10 top20
FOM /Pd	0.2273 0.3014 0.3974	0.38 0.4654 0.5662	0.4925 0.5798 0.6751	0.5772 0.6595 0.7441	0.6371 0.7137 0.7905	0.8095 0.8585 0.9031	0.9072 0.9336 0.9561	**0.9573** **0.9707** **0.9816**	**0.9796** **0.9859** **0.9908**	0.9889 0.9924 0.9954	top5 top10 top20
FOM /wPd	0.0971 0.1417 0.2111	0.1913 0.2584 0.3531	0.2835 0.3632 0.4683	0.3645 0.4491 0.5553	0.4339 0.5202 0.6231	0.6645 0.7367 0.8132	0.8345 0.8787 0.9197	0.9297 0.9507 0.9691	0.9694 0.9798 0.9878	0.9857 0.9905 0.9946	top5 top10 top20
FOM /GS	0.228 0.3077 0.4112	0.4026 0.494 0.5994	0.5258 0.6175 0.7134	0.617 0.7001 0.7837	0.6794 0.7563 0.8273	0.8415 0.8843 0.9216	**0.9177** **0.9397** **0.9585**	0.953 0.9653 0.9756	0.974 0.9814 0.9873	0.9857 0.9899 0.9935	top5 top10 top20
FOM /E	0.2266 0.3017 0.3999	0.3838 0.472 0.5738	0.4993 0.588 0.6825	0.5844 0.6669 0.7521	0.6436 0.7218 0.7962	0.8104 0.8595 0.9026	0.901 0.9275 0.9501	0.9459 0.9604 0.9718	0.9704 0.9785 0.9854	0.9835 0.9881 0.9922	top5 top10 top20
SOM /Pd	0.0999 0.1452 0.2113	0.183 0.2469 0.3328	0.2588 0.3349 0.4285	0.3228 0.4063 0.5028	0.3815 0.4654 0.5619	0.5836 0.6636 0.7453	0.7651 0.8208 0.8733	0.884 0.9149 0.9424	0.9455 0.9622 0.9762	0.9743 0.9827 0.9902	top5 top10 top20
SOM /wPd	0.0655 0.102 0.1586	0.1062 0.1588 0.2334	0.1555 0.221 0.3069	0.2044 0.2781 0.3718	0.2519 0.3332 0.4319	0.4411 0.5306 0.6318	0.6472 0.7213 0.8021	0.8097 0.8612 0.9114	0.9074 0.938 0.9641	0.9577 0.9721 0.985	top5 top10 top20
SOM /GS	0.1119 0.1613 0.2325	0.2168 0.2876 0.3785	0.3071 0.3901 0.4889	0.3853 0.4717 0.568	0.4535 0.5387 0.632	0.6691 0.7408 0.8083	0.8353 0.8749 0.9074	0.9184 0.9367 0.9526	0.9598 0.9688 0.9766	0.9788 0.9839 0.9881	top5 top10 top20
SOM /E	0.1003 0.1462 0.2135	0.1882 0.2542 0.3405	0.2667 0.3434 0.4371	0.3331 0.4168 0.5154	0.392 0.4772 0.5742	0.5953 0.6731 0.7542	0.7693 0.823 0.8724	0.8767 0.9073 0.9343	0.9344 0.952 0.9667	0.9633 0.9742 0.9822	top5 top10 top20
ZOMst /GS	0.1927 0.2544 0.3286	0.237 0.2922 0.3563	0.2564 0.3096 0.369	0.2686 0.3195 0.376	0.274 0.3224 0.3799	0.294 0.3401 0.3923	0.3024 0.3476 0.399	0.3066 0.3492 0.3994	0.3056 0.35 0.3987	0.3049 0.3471 0.3969	top5 top10 top20
ZOMst /E	0.1753 0.2321 0.2992	0.2227 0.2761 0.3351	0.2439 0.2926 0.3487	0.2548 0.3027 0.3559	0.2593 0.3053 0.3599	0.2775 0.3212 0.3705	0.2863 0.3282 0.3776	0.2909 0.3319 0.3794	0.2929 0.3333 0.3796	0.2937 0.3338 0.3823	top5 top10 top20
FOMst /GS	0.2329 0.3135 0.417	0.4059 0.4974 0.6028	0.5285 0.619 0.7149	0.6181 0.7014 0.7845	0.68 0.757 0.8283	0.8408 0.884 0.9211	0.9175 0.9397 0.9585	0.9527 0.9651 0.9755	0.9739 0.9812 0.9871	0.9854 0.9896 0.9932	top5 top10 top20
FOMst /E	0.226 0.3017 0.4	0.3832 0.4714 0.5729	0.4981 0.5867 0.6822	0.5832 0.6662 0.7516	0.643 0.7208 0.7959	0.8093 0.8591 0.9024	0.9006 0.9272 0.95	0.9454 0.9603 0.9717	0.9701 0.9783 0.9853	0.9832 0.988 0.9921	top5 top10 top20
SOMst /GS	0.1065 0.1531 0.2203	0.2139 0.2843 0.3747	0.3055 0.3891 0.487	0.385 0.4714 0.5674	0.4531 0.5388 0.6323	0.6705 0.7423 0.8104	0.837 0.8773 0.9107	0.9194 0.9387 0.9551	0.9593 0.9689 0.9776	0.9777 0.9834 0.9882	top5 top10 top20
SOMst /E	0.1006 0.1477 0.2149	0.1894 0.2557 0.342	0.2672 0.3443 0.4383	0.3334 0.4177 0.5165	0.3921 0.4775 0.5749	0.5953 0.6734 0.7546	0.7696 0.8234 0.8728	0.8776 0.9077 0.9348	0.9352 0.9529 0.967	0.9643 0.9748 0.9829	top5 top10 top20
Zscore /Pd	0.1479 0.2071 0.2874	0.2624 0.3397 0.4349	0.3624 0.449 0.5488	0.4453 0.5335 0.629	0.5167 0.6028 0.6938	0.7318 0.7954 0.8544	0.8797 0.9125 0.9399	0.9506 0.9644 0.9756	0.9793 0.9856 0.9904	**0.9904** **0.9932** **0.9957**	top5 top10 top20
Zscorest /GS	0.0571 0.0803 0.1109	0.1022 0.1347 0.1756	0.1474 0.1852 0.2311	0.191 0.2343 0.2825	0.2285 0.2736 0.3231	0.3752 0.4224 0.4686	0.5228 0.5612 0.5979	0.6474 0.6771 0.7049	0.7407 0.7645 0.7851	0.8109 0.8295 0.8458	top5 top10 top20
Zscorest /E	0.0586 0.0804 0.1088	0.1106 0.1414 0.1804	0.1589 0.1962 0.2383	0.2077 0.2475 0.2917	0.2455 0.2899 0.336	0.3935 0.4369 0.4785	0.5326 0.5692 0.6034	0.6544 0.6822 0.7093	0.7442 0.7669 0.7876	0.8133 0.8319 0.8476	top5 top10 top20