Genomics Data 7 (2016) 202–209

Contents lists available at ScienceDirect

Genomics Data

journal homepage: www.elsevier.com/locate/gdata

# Similarity analysis between chromosomes of *Homo sapiens* and monkeys with correlation coefficient, rank correlation coefficient and cosine similarity measures

Chinta Someswara Rao [a,*], S. Viswanadha Raju [b]

[a] Department of CSE, SRKR Engineering College, Bhimavaram, AP, India
[b] Department of CSE, JNTUHCEJ, JNTUniversity Hyderabad, Telangana, India

ARTICLE INFO

*Article history:*
Received 26 December 2015
Accepted 4 January 2016
Available online 7 January 2016

*Keywords:*
Correlation coefficient
Rank correlation coefficient
Cosine similarity
DNA
Chromosomes

ABSTRACT

In this paper, we consider correlation coefficient, rank correlation coefficient and cosine similarity measures for evaluating similarity between *Homo sapiens* and monkeys. We used DNA chromosomes of genome wide genes to determine the correlation between the chromosomal content and evolutionary relationship. The similarity among the *H. sapiens* and monkeys is measured for a total of 210 chromosomes related to 10 species. The similarity measures of these different species show the relationship between the *H. sapiens* and monkey. This similarity will be helpful at theft identification, maternity identification, disease identification, etc.



## 1. Introduction

Similarity measures are most important operations used in analyzing genomic data. One of the most widely used analysis paradigm is guilt-by-association that requires for measuring the similarity between the pair of genes. Guilt-by-association is important for the analysis of genome interactions because relation of two neighbor genes is often easier to interpret than direct interactions between genes [1,2,3]. A genome interaction is a measure of how surprising a genome feature is similar when compared to phenomenon of another genome [4,5,6,7].

In this study we consider chromosomes of *Homo sapiens* and different kinds of monkeys called *Callithrix jacchus*, *Chlorocebus sabaeus*, *Gorilla gorilla*, *Macaca fascicularis*, *Macaca mulatta*, *Nomascus leucogenys*, *Pan troglodytes*, *Papio anubis* and *Pongo abelli*.

We also develop $2^0$ shaft string matching algorithm that consists of input & output, initialization, main function, search function and shift_left_to_right function. The genome sets and different patterns (TAGA, AGAA,GATA,TCTA,TCAT,GAAT,AGAT,CTTT,TATC,TCTG) are taken as input. The sample_id, sample_name, sample_chromosome_name, lineno, position, noofoccurences, codi are returned as output. multiple_pattern(all patterns in the set), n(text length), m(pattern length) and all the remaining variables required in the process are initialized. In the main function the genome set is read on chromosome by chromosome basis, the individual chromosome is given to shift_left_to_right function. The shift_left_to_right function takes the rightmost character of the pattern and compares it with the characters in the text. If match occurs the position (shift value) of the text is returned to the main function. Once it receives the shift value the search function is called. In the search process character by character is compared from both the directions until a complete match or missmatch occurs. In case match occurs the successive occurrence of the pattern is computed. If the successive occurrence size is greater than 2 then the data is stored in the data base(TandemRepeatDB). If mismatch occurs the same procedure is repeated until end of the text T. The relations created and stored in TandemRepeatDB data base with names of homo_sapiens, callithrix_jacchus, chlorocebus_sabaeus, gorilla_gorilla, macaca_fascicularis, macaca_mulatta, nomascus_leucogenys, pan_troglodytes, papio_anubis and pongo_abelli.

## 2. Materials and methods

In this study, four benchmarked similarity measures are consider and applied on the values of genome datasets of *H. sapiens*, *C. jacchus*, *C. sabaeus*, *G. gorilla*, *M. fascicularis*, *M. mulatta*, *N. leucogenys*, *P. troglodytes*, *P. anubis and P. abelli* [8]. The similarity measures studied in the paper are Correlation coefficient [9,10], Rank correlation coefficient [11,12] and Cosine similarity [13,14].

### 2.1. Correlation coefficient

A correlation coefficient [9,10] is a coefficient that illustrates a quantitative measure of correlation and dependence. It shows the statistical

* Corresponding author.

http://dx.doi.org/10.1016/j.gdata.2016.01.001

relationships between two or more random variables or observed data values. Different correlation coefficients are available in literature, but in this paper, Pearson's correlation coefficient is considered and denoted by $r_{(X,Y)}$ or simply r. The Karl Pearson can be measured by the formula.

$$(\boldsymbol{r}) = \frac{[\mathbf{cov}(\mathbf{X}, \mathbf{Y})]}{\boldsymbol{\sigma_X \sigma_Y}} \quad (1)$$

where cov(X,Y) is the covariance between X and Y variables and is defined as $\mathrm{cov}(X,Y)=\frac{1}{n}\sum(X_i-\overline{X})(Y_i-\overline{Y})$. However, it can also be written as $\mathrm{cov}(X,Y) = \frac{1}{n}\sum(X_iY_i-\overline{X}\overline{Y})$. Further, n is the number of observations used to fit the model, Σ is the summation symbol, $X_i$ is the X value for observation i, $\overline{X}$ is the mean X value, $Y_i$ is the Y value for observation i, $\overline{Y}$ is the mean Y value, $\sigma_X$ and $\sigma_Y$ are standard deviations of X and Y variables and $\boldsymbol{\sigma_X} = \sqrt{\frac{1}{n}\sum(\boldsymbol{X_i}-\overline{\boldsymbol{X}})^2}$ and $\boldsymbol{\sigma_Y} = \sqrt{\frac{1}{n}\sum(\boldsymbol{Y_i}-\overline{\boldsymbol{Y}})^2}$. By executing the SQL query $\pi_{\max(noofoccurences)}$ ($\sigma_{codi=\{TAGA,AGAA,GATA,TCTA,TCAT,GAAT,AGAT,CTTT,TATC,TCTG\}}$ ({*homo_sapiens*, *callithrix_jacchus*, *chlorocebus_sabaeus*, *gorilla_gorilla*, *macaca_fascicularis*, *macaca_mulatta*, *nomascus_leucogenys*, *pan_troglodytes*, *papio_anubis and pongo_abelli*})) on TandemRepeatDB tables, MAXIMUM Tandem Repeats of each repeat in all genome tables are extracted. The queried data is given as input to correlation coefficient measure, the measures are shown in Table 1.

**Notations.** In all the tables rows represent genome data sets and columns represent Tandem Repeats. The data in tables shows similarity measures of corresponding genome data.

Table 1 shows the correlation coefficient measures of *H. sapiens* genomes versus *C. jacchus*, *C. sabaeus*, *G. gorilla*, *M. fascicularis*, *M. mulatta*, *N. leucogenys*, *P. troglodytes*, *P. anubis* and *P. abelli* genomes.

*From* Table 1, *it is observed that every Tandem Repeat has shown the positive correlation, and also observed the following correlations*:

- TATC Tandem Repeat has shown a highest positive correlation(0.4) between *H. sapiens* and *C. jacchus*, whereas TCTG has shown a less positive correlation(0.03).
- TATC Tandem Repeat has shown a highest positive correlation(0.28) between *H. sapiens* and *C. sabaeus*, whereas AGAT has shown a less positive correlation(0.001087).
- TCTA Tandem Repeat has shown a highest positive correlation(0.74) between *H. sapiens* and *G. gorilla*, whereas GAAT has shown a less positive correlation(0.01365).
- TCTG Tandem Repeat has shown a highest positive correlation (0.266) between *H. sapiens* and *M. fascicularis*, whereas TAGA has shown a less positive correlation(0.1079).
- TCTA Tandem Repeat has shown the highest positive correlation(0.25) between *H. sapiens* and *M. mulatta*, whereas TAGA has shown a less positive correlation(0.018).
- AGAT Tandem Repeat had shown the highest positive correlation(0.3147) between *H. sapiens* and *N. leucogenys*, whereas CTTT has shown a less positive correlation(0.089).
- TATC Tandem Repeat has shown a highest positive correlation(0.2737) between *H. sapiens* and *Pantroglodytes*, whereas AGAT has shown a less positive correlation(0.052729).
- GAAT Tandem Repeat has shown the highest positive correlation(0.464) between *H. sapiens* and *P. anubis*, whereas TCAT has shown a less positive correlation(0.010851).
- TAGA Tandem Repeat has shown a highest positive correlation(0.537) between *H. sapiens* and *P. abelli*, whereas GATA has shown a less positive correlation(0.013134).

**Inference.** The overall highest value 0.74 occurred at TCTA Tandem Repeat of *G. gorilla* shows a positive correlation between the sets of *H. sapiens* and *G. gorilla*.

Tables 2, 3, 4, 5, 6, 7, 8 and 9 have shown the correlation coefficient measures among the different genome data sets. Observations which are very similar to those from Table 1 can also be made from the other Tables 2, 3, 4, 5, 6, 7, 8 and 9. Some of the observations are:

- The highest value 0.8307 corresponding to TCTG Tandem Repeat of *P. troglodytes* from the Table 2 shows a positive correlation between the sets of *C. jacchus* and *P. troglodytes*.
- The highest value 0.93 corresponding to TATC Tandem Repeat of *M. mulatta* from the Table 3 shows a positive correlation between the sets of *C. sabaeus* and *M. mulatta*.
- The highest value 0.68 corresponding to GATA Tandem Repeat of *N. leucogenys* from the Table 4 shows a positive correlation between the sets of *G. gorilla* and *N. leucogenys*.
- The highest value 0.72 corresponding to GAAT Tandem Repeat of *N. leucogenys* from the Table 5 shows a positive correlation between the sets of *M. fascicularis* and *N. leucogenys*.
- The highest value 0.916 corresponding to TAGA Tandem Repeat of *P. troglodytes* from the Table 6 shows a positive correlation between the sets of *M. mulatta* and *P. troglodytes*.
- The highest value 0.840 corresponding to TAGA Tandem Repeat of *P. abelli* from the Table 7 shows a positive correlation between the sets of *N. leucogenys* and *P. abelli*.
- The highest value 0.686 corresponding to TAGA Tandem Repeat of *P. anubis* from the Table 8 shows a positive correlation between the sets of *P. troglodytes* and *P. anubis*.
- The highest value 0.56 corresponding to TAGA Tandem Repeat of *Pongo abelli* from the Table 9 shows a positive correlation between the sets of *P. anubis* and *P. abelli*.

### 2.2. Rank correlation coefficient

A rank correlation coefficient [11,12] measures the degree of similarity between two sets of data, and can be used to assess the significance of the

**Table 1**
Correlation Coefficient measures of *Homo sapiens* genomes versus *Callithrix jacchus*, *Chlorocebus sabaeus*, *Gorilla gorilla*, *Macaca fascicularis*, *Macaca mulatta*, *Nomascus leucogenys*, *Pan troglodytes*, *Papio anubis* and *Pongo abelli*.

| *Homo sapiens* (VS) | TAGA | AGAA | GATA | TCTA | TCAT | GAAT | AGAT | CTTT | TATC | TCTG |
|---|---|---|---|---|---|---|---|---|---|---|
| *Callithrix jacchus* | 0.200446 | 0.102062 | 0.171365 | 0.123596 | 0.176777 | 0.103889 | 0.127986 | 0.14017 | 0.406061 | 0.032686 |
| *Chlorocebus sabaeus* | 0.019488 | 0.072169 | 0.162614 | 0.192739 | 0.081111 | 0.029802 | 0.001087 | 0.147246 | 0.285724 | 0.278168 |
| *Gorilla gorilla* | 0.013941 | 0.369925 | 0.202242 | 0.749865 | 0.179746 | 0.01365 | 0.199109 | 0.213699 | 0.037247 | 0.152294 |
| *Macaca fascicularis* | 0.107922 | 0.131794 | 0.286145 | 0.194849 | 0.136482 | 0.238217 | 0.13257 | 0.211702 | 0.249029 | 0.266628 |
| *Macaca mulatta* | 0.018966 | 0.139963 | 0.084173 | 0.250192 | 0.139573 | 0.042875 | 0.043906 | 0.137929 | 0.23994 | 0.004386 |
| *Nomascus leucogenys* | 0.108512 | 0.290926 | 0.232048 | 0.278772 | 0.312555 | 0.331841 | 0.314733 | 0.089229 | 0.040664 | 0.14093 |
| *Pan troglodytes* | 0.131857 | 0.185799 | 0.143149 | 0.184133 | 0.272337 | 0.095368 | 0.052729 | 0.124725 | 0.273724 | 0.097109 |
| *Papio anubis* | 0.321465 | 0.154335 | 0.029247 | 0.092762 | 0.010851 | 0.46405 | 0.158686 | 0.115516 | 0.157341 | 0.117418 |
| *Pongo abelli* | 0.537383 | 0.241432 | 0.013134 | 0.47516 | 0.230636 | 0.140526 | 0.070296 | 0.263892 | 0.212457 | 0.268534 |