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, σ_X and σ_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

Table 2
Correlation coefficient measures of *Callithrix jacchus* genomes versus *Chlorocebus sabaeus*, *Gorilla gorilla*, *Macaca fascicularis*, *Macaca mulatta*, *Nomascus leucogenys*, *Pan troglodytes*, *Papio anubis* and *Pongo abelli* genomes.

Callithrix jacchus (VS)	TAGA	AGAA	GATA	TCTA	TCAT	GAAT	AGAT	CTTT	TATC	TCTG
Chlorocebus sabaeus	0.328125	0.176777	0.218182	0.066319	0.043015	0.200805	0.032511	0.045434	0.356406	0.389536
Gorilla gorilla	0.019061	0.080128	0.055201	0.012361	0.234333	0.33094	0.282006	0.060398	0.144457	0.097763
Macaca fascicularis	0.053087	0.145896	0	0.151402	0.095871	0.147296	0.117555	0.102869	0.130212	0.127185
Macaca mulatta	0.098304	0.184999	0.157378	0.08269	0.157026	0.196533	0.257248	0.185386	0.219581	0.112572
Nomascus leucogenys	0.076547	0.076547	0.076547	0.076547	0.076547	0.076547	0.076547	0.076547	0.076547	0.076547
Pan troglodytes	0.681621	0.531428	0.324655	0.369584	0.116514	0.171631	0.326814	0.454315	0.333299	0.8307
Papio anubis	0.170735	0.180679	0.099669	0.036943	0.094883	0.034143	0.329083	0.202311	0.107175	0.384421
Pongo abelli	0.166473	0.181577	0.10488	0.257085	0.089173	0.21143	0.267057	0.015841	0.026602	0.169649

Table 3
Correlation coefficient measures of *Chlorocebus sabaeus* genomes versus *Gorilla gorilla*, *Macaca fascicularis*, *Macaca mulatta*, *Nomascus leucogenys*, *Pan troglodytes*, *Papio anubis* and *Pongo abelli* genomes.

Chlorocebus sabaeus (VS)	TAGA	AGAA	GATA	TCTA	TCAT	GAAT	AGAT	CTTT	TATC	TCTG
Gorilla gorilla	0.640857	0.88212	0.405542	0.634391	0.662396	0.435253	0.106101	0.791399	0.688471	0.599238
Macaca fascicularis	0.349842	0.598501	0.499025	0.727016	0.768771	0.353851	0.223126	0.621981	0.478913	0.619823
Macaca mulatta	0.770189	0.49128	0.839825	0.381185	0.44722	0.714002	0.277369	0.531669	0.939179	0.620586
Nomascus leucogenys	0.567134	0.542375	0.437535	0.586107	0.449786	0.487557	0.495324	0.445657	0.715455	0.250417
Pan troglodytes	0.349779	0.413092	0.575629	0.472225	0.381874	0.574563	0.452958	0.503647	0.520591	0.411884
Papio anubis	0.585966	0.384426	0.151903	0.378309	0.470933	0.086993	0.562842	0.341685	0.195816	0.452875
Pongo abelli	0.332388	0.400708	0.175277	0.393403	0.278554	0.346856	0.30668	0.223723	0.337736	0.252836

Table 4
Correlation coefficient measures of *Gorilla gorilla* genomes versus M*acaca fascicularis*, *Macaca mulatta*, *Nomascus leucogenys*, *Pan troglodytes*, *Papio anubis* and *Pongo abelli* genomes.

Gorilla gorilla (VS)	TAGA	AGAA	GATA	TCTA	TCAT	GAAT	AGAT	CTTT	TATC	TCTG
Macaca fascicularis	0.491304	0.264215	0.109273	0.161771	0.260462	0.250062	0.276295	0.288584	0.294069	0.460102
Macaca mulatta	0.422999	0.677541	0.550816	0.303802	0.48935	0.293201	0.085579	0.500019	0.447214	0.3397
Nomascus leucogenys	0.317612	0.271708	0.684641	0.319758	0.374766	0.295869	0.358296	0.14395	0.448556	0.009342
Pan troglodytes	0.107299	0.147788	0.220763	0.530794	0.033686	0.305351	0.004551	0.146801	0.265016	0.0804
Papio anubis	0.400278	0.052511	0.088327	0.175621	0.037176	0.09547	0.124944	7.20E-16	0.029802	0.015433
Pongo abelli	0.408863	0.139876	0.09249	0.173788	0.189805	0.080632	0.128492	0.155725	0.213405	0.021678

Table 5
Correlation coefficient measures of *Macaca fascicularis* genomes versus *Macaca mulatta*, *Nomascus leucogenys*, *Pan troglodytes*, *Papio anubis* and *Pongo abelli* genomes.

Macaca fascicularis (VS)	TAGA	AGAA	GATA	TCTA	TCAT	GAAT	AGAT	CTTT	TATC	TCTG
Macaca mulatta	0.676184	0.091145	0.329204	0.136178	0.565936	0.291937	0.524909	0.282913	0.456832	0.112097
Nomascus leucogenys	0.299786	0.018588	0.349482	0.089532	0.631784	0.720937	0.192511	0.070829	0.423114	0.200311
Pan troglodytes	0.182887	0.008755	0.339561	0.047286	0.247775	0.016399	0.224782	0.286707	0.083361	0.238254
Papio anubis	0.114517	0.597851	0.216025	0.048224	0.187767	0.254894	0.403582	0.118345	0.354019	0.326006
Pongo abelli	0.021341	0.107491	0.009068	5.08E-16	0.26998	0.114633	0.409001	0.2095	4.96E-16	0.070376

Table 6
Correlation coefficient measure of *Macaca mulatta* genomes versus *Nomascus leucogenys*, *Pan troglodytes*, *Papio anubis* and *Pongo abelli* genomes.

Macaca mulatta (VS)	TAGA	AGAA	GATA	TCTA	TCAT	GAAT	AGAT	CTTT	TATC	TCTG
Nomascus leucogenys	0	0.084895	0.5	0.307095	0.330701	0.39615	0.180568	0.205161	0.385654	0.228129
Pan troglodytes	0.91663	0.285989	0.311647	0.03032	0.202775	0.337225	0.072509	0.349482	0.226221	0.313863
Papio anubis	0.01194	0.447368	0.046127	0.107604	0.041417	0.237945	0.199931	0.4	0.030682	0.236297
Pongo abelli	0	0.132221	0.114401	0.273635	0.243222	0.114708	0.05698	0.236743	0.188163	0.395974

Table 7
Correlation coefficient measures of *Nomascus leucogenys* versus *Pan troglodytes*, *Papio anubis* and *Pongo abelli* genomes.

Nomascus leucogenys (VS)	TAGA	AGAA	GATA	TCTA	TCAT	GAAT	AGAT	CTTT	TATC	TCTG
Pan troglodytes	0.443135	0.41762	0.360597	0.586238	0.65982	0.50545	0.223462	0.699197	0.541433	0.496358
Papio anubis	0.441707	0.688247	0.707424	0.352673	0.637482	0.35767	0.617901	0.519875	0.459358	0.486299
Pongo abelli	0.185251	0.707947	0.472408	0.35465	0.325077	0.542945	0.425792	0.609597	0.84058	0.611654

Table 8
Correlation coefficient measures of *Pan troglodytes* genomes versus *Papio anubis* and *Pongo abelli*.

Pan troglodytes (VS)	TAGA	AGAA	GATA	TCTA	TCAT	GAAT	AGAT	CTTT	TATC	TCTG
Papio anubis	0.68698	0.62301	0.471927	0.510857	0.216995	0.038472	0.186636	0.323029	0.441541	0.494709
Pongo abelli	0.143491	0.137813	0.01815	0.071753	0.081519	0.047088	0.151872	0.120074	0.129268	0.224163

Table 9
Correlation coefficient measures of *Papio anubis* genomes versus *Pongo abelli* genomes.

Papio anubis (VS)	TAGA	AGAA	GATA	TCTA	TCAT	GAAT	AGAT	CTTT	TATC	TCTG
Pongo abelli	0.560137	0.023509	0.264867	0.180269	0.136078	0.48437	0.208681	0.023444	0.090434	0.018164

relation between them. Different rank correlation coefficients are available in the literature. The Spearman's Rank correlation coefficient is considered and denoted by r, in this paper. It can be measured by the formula

$$(r) = 1 - \frac{6\sum d_i^2}{n(n^2-1)} \quad (2)$$

where $d_i = (R_X\text{-}R_Y)$ is the difference of ranks of X_i and Y_i for each i, and n is the number of pairs of observations.

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 has been arranged in the form of ranks. The ranks are given as input to rank correlation coefficient measure; the measures are shown in Table 10.

Table 10 shows the rank correlation coefficient measures of *H. sapiens* genomes versus *C. jacchus*, *C. sabaeus*, *G. gorilla*, *M. fascicularis*, *M. mulatta*, *N. leucogenys*, *Pan troglodytes*, *P. anubis and P. abelli* genomes.

From the Table 10, *it is observed that every Tandem Repeat has shown a positive rank correlation, and also observed the following correlations*:

- AGAA Tandem Repeat has shown a highest positive correlation(0.997) between *H. sapiens* and *C. jacchus*, whereas TCTG has shown a less positive correlation(0.903).
- CTTT Tandem Repeat has shown a highest positive correlation(0.993) between *H. sapiens* and *C. sabaeus*, whereas TCTG has shown a less positive correlation(0.911).
- GATA Tandem Repeat has shown a highest positive correlation(0.990) between *H. sapiens* and *G. gorilla*, whereas TATC has shown a less positive correlation(0.944).
- TATC Tandem Repeat has shown a highest positive correlation(0.993) between *H. sapiens* and *M. fascicularis*, whereas GATA has shown a less positive correlation(0.9610).
- AGAA Tandem Repeat has shown a highest positive correlation(0.998) between *H. sapiens* and *M. mulatta*, whereas GAAT has shown a less positive correlation(0.883).
- AGAA Tandem Repeat has shown a highest positive correlation(0.998) between *H. sapiens* and *N. leucogenys*, whereas TATC has shown a less positive correlation(0.934).
- CTTT Tandem Repeat has shown a highest positive correlation(0.996) between *H. sapiens* and *Pan troglodytes*, whereas TCAT has shown a less positive correlation(0.943).
- AGAA Tandem Repeat has shown a highest positive correlation(0.998) between *H. sapiens* and *Papio anubis*, whereas TCAT has shown a less positive correlation(0.907).
- AGAA Tandem Repeat had shown a highest positive correlation(0.998) between *H. sapiens* and *Pongo abelli*, whereas GATA has shown a less positive correlation(0.899).

Inference. The overall highest value 0.998 occurred at AGAA Tandem Repeat of *pongo abelli*, *P. anubis*, *N. leucogenys* and *M. mulatta* shows a positive correlation between the sets of *H. sapiens* and *P. abelli*, *P. anubis*, *N. leucogenys*, *M. mulatta*.

Tables 11, 12, 13, 14, 15, 16, 17, and 18 have shown the *Rank Correlation Coefficient measures* among the different genome data sets. Observations which are very similar to those from Table 10 can also be made from the other Tables 11,12, 13, 14, 15, 16, 17, and 18. Some of the observations are:

- The highest value 0.997 corresponding to AGAA Tandem Repeat of *P. abelli* from the Table 11 shows a positive correlation between the sets of *C. jacchus* and *P. abelli*.
- The highest value 0.997 corresponding to AGAA Tandem Repeat of *M. mulatta* from the Table 12 shows a positive correlation between the sets of *C. sabaeus* and *M. mulatta*.
- The highest value 0.997 corresponding to AGAA Tandem Repeat of *P. anubis* from the Table 13 shows a positive correlation between the sets of *G. gorilla* and *P. anubis*.
- The highest value 0.997 corresponding to AGAA Tandem Repeat of *P. anubis* from the Table 14 shows a positive correlation between the sets of *M. fascicularis* and *P. anubis*.
- The highest value 0.998 corresponding to AGAA Tandem Repeat of *P. anubis* from the Table 15 shows a positive correlation between the sets of *M. mulatta* and *P. anubis*.
- The highest value 0.996 corresponding to AGAA Tandem Repeat of *P. abelli* from the Table 16 shows a positive correlation between the sets of *N. leucogenys* and *P. abelli*.
- The highest value 0.986 corresponding to AGAA Tandem Repeat of *P. anubis* from the Table 17 shows a positive correlation between the sets of *Pan troglodytes* and *P. anubis*.
- The highest value 0.997 corresponding to AGAA Tandem Repeat of *P. abelli* from the Table 18 shows a positive correlation between the sets of *P. anubis* and *P. abelli*.

Table 10
Rank 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* genomes.

Homo sapiens (VS)	TAGA	AGAA	GATA	TCTA	TCAT	GAAT	AGAT	CTTT	TATC	TCTG
Callithrix jacchus	0.963462	0.997692	0.986154	0.981538	0.979231	0.983462	0.976154	0.994231	0.986923	0.903846
Chlorocebus sabaeus	0.973462	0.989615	0.990385	0.970385	0.982692	0.949231	0.953462	0.993462	0.979231	0.911923
Gorilla gorilla	0.979249	0.988142	0.990119	0.98419	0.979743	0.974802	0.980731	0.977767	0.94417	0.980731
Macaca fascicularis	0.976849	0.986448	0.961039	0.988142	0.976285	0.980802	0.961604	0.987578	0.993224	0.980802
Macaca mulatta	0.983766	0.998052	0.975325	0.975325	0.933766	0.883117	0.964935	0.986364	0.986364	0.977273
Nomascus leucogenys	0.971429	0.998701	0.977922	0.974026	0.964286	0.969481	0.985065	0.99026	0.934416	0.976623
Pan troglodytes	0.985217	0.965217	0.984348	0.973043	0.943913	0.98087	0.951304	0.996087	0.96	0.95087
Papio anubis	0.978543	0.998306	0.957651	0.981366	0.981366	0.981366	0.980802	0.98419	0.907962	0.994918
Pongo abelli	0.982213	0.998518	0.977767	0.979743	0.975296	0.987648	0.985178	0.987154	0.972826	0.899209

Table 11
Rank correlation coefficient measures of Callithrix jacchus genomes versus *Chlorocebus sabaeus*, *Gorilla gorilla*, *Macaca fascicularis*, *Macaca mulatta*, *Nomascus leucogenys*, *Pan troglodytes*, *Papio anubis* and *Pongo abelli* genomes.

Callithrix jacchus(VS)	TAGA	AGAA	GATA	TCTA	TCAT	GAAT	AGAT	CTTT	TATC	TCTG
Chlorocebus sabaeus	0.937692	0.986538	0.986538	0.943462	0.990385	0.921154	0.967308	0.995385	0.991538	0.987308
Gorilla gorilla	0.986166	0.985178	0.987154	0.972826	0.981225	0.975296	0.969862	0.981719	0.969368	0.971838
Macaca fascicularis	0.950875	0.981366	0.939018	0.985319	0.961604	0.981366	0.970638	0.988142	0.979108	0.966685
Macaca mulatta	0.980519	0.996104	0.988312	0.971429	0.873377	0.887662	0.921429	0.987013	0.97013	0.967532
Nomascus leucogenys	0.952597	0.995455	0.977273	0.958442	0.980519	0.977922	0.95974	0.994156	0.963636	0.991558
Pan troglodytes	0.955652	0.962174	0.982609	0.94913	0.88087	0.976522	0.973913	0.994348	0.984348	0.973478
Papio anubis	0.985319	0.996612	0.953698	0.971767	0.965556	0.955957	0.961604	0.983625	0.945793	0.944664
Pongo abelli	0.91996	0.99753	0.973814	0.976285	0.969862	0.990119	0.950593	0.988142	0.982213	0.974308

Table 12
Rank correlation coefficient measures of *Chlorocebus sabaeus* genomes versus *Gorilla gorilla*, *Macaca fascicularis*, *Macaca mulatta*, *Nomascus leucogenys*, *Pan troglodytes*, *Papio anubis* and *Pongo abelli* genomes.

Chlorocebus sabaeus (VS)	TAGA	AGAA	GATA	TCTA	TCAT	GAAT	AGAT	CTTT	TATC	TCTG
Gorilla gorilla	0.952569	0.994565	0.990613	0.972332	0.988142	0.920949	0.932806	0.98419	0.978261	0.974308
Macaca fascicularis	0.963298	0.992095	0.964992	0.952005	0.981366	0.968944	0.945793	0.987013	0.976285	0.968944
Macaca mulatta	0.95	0.997403	0.974675	0.980519	0.888961	0.946753	0.856494	0.985714	0.966883	0.971429
Nomascus leucogenys	0.98	0.995652	0.989565	0.99087	0.983913	0.962174	0.948261	0.99087	0.972609	0.983043
Pan troglodytes	0.971739	0.984348	0.987391	0.973043	0.903043	0.944783	0.961304	0.994783	0.97913	0.974783
Papio anubis	0.957086	0.984754	0.960474	0.987013	0.972897	0.966121	0.937888	0.980237	0.946358	0.952569
Pongo abelli	0.967391	0.967391	0.967391	0.967391	0.967391	0.967391	0.967391	0.967391	0.967391	0.967391

Table 13
Rank correlation coefficient measures of *Gorilla gorilla* genomes versus *Macaca fascicularis*, *Macaca mulatta*, *Nomascus leucogenys*, *Pan troglodytes*, *Papio anubis* and *Pongo abelli* genomes.

Gorilla gorilla (VS)	TAGA	AGAA	GATA	TCTA	TCAT	GAAT	AGAT	CTTT	TATC	TCTG
Macaca fascicularis	0.963636	0.983117	0.950649	0.947403	0.961039	0.971429	0.963636	0.996753	0.936364	0.984416
Macaca mulatta	0.984962	0.996992	0.984211	0.972932	0.869925	0.909774	0.966165	0.996241	0.915038	0.981203
Nomascus leucogenys	0.966165	0.996992	0.977444	0.972932	0.980451	0.969925	0.972932	0.988722	0.95188	0.987218
Pan troglodytes	0.978077	0.985	0.97	0.988462	0.849231	0.986923	0.973077	0.989615	0.965385	0.979615
Papio anubis	0.99026	0.997403	0.953247	0.970779	0.96039	0.964286	0.974675	0.994156	0.951948	0.983117
Pongo abelli	0.963846	0.989615	0.977692	0.989615	0.976538	0.985385	0.980385	0.967308	0.972692	0.952308

Table 14
Rank correlation coefficient measures of *Macaca fascicularis* genomes versus *Macaca mulatta*, *Nomascus leucogenys*, *Pan troglodytes*, *Papio anubis* and *Pongo abelli* genomes.

Macaca fascicularis (VS)	TAGA	AGAA	GATA	TCTA	TCAT	GAAT	AGAT	CTTT	TATC	TCTG
Macaca mulatta	0.978571	0.985714	0.933766	0.959091	0.93961	0.944156	0.924026	0.987013	0.987662	0.987013
Nomascus leucogenys	0.980451	0.978947	0.954887	0.950376	0.97218	0.975188	0.948872	0.986466	0.933835	0.985714
Pan troglodytes	0.981169	0.977922	0.957143	0.918182	0.924675	0.972078	0.961688	0.985065	0.949351	0.965584
Papio anubis	0.977979	0.987013	0.971767	0.966121	0.984754	0.965556	0.971767	0.996612	0.916996	0.988142
Pongo abelli	0.971429	0.983117	0.951299	0.956494	0.973377	0.976623	0.954545	0.998052	0.972078	0.932468

Table 15
Rank correlation coefficient measures of *Macaca mulatta* genomes versus *Nomascus leucogenys*, *Pan troglodytes*, *Papio anubis* and *Pongo abelli* genomes.

Macaca mulatta (VS)	TAGA	AGAA	GATA	TCTA	TCAT	GAAT	AGAT	CTTT	TATC	TCTG
Nomascus leucogenys	0.969298	0.996491	0.968421	0.971053	0.854386	0.913158	0.959649	0.985088	0.874561	0.994737
Pan troglodytes	0.946617	0.942857	0.969925	0.95188	0.932331	0.907519	0.957895	0.983459	0.940602	0.966165
Papio anubis	0.992208	0.998701	0.93961	0.979221	0.937013	0.90974	0.967532	0.997403	0.895455	0.985714
Pongo abelli	0.904511	0.996992	0.944361	0.971429	0.903008	0.890226	0.968421	0.997744	0.966917	0.934586

Table 16
Rank correlation coefficient measures of *Nomascus leucogenys* genomes versus *Pan troglodytes*, *Papio anubis* and *Pongo abelli* genomes.

Nomascus leucogenys (VS)	TAGA	AGAA	GATA	TCTA	TCAT	GAAT	AGAT	CTTT	TATC	TCTG
Pan troglodytes	0.978947	0.941353	0.97218	0.981955	0.871429	0.972932	0.950376	0.986466	0.966917	0.970677
Papio anubis	0.97218	0.996241	0.965414	0.972932	0.957895	0.960902	0.984962	0.981955	0.953383	0.981955
Pongo abelli	0.973684	0.996241	0.96015	0.978195	0.961654	0.973684	0.971429	0.983459	0.978195	0.945113

Table 17
Rank correlation coefficient measures of *Pan troglodytes* genomes versus *Papio anubis* and *Pongo abelli* genomes.

Pan troglodytes (VS)	TAGA	AGAA	GATA	TCTA	TCAT	GAAT	AGAT	CTTT	TATC	TCTG
Papio anubis	0.979221	0.951948	0.974026	0.967532	0.929221	0.966234	0.951948	0.977273	0.951948	0.947403
Pongo abelli	0.978846	0.965385	0.974615	0.978846	0.896538	0.986154	0.954231	0.963077	0.981154	0.975

Table 18
Rank correlation coefficient measures of *Papio anubis* genomes versus *Pongo abelli* genomes.

Papio anubis (VS)	TAGA	AGAA	GATA	TCTA	TCAT	GAAT	AGAT	CTTT	TATC	TCTG
Pongo abelli	0.952597	0.997403	0.953896	0.974675	0.987013	0.966883	0.968182	0.995455	0.937013	0.9

2.3. Cosine similarity

Cosine similarity [13,14] is a measure of similarity between two data sets. The cosine of two sets can be derived by the Euclidean dot product formula as

$$\cos(\theta) = \frac{X.Y}{||X||||Y||} = \frac{\sum_{i=1}^{n} X_i \times Y_i}{\sqrt{\sum_{i=1}^{n} X_i^2}\sqrt{\sum_{i=1}^{n} Y_i^2}} \quad (3)$$

where n is the number of observations, Σ is the summation symbol, X_i is the X value for observation i, Y_i is the Y value for observation i.

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 has been given as input to cosine similarity measure; the measures are shown in Table 19.

Table 19 shows the cosine similarity 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* genomes.

From Table 19, *it is observed that every Tandem Repeat has shown a good relation, and also observed the following relations*:

- AGAA Tandem Repeat has shown a good relation (0.926) between *H. sapiens* and *C. jacchus*, whereas GATA has shown a weak relation(0.608).
- AGAA Tandem Repeat has shown a good relation (0.883) between *H. sapiens* and *C. sabaeus*, whereas AGAT has shown a weak relation (0.662).
- TAGA Tandem Repeat has shown a good relation (0.866) between *H. sapiens* and *G. gorilla*, whereas AGAT has shown a weak relation (0.567).
- AGAA Tandem Repeat has shown a good relation (0.905) between *H. sapiens* and *Macaca fascicularis*, whereas AGAT has shown a weak relation (0.562).
- AGAA Tandem Repeat has shown a good relation (0.942) between *H. sapiens* and *M. mulatta*, whereas TCAT has shown a weak relation (0.594).
- AGAA Tandem Repeat has shown a good relation (0.968) between *H. sapiens* and *N. leucogenys*, whereas TAGA has shown a weak relation (0.586).
- CTTT Tandem Repeat has shown a good relation (0.847) between *H. sapiens* and *Pan troglodytes*, whereas GATA has shown a weak relation (0.666).
- AGAA Tandem Repeat has shown a good relation (0.944) between *H. sapiens* and *P. anubis*, whereas GATA has shown a weak relation(0.556).
- AGAA Tandem Repeat has shown a good relation (0.946) between *H. sapiens* and *pongo abelli*, whereas TCTA has shown a weak relation (0.498).

Inference. The overall highest value 0.968 occurred at AGAA Tandem Repeat of *N. leucogenys* shows a good relation between the sets of *H. sapiens* and *N. leucogenys*.

Tables 20, 21, 22, 23, 24, 25, 26 and 27 have shown the *cosine similarity measures* among the different genome data sets. Observations which are very similar to those from Table 19 can also be made from the other Tables 20,21,22, 23, 24, 25, 26, and 27. Some of the observations are:

- The highest value 0.919 corresponding to AGAA Tandem Repeat of *P. abelli* from the Table 20 shows a good relation between the sets of *C. jacchus* and *P. abelli*.
- The highest value 0.910 corresponding to CTTT Tandem Repeat of *N. leucogenys* from the Table 21 shows a good relation between the sets of *C. sabaeus* and *N. leucogenys*.
- The highest value 0.929 corresponding to AGAA Tandem Repeat of *N. leucogenys* from the Table 22 shows a good relation between the sets of *G. gorilla* and *N. leucogenys*.
- The highest value 0.979 corresponding to CTTT Tandem Repeat of *M. mulatta* from the Table 23 shows a good relation between the sets of *M. fascicularis* and *M. mulatta*.
- The highest value 0.962 corresponding to AGAA Tandem Repeat of *P. anubis* from the Table 24 shows a good relation between the sets of *M. mulatta* and *P. anubis*.
- The highest value 0.910 corresponding to AGAA Tandem Repeat of *P. anubis* and *P. abelli* from the Table 25 shows a good relation between the sets of *N. leucogenys*, *P. anubis* and *P.abelli*.
- The highest value 0.910 corresponding to AGAA Tandem Repeat of *P. anubis* from the Table 26 shows a good relation between the sets of *P. troglodytes* and *P. anubis*.
- The highest value 0.925 corresponding to TCTA Tandem Repeat of *P. abelli* from the Table 27 shows a good relation between the sets of *P. anubis* and *P. abelli*.

Table 19
Cosine similarity 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* genomes.

Homo sapiens (VS)	TAGA	AGAA	GATA	TCTA	TCAT	GAAT	AGAT	CTTT	TATC	TCTG
Callithrix jacchus	0.790569	0.926302	0.608522	0.822881	0.77594	0.820254	0.668472	0.796715	0.850781	0.768615
Chlorocebus sabaeus	0.73095	0.883194	0.723515	0.680545	0.714435	0.698323	0.662651	0.771503	0.753182	0.668943
Gorilla gorilla	0.866169	0.857403	0.805146	0.702474	0.818765	0.696725	0.567734	0.704714	0.674307	0.81657
Macaca fascicularis	0.650203	0.905204	0.606275	0.873273	0.60911	0.776316	0.56296	0.87519	0.858116	0.859676
Macaca mulatta	0.867461	0.942809	0.68973	0.790756	0.594661	0.744622	0.768633	0.872797	0.801784	0.860689
Nomascus leucogenys	0.586952	0.968963	0.690543	0.692308	0.643172	0.684579	0.784157	0.84591	0.696867	0.801784
Pan troglodytes	0.726658	0.806255	0.666597	0.72501	0.666067	0.639094	0.701358	0.847566	0.785118	0.808019
Papio anubis	0.763381	0.944911	0.55688	0.810122	0.645497	0.863294	0.727518	0.887262	0.745054	0.906493
Pongo abelli	0.808176	0.946864	0.75371	0.498557	0.559773	0.88632	0.828177	0.85769	0.702439	0.7534

Table 20
Cosine similarity measures of *Callithrix jacchus* genomes versus *Chlorocebus sabaeus*, *Gorilla gorilla*, *Macaca fascicularis*, *Macaca mulatta*, *Nomascus leucogenys*, *Pan troglodytes*, *Papio anubis* and *Pongo abelli* genomes.

Callithrix jacchus (VS)	TAGA	AGAA	GATA	TCTA	TCAT	GAAT	AGAT	CTTT	TATC	TCTG
Chlorocebus sabaeus	0.851041	0.881953	0.646811	0.708813	0.796599	0.767146	0.674679	0.736242	0.828775	0.769682
Gorilla gorilla	0.837957	0.826153	0.635438	0.763559	0.730437	0.739975	0.703337	0.652438	0.646129	0.717975
Macaca fascicularis	0.776493	0.865768	0.666252	0.829515	0.700071	0.850842	0.652789	0.8141	0.741059	0.712274
Macaca mulatta	0.847174	0.904534	0.737931	0.742611	0.662729	0.684532	0.748598	0.811107	0.736571	0.722718
Nomascus leucogenys	0.878945	0.889898	0.722544	0.677354	0.7542	0.862116	0.723434	0.881662	0.666541	0.790981
Pan troglodytes	0.792183	0.830336	0.675053	0.728881	0.608675	0.65467	0.69786	0.67868	0.876223	0.711305
Papio anubis	0.832495	0.907485	0.765345	0.683599	0.775203	0.74784	0.697669	0.841879	0.771757	0.749613
Pongo abelli	0.737014	0.919145	0.696358	0.627494	0.731126	0.807957	0.672312	0.818881	0.68206	0.650523

Table 21
Cosine similarity measures of *Chlorocebus sabaeus* genomes versus *Gorilla gorilla*, *Macaca fascicularis*, *Macaca mulatta*, *Nomascus leucogenys*, *Pan troglodytes*, *Papio anubis* and *Pongo abelli* genomes.

Chlorocebus sabaeus (VS)	TAGA	AGAA	GATA	TCTA	TCAT	GAAT	AGAT	CTTT	TATC	TCTG
Gorilla gorilla	0.791563	0.814174	0.724469	0.673909	0.771869	0.605705	0.472456	0.73605	0.81478	0.713223
Macaca fascicularis	0.620089	0.815591	0.69455	0.78072	0.909416	0.735456	0.532952	0.853992	0.748534	0.690281
Macaca mulatta	0.730286	0.854242	0.730159	0.721117	0.639382	0.629416	0.657571	0.85042	0.732143	0.720577
Nomascus leucogenys	0.764996	0.783604	0.795472	0.838235	0.809963	0.677192	0.743919	0.910877	0.67465	0.7396
Pan troglodytes	0.692144	0.777018	0.785646	0.662503	0.681598	0.643622	0.661892	0.723627	0.759369	0.630437
Papio anubis	0.780443	0.855908	0.607551	0.810191	0.784063	0.605473	0.765513	0.80226	0.734117	0.767217
Pongo abelli	0.691808	0.859602	0.674541	0.629524	0.778604	0.68156	0.628542	0.839839	0.739375	0.683537

Table 22
Cosine similarity measures of *Gorilla gorilla* genomes versus *Macaca fascicularis*, *Macaca mulatta*, *Nomascus leucogenys*, *Pan troglodytes*, *Papio anubis* and *Pongo abelli* genomes.

Gorilla gorilla (VS)	TAGA	AGAA	GATA	TCTA	TCAT	GAAT	AGAT	CTTT	TATC	TCTG
Macaca fascicularis	0.682863	0.884538	0.540875	0.776701	0.71808	0.728652	0.572434	0.92387	0.657018	0.790569
Macaca mulatta	0.860858	0.923077	0.792747	0.670078	0.69739	0.744352	0.568574	0.921512	0.578399	0.752512
Nomascus leucogenys	0.689134	0.929284	0.66519	0.750306	0.780671	0.737417	0.509615	0.783349	0.480125	0.781408
Pan troglodytes	0.825765	0.702731	0.608845	0.869048	0.701742	0.793107	0.568802	0.761979	0.531234	0.806872
Papio anubis	0.842651	0.925926	0.536175	0.602911	0.726844	0.702112	0.525888	0.870388	0.568787	0.852803
Pongo abelli	0.847671	0.873885	0.697136	0.611887	0.703211	0.795949	0.568473	0.793364	0.548877	0.755865

Table 23
Cosine similarity measures of *Macaca fascicularis* genomes versus *Macaca mulatta*, *Nomascus leucogenys*, *Pan troglodytes*, *Papio anubis* and *Pongo abelli* genomes.

Macaca fascicularis (VS)	TAGA	AGAA	GATA	TCTA	TCAT	GAAT	AGAT	CTTT	TATC	TCTG
Macaca mulatta	0.757033	0.952579	0.719847	0.861858	0.71202	0.86069	0.780721	0.979958	0.781947	0.823532
Nomascus leucogenys	0.666717	0.867149	0.749785	0.882523	0.902698	0.775528	0.628906	0.875936	0.720838	0.79758
Pan troglodytes	0.686803	0.729204	0.722716	0.84678	0.686352	0.629253	0.572883	0.830868	0.734358	0.759072
Papio anubis	0.744529	0.95403	0.737822	0.815374	0.786357	0.652063	0.8	0.920634	0.834415	0.900284
Pongo abelli	0.655447	0.884538	0.649184	0.598764	0.677208	0.773021	0.745995	0.942809	0.699395	0.770675

Table 24
Cosine similarity measures of *Macaca mulatta* genomes versus *Nomascus leucogenys*, *Pan troglodytes*, *Papio anubis* and *Pongo abelli* genomes.

Macaca mulatta (VS)	TAGA	AGAA	GATA	TCTA	TCAT	GAAT	AGAT	CTTT	TATC	TCTG
Nomascus leucogenys	0.666667	0.92	0.759737	0.736024	0.698072	0.727324	0.775404	0.874383	0.682732	0.924785
Pan troglodytes	0.901296	0.765958	0.745356	0.679873	0.599265	0.60455	0.762713	0.826752	0.790277	0.776736
Papio anubis	0.81776	0.962963	0.718648	0.754247	0.64515	0.703101	0.816345	0.952579	0.728912	0.893188
Pongo abelli	0.779773	0.923077	0.697374	0.504772	0.513744	0.726345	0.803739	0.94054	0.78009	0.825029

Table 25
Cosine similarity measures of *Nomascus leucogenys* genomes versus *Pan troglodytes*, *Papio anubis* and *Pongo abelli* genomes.

Nomascus leucogenys (VS)	TAGA	AGAA	GATA	TCTA	TCAT	GAAT	AGAT	CTTT	TATC	TCTG
Pan troglodytes	0.655186	0.754298	0.748455	0.824958	0.796243	0.659346	0.673909	0.707396	0.717218	0.708064
Papio anubis	0.71294	0.910446	0.80111	0.70548	0.743937	0.673451	0.857931	0.875755	0.634733	0.850923
Pongo abelli	0.602534	0.910446	0.717765	0.593848	0.701561	0.777245	0.731263	0.849208	0.828775	0.781918

Table 26
Cosine similarity measures of *Pan troglodytes* genomes versus *Papio anubis* and *pongo abelli* genomes.

Pan troglodytes (VS)	TAGA	AGAA	GATA	TCTA	TCAT	GAAT	AGAT	CTTT	TATC	TCTG
Papio anubis	0.884779	0.866025	0.855337	0.74885	0.652041	0.597355	0.617213	0.778904	0.75963	0.837436
Pongo abelli	0.689658	0.79339	0.715203	0.583562	0.598149	0.733333	0.659456	0.818392	0.740033	0.672977

Table 27
Cosine similarity measures of *Papio anubis* genomes versus *Pongo abelli* genomes.

Papio anubis (VS)	TAGA	AGAA	GATA	TCTA	TCAT	GAAT	AGAT	CTTT	TATC	TCTG
Pongo abelli	0.70274	0.925926	0.669746	0.501648	0.850758	0.84046	0.712389	0.8981	0.661989	0.835053

2.3.1. Purpose of the research

To perform a DNA analysis, DNA is first extracted from a sample. Just one nano-gram of DNA is usually a sufficient quantity to provide good data. In order to match the two DNA sequences, for example, theft evidence to a suspect, a string matching algorithm would search the allele of the 10 STRs [15] for both the evidence sample and the suspect's sample, data base is prepared. If Suspect A is the source of theft sample and Suspect B is in other side, then the similarity between the evidence and suspect is measured from the extracted data with database. This similarity value tells the similarity between A and B. Basing on the resultant values the decision will be taken.

3. Conclusions

This study measures the similarity between the *Homo sapiens* and monkeys by considering correlation coefficient, rank correlation coefficient and cosine similarity. From the Tables 1, 10 and 19, the linear increasing relationship for all the considered similarity measures can be observed. It is also observed that monkeys have a close correlation with *H. sapiens*.

References

[1] M. Costanzo, A. Baryshnikova, J. Bellay, Y. Kim, E.D. Spear, et al., The genetic landscape of a cell. Science (2010) 425–431.
[2] J. Bellay, G. Atluri, T.L. Sing, K. Toufighi, M. Costanzo, et al., Putting genetic interactions in context through a global modular decomposition. Genome Res. (2011) 1375–1387.
[3] L. Avery, S. Wasserman, Ordering gene function: the interpretation of epistasis in regulatory hierarchies. Trends Genet. (1992) 312–316.
[4] R. Mani, R.P. St Onge, J.L. Hartman, G. Giaever, F.P. Roth, Defining genetic interaction. Proc. Natl. Acad. Sci. (2008) 3461–3466.
[5] C.J. Ryan, A. Roguev, K. Patrick, J. Xu, H. Jahari, et al., Hierarchical modularity and the evolution of genetic interactions across species. Mol. Cell (2012) 691–704.
[6] A. Typas, R.J. Nichols, D.A. Siegele, M. Shales, S.R. Collins, B. Lim, H. Braberg, et al., High-throughput, quantitative analyses of genetic interactions in *E. coli*. *Nat. Methods* (2008) 781–787.
[7] B. Lehner, C. Crombie, J. Tischler, A. Fortunato, A.G. Fraser, Systematic mapping of genetic interactions in *Caenorhabditis elegans* identifies common modifiers of diverse signaling pathways. *Nat. Genet.* (2006) 896–903.
[8] http://www.ncbi.nlm.nih.gov/.
[9] Pearson, Karl., "Notes on the history of correlation", Biometrika, pp.25–45.
[10] P.Y. Chen, P.M. Popovich, Correlation: Parametric and Nonparametric Measures. Sage, 2002 137–139.
[11] Z. Govindarajulu, Rank correlation methods. Technometrics (1992) 108.
[12] P. Bobko, Correlation and Regression: Applications for Industrial Organizational Psychology and Management. Sage Publications, 2001.
[13] G. Sidorov, A. Gelbukh, H. Gómez-Adorno, D. Pinto, Soft similarity and soft cosine measure: similarity of features in vector space model. Comput. Syst. (2014) 491–504.
[14] B. Li, L. Han, Distance Weighted Cosine Similarity Measure for Text Classification. Intelligent Data Engineering and Automated Learning 2013, pp. 611–618.
[15] K. Norrgard, Forensics, DNA Fingerprinting, and CODIS. Nature Education, no. 1, 2008.