Barcodes for genomes and applications

Fengfeng Zhou, Victor Olman and Ying Xu

BMC Bioinformatics 2008, 9:546

I609-Week 8th

3/2/2010

Presented by Simo Zhang

Outline

- Background:
  - Problem Description
  - Methods Review
  - Motivation
- Method:
  - Barcode Matrix
  - Barcode Image
- Experiments:
  - Experiments on Prokaryotes
  - Extended Experiments
  - Barcodes in Feature Space
- Application:
  - Identification of abnormal fragments
  - Binning

Background: Problem

- Given a collection of short genomic fragments generated by metagenomic sequencing projects, bin these reads such that DNA fragments from common clade can be grouped together and assembled.

- This is a problem of binning, not classification.

Background: Methods Review

- Phymm/PhymmBL
- Phylopythia

Background: Motivation

- Earlier works[1] have observed the dinucleotide (AT, TA, CG, GC) distribution property of genomes, which is called “general design.”

- Another observation from earlier work[2] revealed that some of the dinucleotides tend to repeat along the genomes and the periodicity is 10.4-10.5 bases.

- Inspired by the above idea, this barcode-based approach inspects all k-mers distribution, where $k>2$.

1. Trifonov EN, Sussman JL: Dinucleotide relative abundance extremes: a genomic signature.

2. Karlin S, Burge C: The pitch of chromatin DNA is reflected in its nucleotide sequence.

Content Overview

- Genome Barcode
- Application 1: Identification of Abnormal Fragments
- Application 2: Binning of Metagenomic Sequences

Method: Genome Barcodes

- First, calculating the barcode for each genome.
- Second, mapping the barcodes to grey levels.
- Third, getting each genome an barcode image.

Method: Generating Barcodes

- Partitioning the genome into non-overlapping fragments of length M bps, and then for each k-mer, calculating the "combined frequency" of the k-mer and its reverse complement within in each fragment.

- A barcode for one genome is defined as a matrix *M*, in which columns represent all possible k-mers, and rows represent all fragments within the genome. The value for each element *M(i, j)* corresponds to the combined frequency of a particular k-mer within the current fragment.

- There are $4^k/2$ or $(4^k+4^{K/2})/2$ k-mers. The number of rows is the total length of genome divided by M.

Method: Mapping Frequency to grey-levels

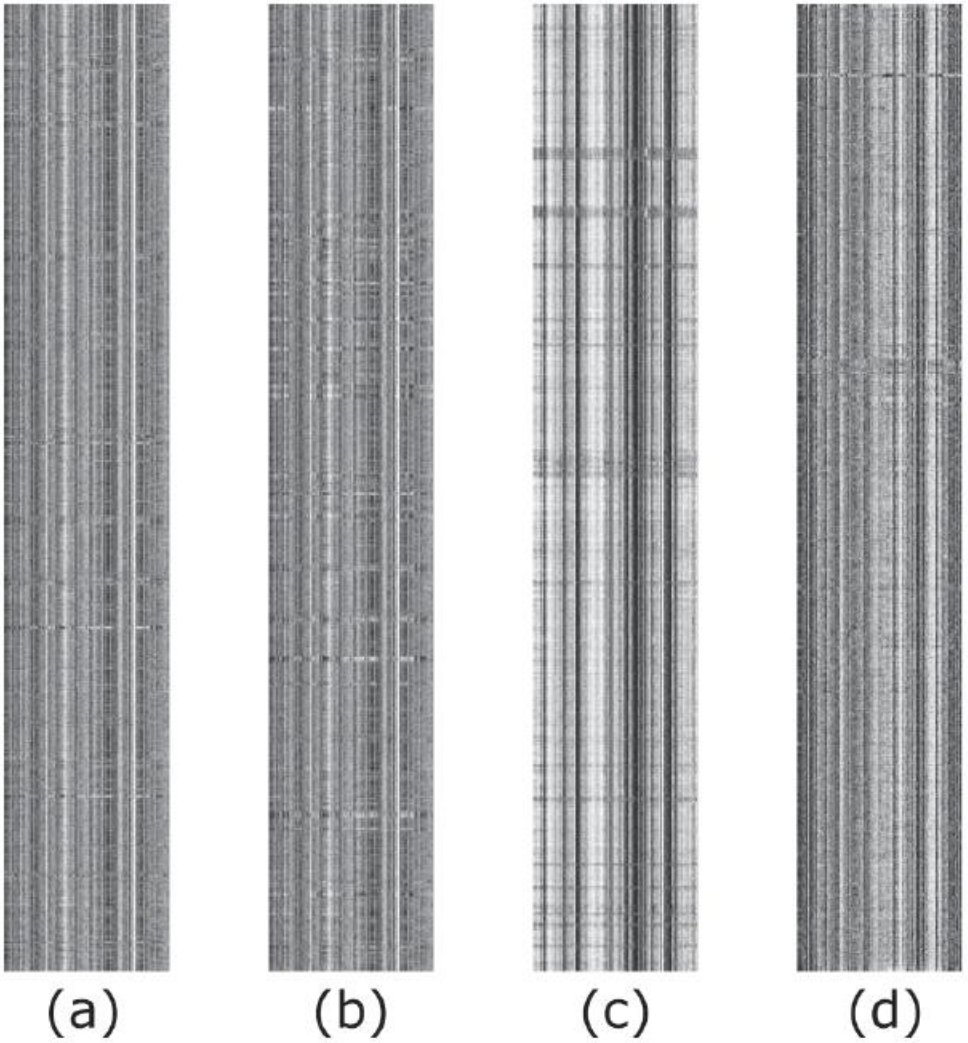
- Grey-level is defined as a vector, in which each of its element represents a frequency range. The lower the frequency is, the darker the grey level would display in the image.

- The number of grey-levels is calculated in the following step:
  - Counting the frequency of each k-mer across all genomes.
  - Sorting the frequency list *S* in the increasing order.
  - Partitioning the list into L sub-lists, such that L should minimize the following formula:

$$\sum_{i=1}^{L}(S_i - \bar{S})$$

where S_i represents for the sum of all frequencies in the ith sub-list, and $\bar{S}$ is the average frequency of S.

Method: Barcode Image



E.Coli K-12 E.Coli O157 B.pseudomallei P.furiosus

Method: Combined Frequency

- Combined frequency is calculated from k-mer and its reverse complement.

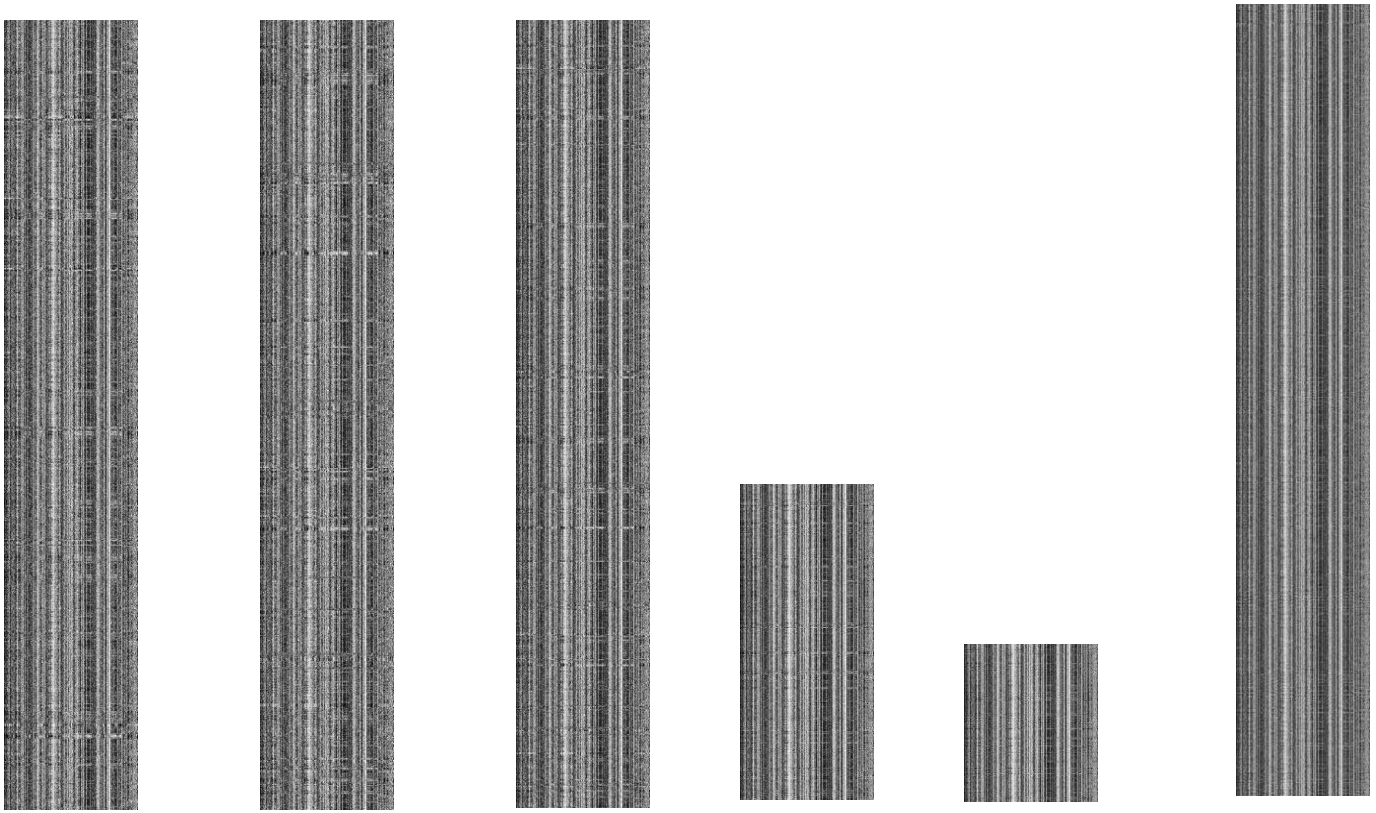
- The reason of not using single frequency based barcodes is because the combined frequency gives a more stable frequency distribution.

Fragment size	Ratio of combined 4mer/singe 4mer frequency variations
1000bps	0.7065452
2000bps	0.6958942
5000bps	0.6792713
10000bps	0.6590242

Method: Choice for M

- An appropriate value for M should take into consideration of the following two competing factors:
  - the stability of the k-mer frequencies.
  - the ability to identify the abnormal fragments.
- The longer the fragment size is, the more stable the frequencies will be.
- It is not necessary to divide the genome into “equal sized” fragments.

Method: Barcode Images with Different M



M = 1000 M = 2000 M =5000 M = 10000 Random length

Method: Barcode Distance Matrix

…	…	…	…
…	…	…	…
…	…	…	…
…	…	…	…
…	…	…	…
11	9	…	…
10	8	…	…
3	4	…	…
22	18	…	…
19	33	…	…
ATCG	ATAT	ACGT	ACGT

…	…	…		…
Lv5: …	…	…	…	…
Lv4:18-28	2	1	…	…
Lv3:13-17	3	0	…	…
Lv2:7-12	2	2	…	…
Lv1:1-6	1	1	…	…
	ATCG	ATAT	ACGT	ACGT

The columns representing all possible k-mers are the same as the barcode matrix, the rows represent the frequency of the corresponding grey-level.

Method: Barcode Distance

- Thus, the barcode distance between two barcodes is defined as

$$\sqrt{\sum_{i=1}^{k}\sum_{j=1}^{L}(M_1(i,j)-M_2(i,j))^2}$$

Method: Choice for k

- An appropriate choice for k should give a barcode of the highest discerning power, which means that fragments from same genome should have highly similar barcodes while fragments from different genome should have different barcodes. A cut-off value *d* is used as this purpose.

Method: Choice for k

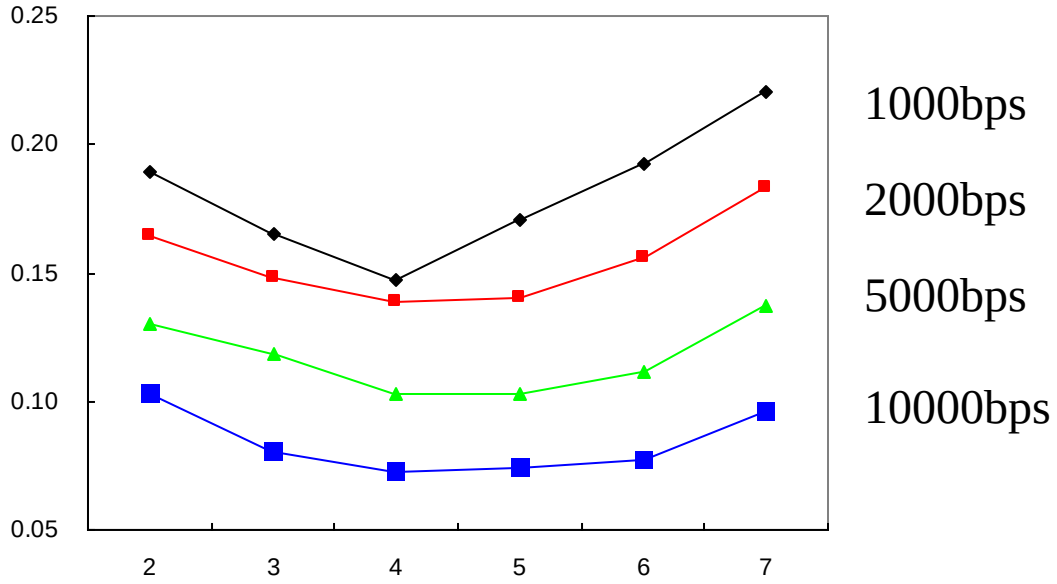
- In fact, the highest discerning power is defined in terms of the lowest total error given *d*,

$$D(k,M) = \min_{d>0}(F_k(diff,M) + 1 - F_k(same,M))$$

  - where F_k (*diff*,*M*) is the total probability of two fragments having distance less than *d*, given that they come from the different genome, while $1-F_k$ (*same*,*M*) is the total probability of two fragments having distance larger than d, given that they come from the same genome.

Method: k=4 and M=1000

- As shown in the following graph, k=4 gives barcodes the highest discerning power (minimum total error).



X-axis: The length of k-mer

Y-axis: The value for discerning power

It also shows that the combination of k = 4 and M = 1000 is the best choice among others.

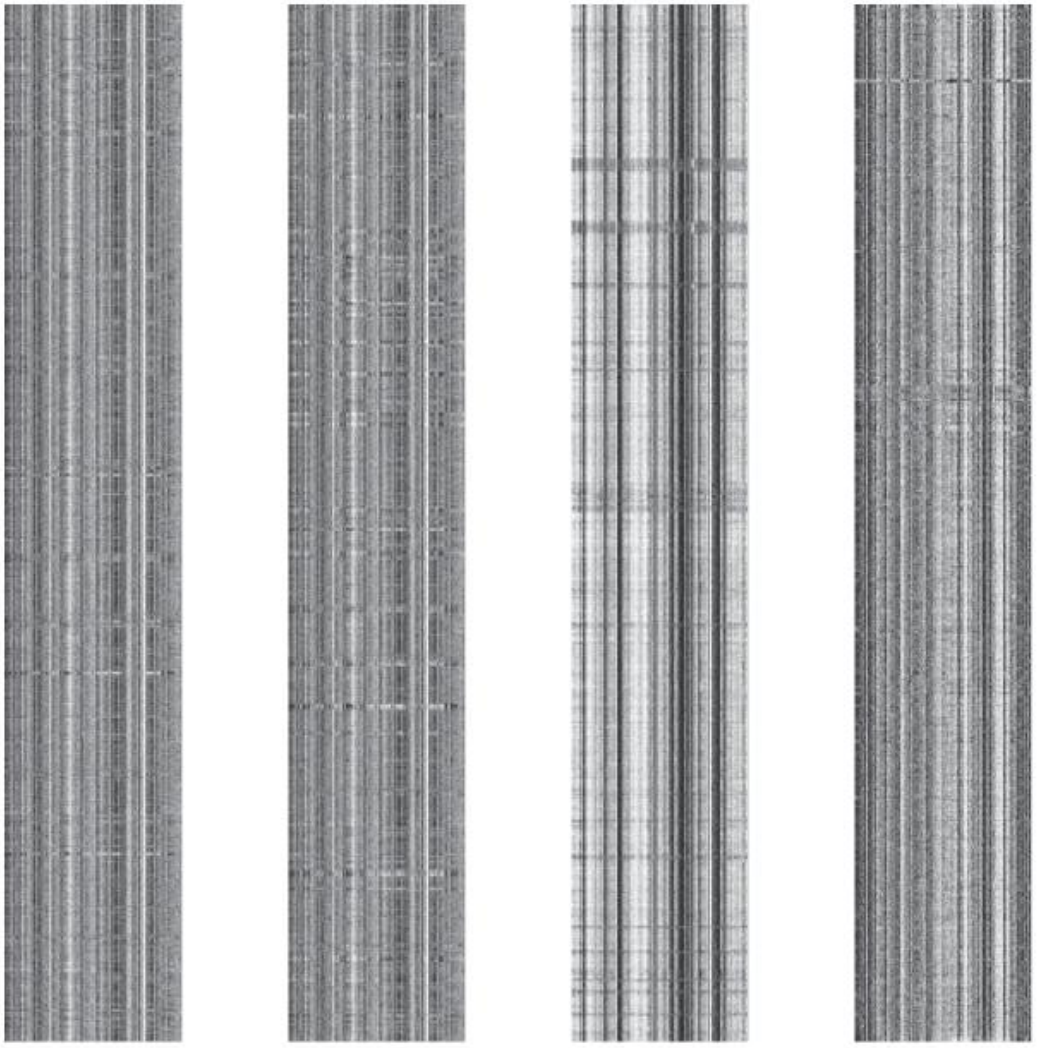
Experiment on Prokaryotes

- Observations of stable 4-mer frequency distributions over all 4-mers across 586 prokaryotic genomes have been detected.

- Consistent grey-level in each vertical band.

- A small fraction of abnormal barcodes have also been observed in the following images, represented by the horizontal stripes. These abnormal fragments have different combined k-mer frequencies from the average of the rest genome.

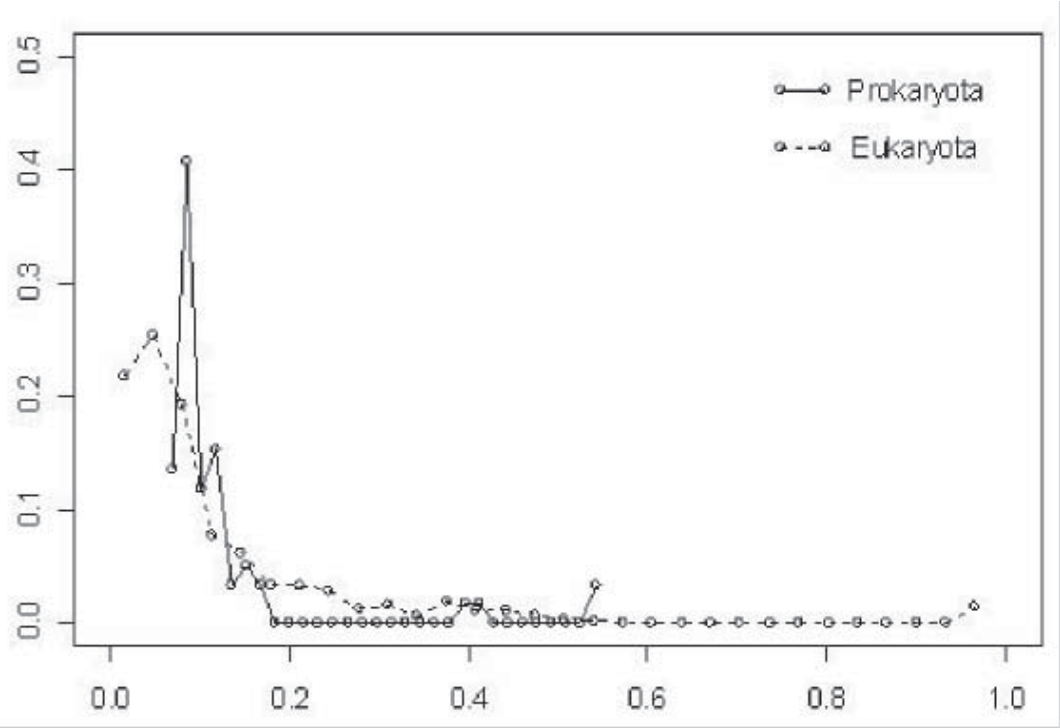
Barcode Images for four Different Genomes



E.Coli K-12 E.Coli O157 B.pseudomallei P.furiosus

Experiment on Prokaryotes

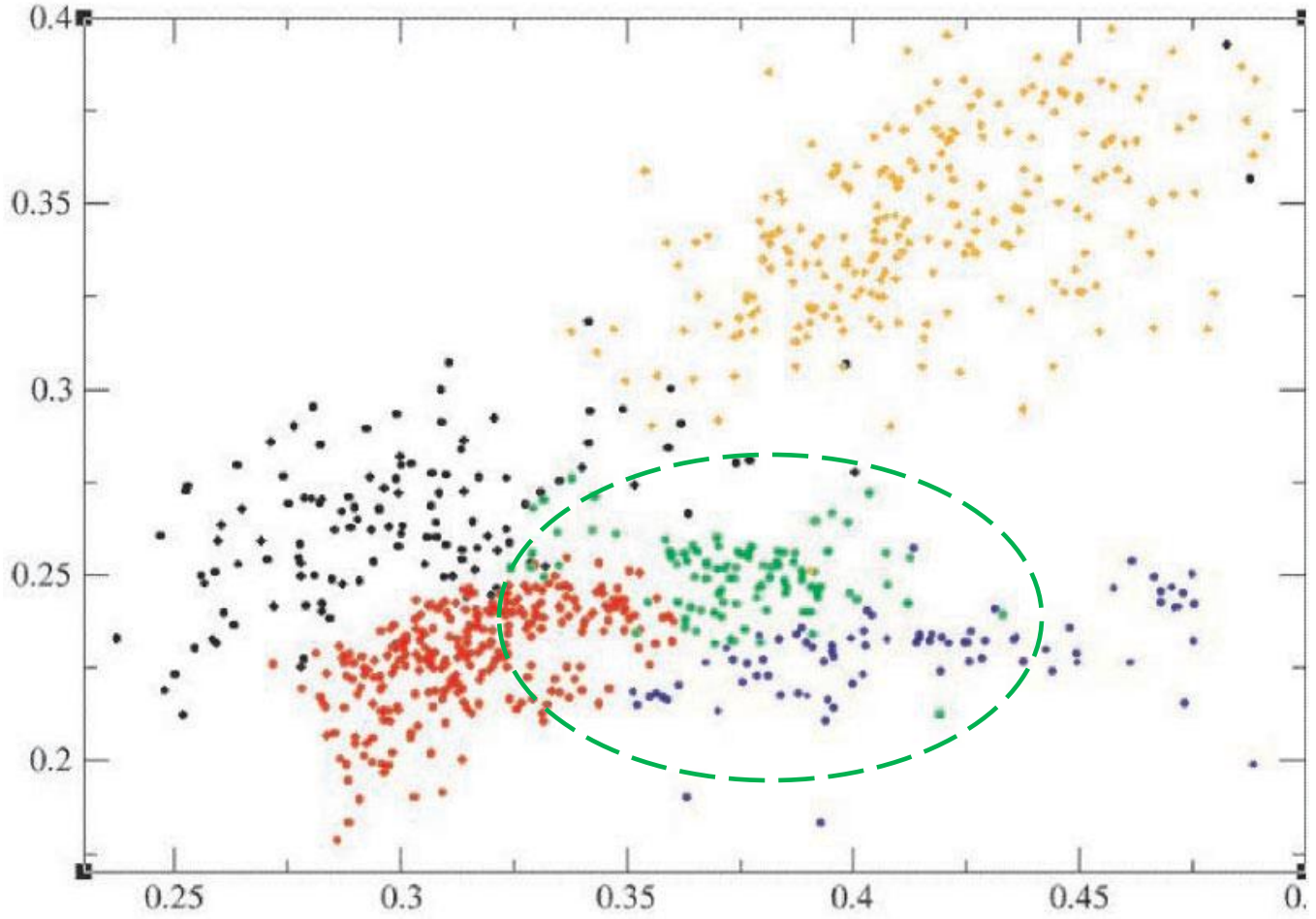
- Genomes of the same clade have highly similar barcodes, but they each also have their unique patterns of abnormal fragments. See the graph below.



X-axis: Barcode distance

Y-axis: The number of times of genome pairs having certain distance within the same organism

Experiment on Prokaryotes

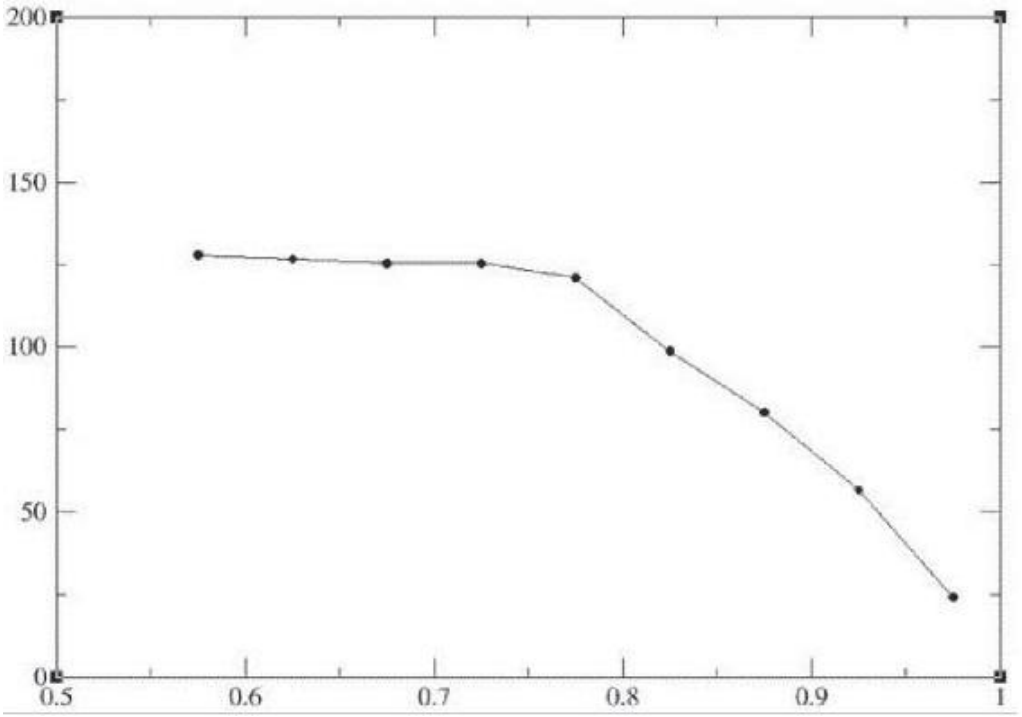


X-axis: average variation of all 4-mers within one genome

Y-axis: average similarity level among fragments within one genome

Experiment on Prokaryotes

Related genomes have closed barcode distance.

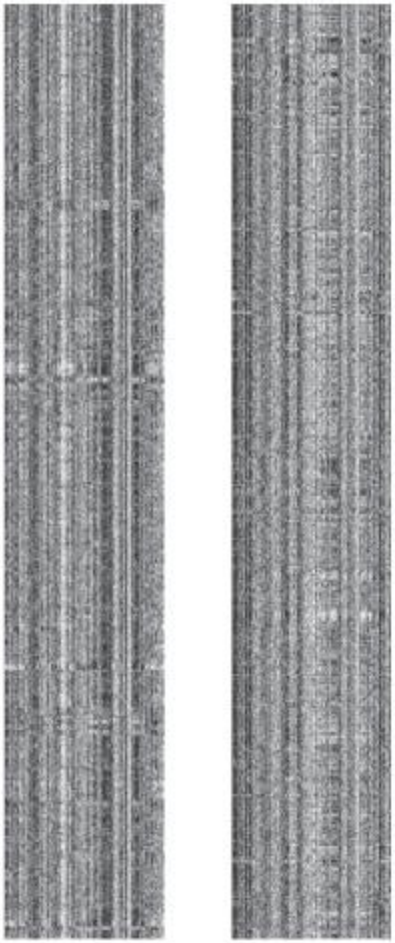


X-axis: Average sequence similarity among genomes

Y-axis: Barcode distance

Extension to Prokaryotes

- Barcodes for coding and non-coding regions are weakly similar.



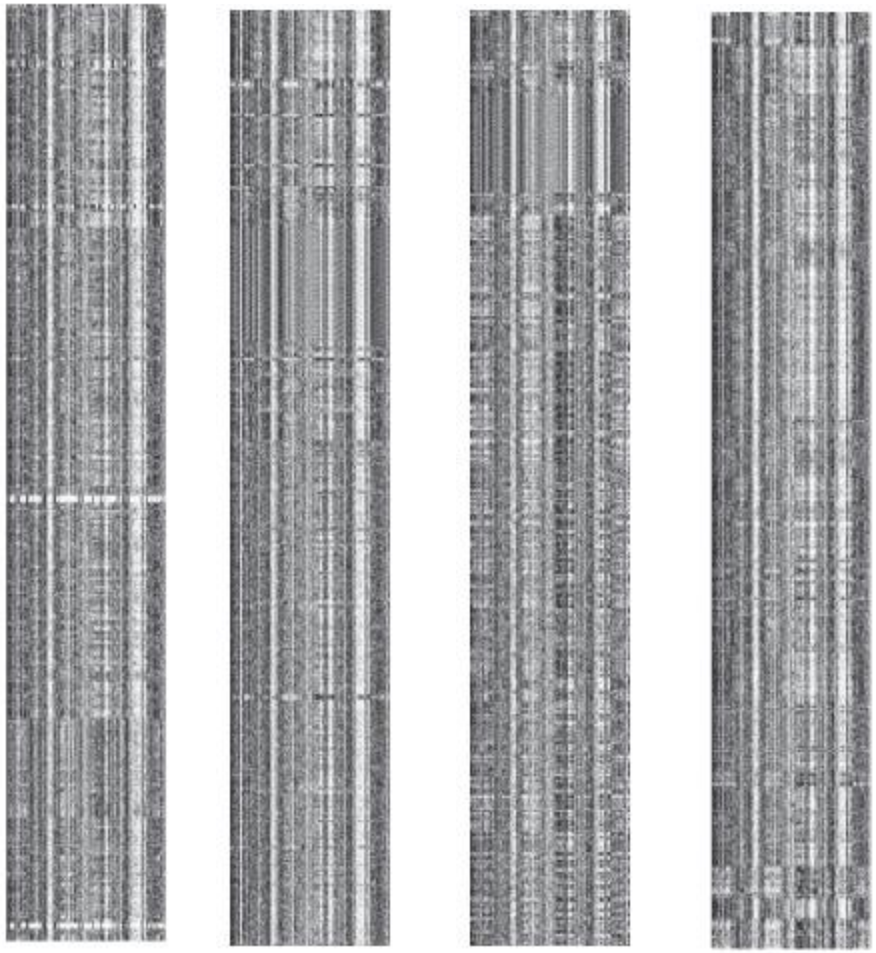
Coding regions E.coli

Non-Coding regions E.coli

Extension to Eukaryotes

- Four types of composite regions are barcoded: repetitive sequences, promoter sequences, coding regions and introns.
  - Similar barcodes have been observed among the four types of regions.
  - Four regions have an increasing “complexity,” going from repetitive sequences to coding regions to introns and promoter sequences. See the images in the following.

Extension to Eukaryotes



Repetitive sequence

Coding regions

Promoter sequence

Introns

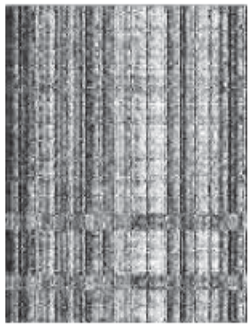
Extension to Mitochondrial and Plastid

(h)

C.elegans

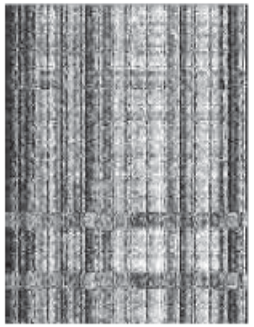
(i)

D.melanogaster



(j)

Ceratophyllum demersum



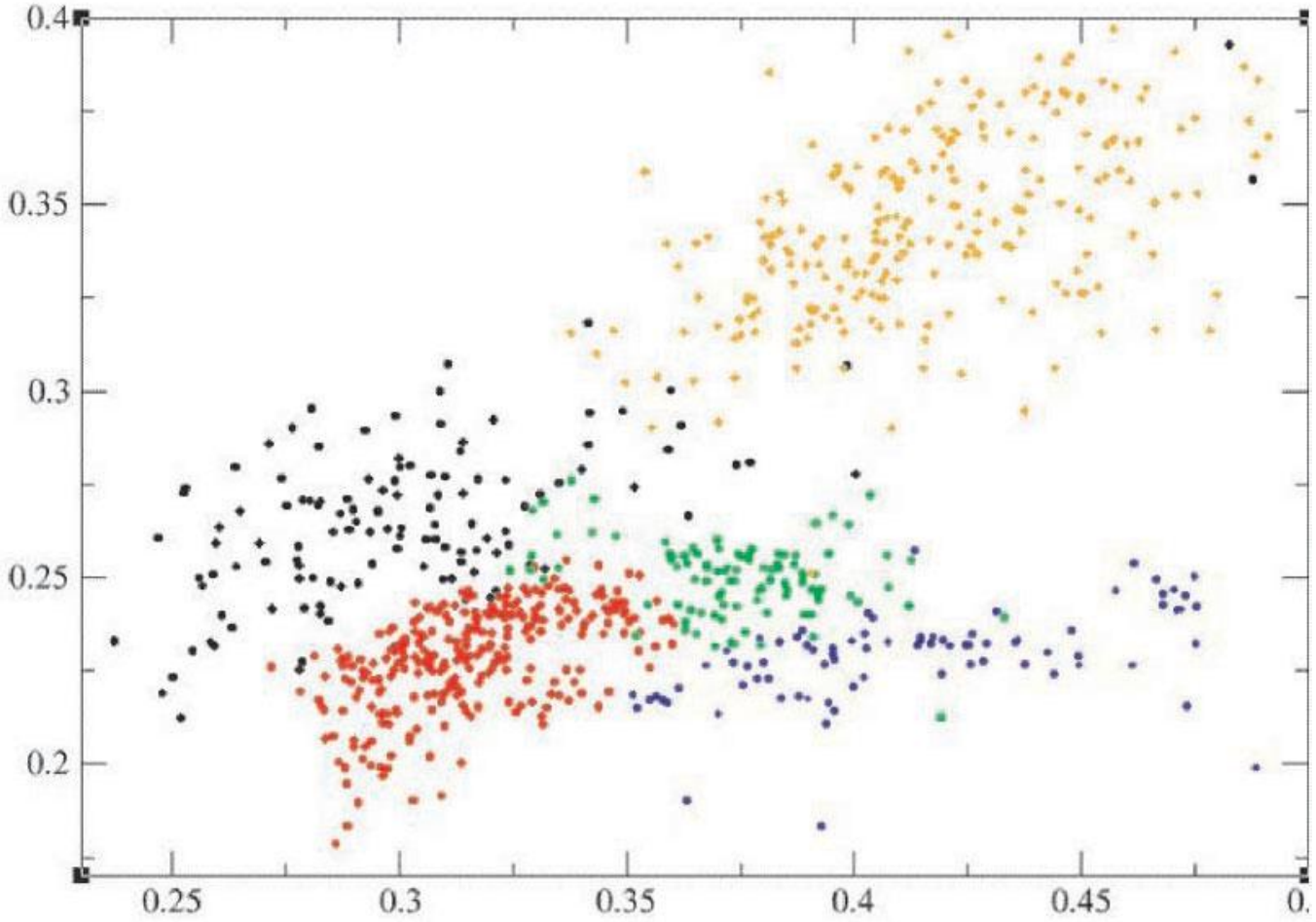
(k)

Populus trichocarpa

Barcodes in Feature Space

- Do different classes of genomes have unique characteristics in their barcodes?
  - A two-dimensional feature space below shows this unique property of barcode. The X-axis represents for the average variations of all 4-mers frequencies within one genome, while the Y-axis represents for the similarity level among all fragments within one genome.
  - The similarity levels among fragments are computed by building a minimum spanning tree.

Barcodes in Feature Space



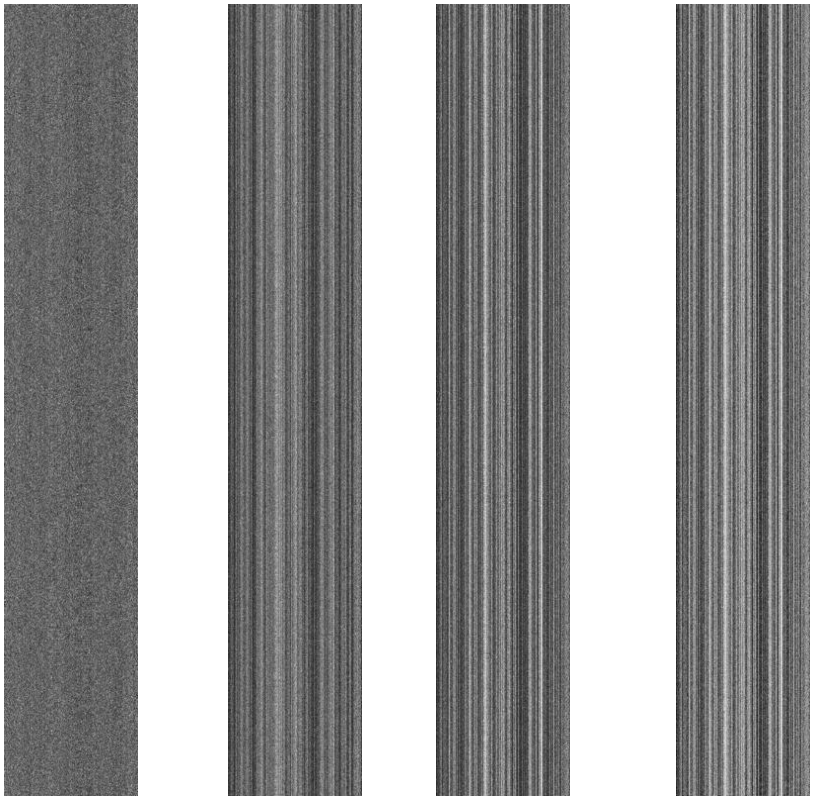
X-axis: average variation of all 4-mers within one genome

Y-axis: average similarity level among fragments within one genome

Barcodes in Feature Space

- Thinking about another question: do all the nucleotide sequences have unique barcodes like genome sequences have?
  - No! A random sequence generated using a zeroth order Markov chain model doesn't have the vertical band structures.
  - The following graph shows nucleotide sequences generated by zeroth, first, third, and fifth order Markov chain model, respectively.

Barcodes in Feature Space



Markov model	0^{th}	1^{st}	3^{rd}	5^{th}
K-mers		2-mers	4-mers	6-mers

Barcodes in Feature Space

- From the graph above, we can see:
  - The sequences generated by third order Markov model captures the barcode property of the genome sequence.
  - Higher order (>4) Markov model do not seem to add much to this property.
- This observation provides us with another reason for choosing k=4.
  - 4-mers: 3^{rd} order Markov model.
  - Reduced complexity: 4^4 vs 4^6

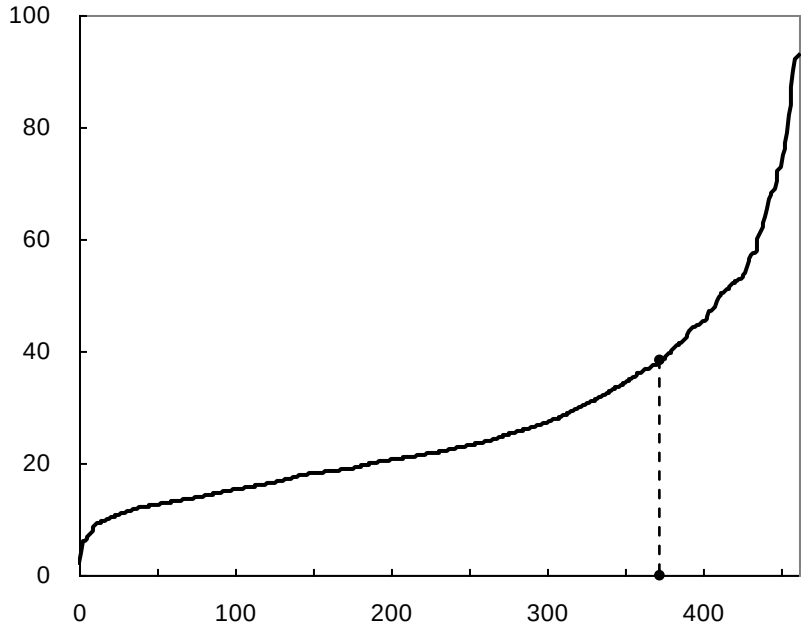
Content Overview

- Application 1: Identification of Abnormal Fragments
- Application 2: Binning of Metagenomic Sequences

Application 1: Abnormal Fragments Identification

- Method:
  - For each k-mer, selecting those fragments where the combined frequency of this k-mer has the highest or the lowest X% of its total frequency over all the fragments.
  - Sorting all the fragments in an increasing order in terms of F(p), where p represents for the index of each fragments, and the function returns the number of times each fragment has been selected by each k-mer.
  - Take $F(p_0)$ as a cut-off and fragments of $F(p_i) > F(p_0)$ are considered as the abnormal fragments, or called “non-native fragments”.

Applications: Abnormal Fragments Identification



X-axis: p, the index of each selected fragment

Y-axis: the number of times those fragments are selected

Application 1: Results

- An average of 12.85% of all the bacteria genomes have the abnormal barcodes, while 13.58% in archaeal genomes.

- So far, 30% of those abnormal fragments have been explained in terms of horizontal gene transfer (6.99%), phage invasions (4.97%), and highly expressed genes (18.90%).

- The estimation of these foreign fragments is generally consistent with the previous works, *Lateral gene transfer and the nature of bacterial innovation*.

- The remaining 70% of those abnormal fragments may fall into the above three categories but it has been proved.

Application 2: Binning of Metagenomic Sequences

- Method (CLUMP + K-means):
  - Input: given a pool of equal sized fragments with **known** number of genomes
  - Steps:
    - Using the CLUMP[1,2,3] program to get the initial clusters, and picking a seed from each cluster randomly.
    - Running the K-means algorithms by using the selected seeds from step 1.
    - Repeating the above two steps multiple times.

1. Olman V, Mao F, Wu H, Xu Y: Parallel Clustering Algorithm for Large Data Sets with applications in Bioinformatics.
2. V. Olman, D. Xu, and Y. Xu: CUBIC: Identification of Regulatory Binding Sites through Data Clustering.
3. Y. Xu, V. Olman, and D. Xu, "Clustering Gene Expression Data Using a Graph-Theoretic Approach: An Application of Minimum Spanning Tree,"

Application 2: Binning of Metagenomic Sequences

- Simulated datasets:
  - Original generated genomes.
  - 10% reduced size of original generated genomes.

Application 2: Results

Table 1: Binning accuracies of our barcode-based clustering algorithm.

	11 genomes		30 genomes		100 genomes	
	Original genomes	Filtered genomes	Original genomes	Filtered genomes	Original genomes	Filtered genomes
FS = 500 bps	71.10%	77.30%	51. 6%	55.70%	40.50%	41.10%
FS = 1000 bps	79.90%	85.90%	65.30%	70.30%	51.10%	52.60%
FS = 2000 bps	86.30%	91.70%	74.80%	80.60%	61.00%	68.53%
FS = 5000 bps	91.10%	98.10%	86.60%	93.20%	79.40%	81.90%
FS = 10000 bps	95.80%	99.30%	91.90%	97.50%	86.60%	89.18%

Application 2: Binning of Metagenomic Sequences

- Comparison with Phylopythia
  - At the species level, barcode-based approach performs better than 50% accuracy on the test set, in which the fragments size are at least 2000bp.
  - At the genus level, barcode-based approach provides more accurate binning results.
  - This comparison result should be taken carefully due to the different experiment parameters (data, the size of the data).
  - There is no training process and no training sets required for the barcode-based approach.

A Few Thoughts—Limitations

- Still unknown of what is a significant value for accuracy.
- No test on the real data, since we generally don’t know the number of genomes beforehand.
- Still cannot alleviate the pain of assembling those unknown genomes with similar barcodes.