Kmer distance: Pairwise distance base on kmer strategy

Cai Huang

Georgia Instutute of Technology,
Atlanta, GA, USA
chuang95@gatech.edu

Fredrik O. Vannberg

Georgia Instutute of Technology,
Atlanta, GA, USA
fredrik.vannberg@biology.gatech.edu

July 10, 2015

Abstract

To date the comparison of genomic DNA sequences have routinely utilized shorter conserved regions for comparative genomics. Current phylogenetic analysis therefore can create divergent results based on which genetic loci are utilized for this analysis. Sequence similarity is also commonly determined by first carrying out gap penalty pairwise alignments for a set of sequences, and the similarity is quantified based upon this alignment. This approach, however, has limitations and is primarily utilized to compare relatively conserved sequences and is also dependent upon the algorithm utilized to create the alignment. From first principles we here develop a framework for encoding nucleic acid sequences into fixed length sections (k-mers) which we utilize to create an invariant pairwise distance of total information content of genome sequences. Here we provide a R package which has two kmer base algorithms to compute the pairwise comparison of genomic DNA sequence. First function is Boolean analysis by using XOR discrete function, second function is linear algebra analysis by using hierarchical PCA.

Contents

1 Loading the R package

First, download and install the *Finch* R package. Launch R and load the package.

```
require(Matrix)
```

```
## Loading required package:  Matrix
## Loading required package:  methods
```

```
library(Finch)
```

The examples in this document are integer kc files generated by using Kanalyze[1] with following command.

$$\mathrm{ls} * \mathrm{fna}|\mathrm{xargs} - \mathrm{Ionejava} - \mathrm{jarkanalyze.jarcount} - \mathrm{k8} - \mathrm{oone\_8.kc} - \mathrm{ffasta} - \mathrm{pkanalyze.outfmt} = \mathrm{int} - \mathrm{rone}$$

The kc files contents kmer and counts for each genome sequence. In our test data set, there are 6 kc files generated from 6 virus genome sequences, and we choose kmer length equal to 8

2 Compute pairwise distance by using XOR

The first function `Finch.dif` compute pairwise distance of kmer data by using XOR discrete function, and we call it boolean analysis. We can call this function.

```
x <- Finch.dif(8,"../../Finch/data") #kmer.length = 8, path.to.data to data folder
```

The x return value is a distance matrix. Each column and row represents one virus genome. We can check the distance matrix.

```
names(x)
```

```
## [1] "Mycobacterium.phage.Phaedrus.complete.genome.fna_8.kc"
## [2] "Mycobacterium.phage.Phlyer.complete.genome.fna_8.kc"
## [3] "Mycobacterium.phage.Pipefish.complete.genome.fna_8.kc"
## [4] "Staphylococcus.phage.55.complete.genome.fna_8.kc"
## [5] "Staphylococcus.phage.69.complete.genome.fna_8.kc"
## [6] "Staphylococcus.phage.71.complete.genome.fna_8.kc"
```

```
names(x) <- c(1:6)
x
```

```
##            1          2         3         4         5         6
## 1 0.00000000 0.04437256 0.1076355 0.6295776 0.6422577 0.6326599
## 2 0.04437256 0.00000000 0.0980835 0.6282349 0.6428986 0.6325989
## 3 0.10763550 0.09808350 0.0000000 0.6255188 0.6399384 0.6290283
## 4 0.62957764 0.62823486 0.6255188 0.0000000 0.2744904 0.2078552
## 5 0.64225769 0.64289856 0.6399384 0.2744904 0.0000000 0.2796173
## 6 0.63265991 0.63259888 0.6290283 0.2078552 0.2796173 0.0000000
```

First 3 virus are *Mycobacterium.phage* and last 3 are *Staphylococcus.phage*. From the matrix, we see the pairwise distance among first 3 genomes and last 3 genomes are lower than the pairwise distance across the them. Base on this, we can clearly cluster them to two family.

3 Compute pairwise distance by using hierarchical PCA

The second function. `Finch.hpca` compute the pairwise distance of kmer data by using hierarchical PCA, and we call it linear algebra analysis. We can call this function.

```
y <- Finch.hpca(8,"../../Finch/data") #kmer.length = 8, path.to.data to data folder
```

This function also returns y value as a distance matrix. Each column and row represents one virus genome. We can check the distance matrix.

```
names(y)
```

```
## [1] "Mycobacterium.phage.Phaedrus.complete.genome.fna_8.kc"
## [2] "Mycobacterium.phage.Phlyer.complete.genome.fna_8.kc"
## [3] "Mycobacterium.phage.Pipefish.complete.genome.fna_8.kc"
## [4] "Staphylococcus.phage.55.complete.genome.fna_8.kc"
## [5] "Staphylococcus.phage.69.complete.genome.fna_8.kc"
## [6] "Staphylococcus.phage.71.complete.genome.fna_8.kc"
```

```
names(y) <- c(1:6)
y
```

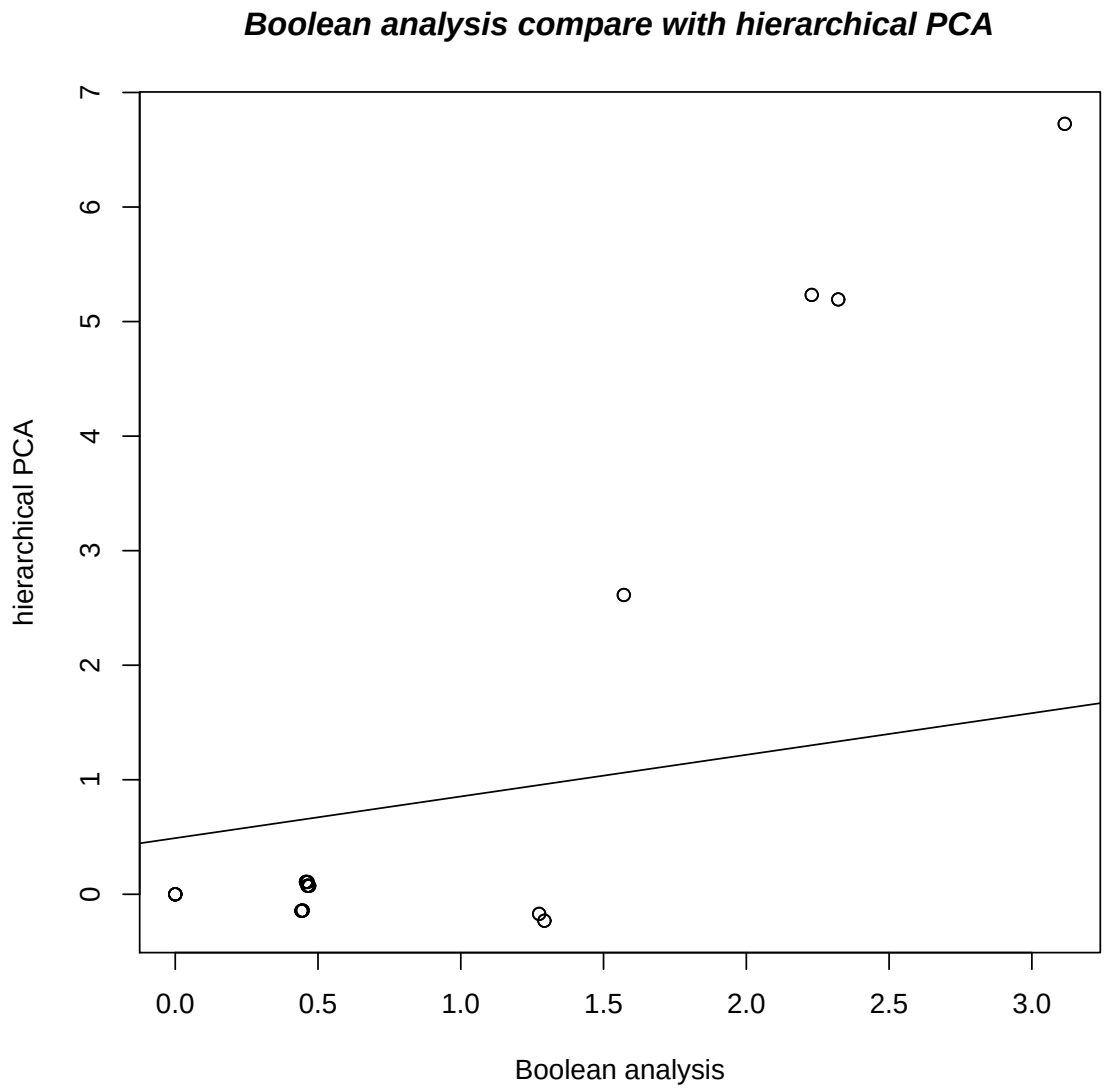
```
##              1           2           3         4         5         6
## 1 0.000000000 0.001199120 0.005338947 0.9282675 1.154281 0.8973571
## 2 0.001199120 0.000000000 0.005554642 0.9287646 1.155395 0.8979285
## 3 0.005338947 0.005554642 0.000000000 0.9291632 1.151979 0.8980785
## 4 0.928267481 0.928764635 0.929163200 0.0000000 1.259683 0.0733131
## 5 1.154280685 1.155395182 1.151979471 1.2596828 0.000000 1.1865236
## 6 0.897357057 0.897928506 0.898078484 0.0733131 1.186524 0.0000000
```

Similarly, we see the pairwise distance are clearly separating the two group of viruses.

4 Compare the results generated by XOR and hierarchical PCA

Further more, we plot the x and y pairwise distance matrix together, and show the nice correlation between our two algorithm.

```
#prepare the result as vector
x=as.vector(data.matrix(x))
y=as.vector(data.matrix(y))
#log transfer data
x[x>0]=-log(x[x>0])
y[y>0]=-log(y[y>0])
#fit a linear regression model to show the correlation between x and y
reg=lm(x~y)
plot(x, y,xlab="Boolean analysis",ylab="hierarchical PCA")
title(main="Boolean analysis compare with hierarchical PCA", col.main="black", font.main=4)
abline(reg)
```




```
#show the coefficient of determination
summary(reg)$r.squared
```

```
## [1] 0.8295085
```

From the plot we can see the linear regression model shows a good correlation between our two function with 0.8295 R squared score . Also the dots are located to two separate parts of the plot, which give us the evidence of two clusters among the genome data.

5 Session info

```
sessionInfo()
```

```
## R version 3.0.3 (2014-03-06)
## Platform: x86_64-apple-darwin10.8.0 (64-bit)
##
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
## attached base packages:
## [1] methods   stats     graphics  grDevices utils     datasets  base
##
## other attached packages:
## [1] Finch_0.0.8  Matrix_1.1-5 knitr_1.9
##
## loaded via a namespace (and not attached):
## [1] evaluate_0.5.5  formatR_1.0     grid_3.0.3      highr_0.4
## [5] lattice_0.20-30 stringr_0.6.2   tools_3.0.3
```

References

[1] Audano, P. and Vannberg, F. (2014). Kanalyze: a fast versatile pipelined k-mer toolkit. *Bioinformatics*, **30**(14), 2070–2.