Genomics Data 7 (2016) 307–317

Contents lists available at ScienceDirect

Genomics Data

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

Next generation sequencing (NGS) database for tandem repeats with multiple pattern 2°-shaft multicore string matching

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

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

ARTICLE INFO

Article history:
Received 14 December 2015
Received in revised form 15 January 2016
Accepted 27 January 2016
Available online 29 January 2016

Keywords:
NGS
SSR
TandemRepeatDB
Genome
String matching
chromosomes

ABSTRACT

Next generation sequencing (NGS) technologies have been rapidly applied in biomedical and biological research in recent years. To provide the comprehensive NGS resource for the research, in this paper , we have considered 10 loci/codi/repeats TAGA, TCAT, GAAT, AGAT, AGAA, GATA, TATC, CTTT, TCTG and TCTA. Then we developed the NGS Tandem Repeat Database (TandemRepeatDB) for all the chromosomes of Homo sapiens, Callithrix jacchus, Chlorocebus sabaeus, Gorilla gorilla, Macaca fascicularis, Macaca mulatta, Nomascus leucogenys, Pan troglodytes, Papio anubis and Pongo abelii genome data sets for all those locis. We find the successive occurence frequency for all the above 10 SSR (simple sequence repeats) in the above genome data sets on a chromosome-by-chromosome basis with multiple pattern 2° shaft multicore string matching.

1. Introduction

Since the completion of the first human genome sequence, demand for cheaper and faster sequencing methods has increased greatly. This demand has driven the development of second-generation sequencing methods, or next-generation sequencing (NGS). In this paper we developed NGS TandemRepeatDB that stores the successive occurence frequency of SSRs from the considered genomes.

Simple sequence repeats (SSRs) are tandemly repeated DNA sequences found in varying abundance in most genomes [1,2]. These repeats have been extensively used for genetic mapping and population studies [3]. SSRs also provide molecular tools to understand spatial relationships between chromosome segments, which in turn, aid in analyzing temporal relationships between species and genera [4]. In humans about 3% of the genome is occupied by SSRs [5].

The study of repeat frequency and its distribution pattern in the genome is expected to help in understanding their significance. There is accumulating evidence to suggest that SSRs function to regulate gene expression [6,7].

The availability of complete genome sequences for many organisms has made it possible to carry out genome-wide analyses. In the present study we have screened all the chromosomes of *Homo sapiens*, *Callithrix jacchus*, *Chlorocebus sabaeus*, *Gorilla gorilla*, *Macaca fascicularis*, *Macaca mulatta*, *Nomascus leucogenys*, *Pan troglodytes*, *Papio anubis* and *Pongo abelii* [8] and studied the distribution and successive occurrence frequency of TAGA, TCAT, GAAT, AGAT, AGAA, GATA, TATC, CTTT, TCTG, and TCTA locis [9,10].

Earlier, few studies [11–14] have attempted to analyze the distribution of tandem repeats in human like genomes but they are confined to a single or a small set of genomes. This tandem repeat mining helps in understanding and addressing biological questions. It is used in various diverse applications like *DNA finger printing, maternity identification*, *paternity identification*, *theft identification*, *suspect findings*, and *disease identification* [15,16,9,10].

2. Methodology

In the present paper, we constructed the TandemRepeatDB with multiple pattern 2° shaft multicore string matching algorithm. String matching [17–20] is a process of identifying the pattern (P) in a given text (T). In the present paper chromosomes of genomes are considered as text (T) and the locis are considered as patterns. The multiple pattern 2° shaft multi core string matching algorithm searches the multiple patterns concurrently in a single part (2° = 1 shaft) with multi core processors.

* Corresponding author.

http://dx.doi.org/10.1016/j.gdata.2016.01.015

In the TandemRepeatDB construction, the text file and the patterns are read and the patterns are searched in text file (T) with multiple pattern 2° shaft multi core string matching algorithm. If perfect tandem repeat occurs then the successive logic is applied. The successive logic means continuous perfect occurrence of similar tandem repeats. If the successive tandem repeat size >1 then the successive occurrence of tandem repeat information is stored in the database. The database is constructed in MySQL using JAVA. The *TandemRepeatDB* construction process comprises four stages and its complete architecture is shown in Fig. 1.

STAGE I: Reading
- (a) The Text file is read.
- (b) The Pattern set is read.

STAGE II: Searching
- (a) All the patterns from the given set are read and categorized basing on their right most characters.
- (b) One of the patterns from one category is selected.
- (c) The shift_left_to_right (Pm-1) function is applied for shift position.
- (d) The multiple pattern 2° shaft multicore string matching algorithm is selected for searching.

STAGE III: Search results
- (a) If a perfect repeat occurs then successive occurrences are searched

STAGE IV: Storing
- (a) If the perfect successive repeat size > 1 then it is stored in the TandemRepeatDB with the following information.
  - sample_name
  - sample_chromosome_name
  - position
  - noofoccurences
  - codi/repeat name

The multiple pattern 2° shaft multicore string matching algorithm consists input and output, initialization, main function, search function and shift_left_to_right function.

In the input and output the genome sets and patterns are taken as input and the sample_id, sample_name, sample_chromosome_name, position, noofoccurences, and codi are returned as output. In the initialization, multi_pattern (all pattern in the set), n (text length), m (pattern length) and all other variables 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. Once the shift value is received the search function is called for all the patterns. The shift_left_to_right function, applies the shift operation with the pattern's rightmost character and the shift position is returned to main function. The search process compares character by character from both the directions until a complete match or mismatch occurs.

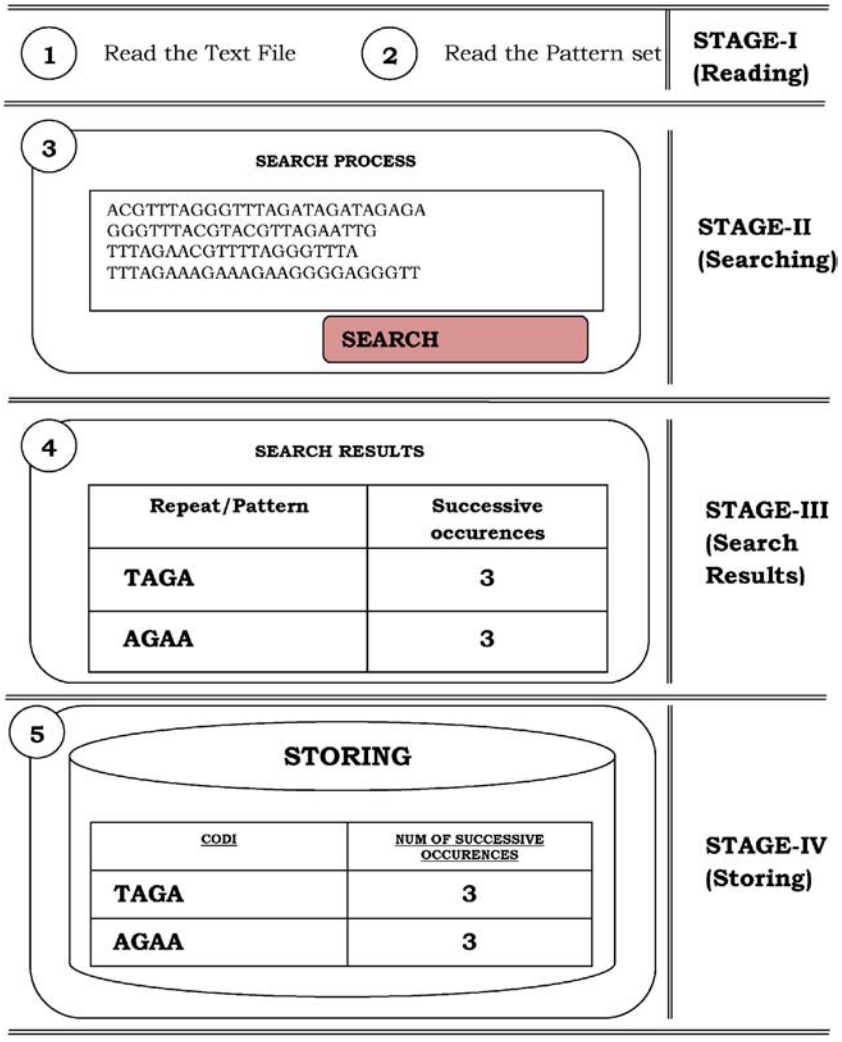

Fig. 1. Architecture of TandemRepeatDB.

If match occurs the successive occurrence of the pattern is searched. If the successive occurrence size is greater than 1 then the data is stored into TandemRepeatDB. If the mismatch or complete match occurs the same procedure is repeated until the end of the T.

Multiple pattern 2^0 shaft multi core string matching algorithm for successive occurrence

Input: *homo sapiens, callithrix jacchus, chlorocebus sabaeus, gorilla gorilla, macaca fascicularis, macaca mulatta, nomascus leucogenys, pan troglodytes, papio anubis*and*pongo abelli genome data sets*
TAGA, TCAT, GAAT, AGAT, AGAA, GATA, TATC, CTTT, TCTG, TCTA patterns
Output: Returns sample_id, sample_name, sample_chromosome_name, lineno, position, noofoccurences and codi

```
/* Initialization */
multi_pattern={"TAGA","AGAA","GATA","TCTA","TCAT",
               "GAAT","AGAT","CTTT","TATC","TCTG"}
n ← T.length, m ← P.length, m1← multi_pattern.length;
i_left ← 0, j ← m-1,k={j-j,j-1,j-(j-1),j-2};
diff_of_two_shafts← 0 , shafts_1←0 , shafts_2←0
/* main function */
for ( File f : genome_files)
begin
        while ((T = f.readLine()) != null)
        T.append(T);
        for i_left←m-1 to n-1 do
        begin
                i_left=shift_left_to_right(T,P,i_left,j,n);
                for i1_left←0 to m1-1 do
                begin
                        count=search(T,multi_pattern[i1_left].toCharArray(),i_left,j,k,count)
                end for;
        end for;
end for;
/* search function */
int search(Char[] T, Char[] P, inti_left, int j, int[] k, int count)
begin
        int j1= j;
        intcomp_pos=0;
        while ( j >=0 &&T[i_left-k[comp_pos]] == P[j-k[comp_pos]] )
        do
                j1--;
                comp_pos ++;
        done ;
        if ( j1 == -1 )
        begin
                diff_of_two_shafts++;
                if((diff_of_two_shafts -(i+1))==-4)
                        shafts_1++
                else
                        shafts_2= shafts_1
              if  shafts_1 > 1
                insert (sample_id, sample_name, sample_chromosome_name, lineno, position,
                noofoccurences, codi)
                position =  " i_left-j";
                count++;
        end;
        return count;
end search;
/* shift_left_to_rightfunction */
intshift_left_to_right (Char[] T, Char[] P, inti_left, int j, int n)
begin
        while(  ( T[i_left] != P[j] ) && (i_left<=n-1) )
        i_left++;
        return i_left;
end shift;
```

2.1. Structure of the database

In this paper a table is created with the given sample name. The table contains sample_id, sample_name, sample_chromosome_name, position indicating the occurrence position of the codi, noofoccurences indicating the number of occurrences of the repeat and codi indicating the name of the repeat. The table structure is shown in Table 1.

Table 1
Table Structure.

Type	Collation
sample_id	text
sample_name	text
sample_chromosome_name	text
position	int(10)
noofoccurences	int(10)
codi	text

Table 2
Genome sequences used in the study.

Genome sequence name	Name & number of chromosomes	Total number of tandem repeats extracted (>1)
Homo sapiens	1 to 22, MT, X, Y and Un (26)	11,99,985
Callithrix jacchus	1 to 22, X, Y and Un (25)	11,40,529
Chlorocebus sabaeus	1 to 29, MT, X, Y and Un (33)	11,13,445
Gorilla gorilla	1, 2A, 2B, 3 to 22, MT, X and Un (26)	11,63,843
Macaca fascicularis	1 to 20, MT, X and Un (23)	12,31,029
Macaca mulatta	1 to 20, MT, X and Un (23)	12,74,556
Nomascus leucogenys	1a 2 to 6, 7b, 8 to 21, 22a, 23 to 25, X and Un (27)	11,71,594
Pan troglodytes	1, 2A, 2B, 3 to 22, MT, X, Y and Un (27)	12,76,766
Papio anubis	1 to 20, MT, X and Un (23)	13,51,393
Pongo abelii	1, 2A, 2B, 3 to 22, MT, X and Un (26)	13,81,887
10	259	**1,23,05,027**

Availability of NGS techniques leads to the accessibility of genome sequences. Studying the perfect successive occurrences of the tandem repeats using Bioinformatics approach would be very interesting and informative.

In the remaining part of the study, perfect tandem repeats of all chromosomes in *H. sapiens*, *C. jacchus*, *C. sabaeus*, *G. gorilla*, *M. fascicularis*, *M. mulatta*, *N. leucogenys*, *P. troglodytes*, *P. anubis* and *P. abelii* genomes are analyzed and a brief note on the successive occurrence frequency of TAGA, TCAT, GAAT, AGAT, AGAA, GATA, TATC, CTTT, TCTG and TCTA repeats have been presented.

In this study 259 chromosomes of *H.sapiens, C.jacchus, C.sabaeus, G.gorilla, M.fascicularis, M.mulatta, N.leucogenys, P.troglodytes, P.anubis* and *P.abelii* genomes have been used as shown in Table 2.

3. Tandem repeat size analysis

In this section, the perfect successive tandem repeats are extracted and analyzed by executing SQL queries on the TandemRepeatDB for all the chromosomes of the considered genomes corresponding to TAGA, TCAT, GAAT, AGAT, AGAA, GATA, TATC, CTTT, TCTG and TCTA repeats.

Notations used in tables:

- First column refers to codi/repeat name.
- Second column refers to MAX number of codi in the successive occurrences.
- Third column refers to number of times the MAX number appeared.

3.1. H. sapiens

H. sapiens are the binomial nomenclature for the human species. *Homo* is the human genus, which also includes Neanderthals and many other extinct species of hominid.

Table 3
Tandem repeat successive occurrences for all chromosomes of *H.sapiens.*

codi/Repeat name	MAX number of codi in the successive occurrences	Number of times the MAX number appeared
TAGA	21	Once
AGAA	42	Twice
GATA	22	Once
TCTA	25	Once
TCAT	12	Twice
GAAT	12	Once
AGAT	21	Once
CTTT	78	Once
TATC	25	Once
TCTG	12	Four Times

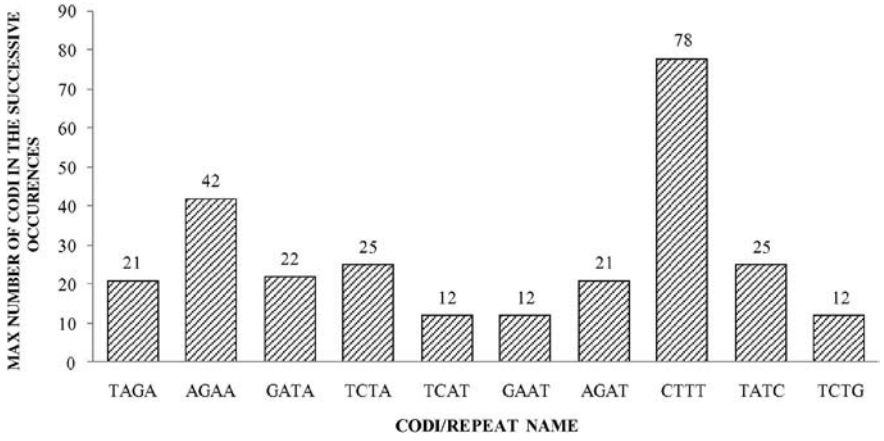

Fig. 2. Max number of successive occurrences of all repeats for all chromosomes of *H.sapiens*.

In the paper, multiple pattern 2° shaft multicore string matching algorithm is used to retrieve the perfect successive tandem repeats from *H. sapiens* genomes which consists *1 to 22, MT, X, Y and Un* chromosomes as shown in Table 2. A total of *11,99,985* perfect successive repeats are extracted from the above chromosomes, which are stored in the *homo_sapiens* table.

Table 3 gives the summary of extracted MAX number of successive occurrences from the *homo_sapiens* table for TAGA, TCAT, GAAT, AGAT, AGAA, GATA, TATC, CTTT, TCTG and TCTA repeats. The *TAGA* results are extracted from the table by executing $Query_1$ and $Query_2$

$$Query_1 = \pi_{\max(noofoccurences)}(\sigma_{codi=TAGA} \quad (homo\_sapiens))$$

$$Query_2 = \pi_{\max(noofoccurences)}(\sigma_{codi=TAGA \wedge noofoccurences=Query1} \quad (homo\_sapiens))$$

$Query_1$ and $Query_2$ are executed for the remaining repeats TCAT, GAAT, AGAT, AGAA, GATA, TATC, CTTT, TCTG and TCTA.

The extracted MAX number of successive occurrences from the *homo_sapiens* table for TAGA, TCAT, GAAT, AGAT, AGAA, GATA, TATC, CTTT, TCTG, and TCTA repeats is graphically shown in Fig. 2.

From the Fig. 2, the following observations can be made:

- CTTT tandem repeat has maximum of 78 successive base pairs,
- AGAA tandem repeat has maximum of 42 successive base pairs twice,
- TCTG tandem repeat has maximum of 12 successive base pairs four times,
- The remaining Tandem repeats have successive base pairs from a minimum of 12 to a maximum of 25,
- All the above observations have a significant role in the bio-informatic studies.

3.2. C. jacchus

The common marmoset is a New World monkey. It originally lived in the Northeastern coast of Brazil.

In the paper, the proposed string matching algorithm is used retrieve the perfect successive tandem repeats from *C. jacchus* genomes which consist *1 to 22, X, Y and Un* chromosomes as shown in Table 2. A total of *11,40,529* perfect successive repeats are extracted from the above chromosomes, which are stored in the *callithrix_jacchus* table.

Table 4 gives the summary of the extracted MAX number of successive occurrences from the *callithrix_jacchus* table for TAGA, TCAT, GAAT, AGAT, AGAA, GATA, TATC, CTTT, TCTG and TCTA repeats.

The *TAGA* results are extracted from the table by executing $Query_1$ and $Query_2$

$$Query_1 = \pi_{\max(noofoccurences)}(\sigma_{codi=TAGA} \quad (callithrix\_jacchus))$$

$$Query_2 = \pi_{\max(noofoccurences)}(\sigma_{codi=TAGA \wedge noofoccurences=Query1} \quad (callithrix\_jacchus))$$

$Query_1$ and $Query_2$ are executed for remaining repeats TCAT, GAAT, AGAT, AGAA, GATA, TATC, CTTT, TCTG and TCTA t.

The extracted MAX number of successive occurrences from the callithrix_jacchus table for TAGA, TCAT, GAAT, AGAT, AGAA, GATA, TATC, CTTT, TCTG, and TCTA repeats is graphically shown in Fig. 3.

From the Fig. 3, the following observations can be made:

- AGAA tandem repeat has maximum of 57 successive base pairs,
- CTTT tandem repeat has maximum of 51 successive base pairs,
- TATC tandem repeat has maximum of 18 successive base pairs thrice,
- Th remaining Tandem repeats have successive base pairs from a minimum of 13 to a maximum of 21,
- All the above observations have a significant role in the bio-informatics studies.

3.3. C. sabaeus

The green monkey, also known as the *sabaeus* monkey or the *Callithrix* monkey, is an Old World monkey with golden-green fur and pale hands and feet. The tip of the tail is golden yellow as are the backs of the thighs and cheek whiskers.

In the paper, the proposed string matching algorithm is used to retrieve the perfect successive tandem repeats from *C. sabaeus* genomes

Table 4
Tandem repeat successive occurrences for all chromosomes of *C.jacchus*.

codi/Repeat name	MAX number of codi in the successive occurrences	Number of times the MAX number appeared
TAGA	21	Once
AGAA	57	Once
GATA	20	Once
TCTA	18	Twice
TCAT	14	Once
GAAT	13	Once
AGAT	21	Once
CTTT	51	Once
TATC	18	Thrice
TCTG	14	Once

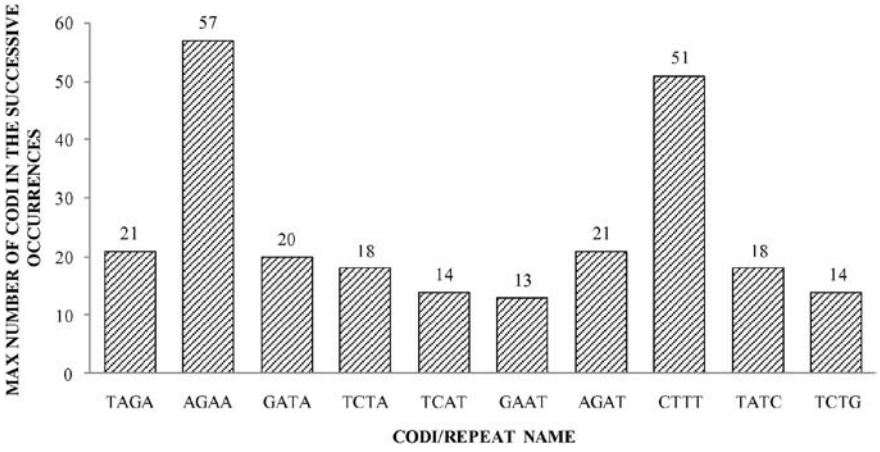

Fig. 3. Max number of successive occurrences of all repeats for all chromosomes of *C.jacchus*.

which consists *1 to 29, MT, X, Y and Un* chromosomes as shown in Table 2. A total of *11,13,445* perfect successive repeats are extracted from the above chromosomes, which are stored in the *chlorocebus_sabaeus* table.

Table 5 gives the summary of extracted MAX number of successive occurrences from the *chlorocebus_sabaeus* table for TAGA, TCAT, GAAT, AGAT, AGAA, GATA, TATC, CTTT, TCTG and TCTA repeats. The *TAGA* results are extracted from the table by executing $Query_1$ and $Query_2$.

$$Query_1 = \pi_{\max(noofoccurences)}(\sigma_{codi=TAGA} \quad (chlorocebus\_sabaeus))$$

$$Query_2 = \pi_{\max(noofoccurences)}(\sigma_{codi=TAGA \wedge noofoccurences=Query1} \quad (chlorocebus\_sabaeus))$$

$Query_1$ and $Query_2$ are executed for remaining repeats TCAT, GAAT, AGAT, AGAA, GATA, TATC, CTTT, TCTG and TCTA.

The extracted MAX number of successive occurrences from the *chlorocebus_sabaeus* table for TAGA, TCAT, GAAT, AGAT, AGAA, GATA, TATC, CTTT, TCTG and TCTA repeats is graphically shown in Fig. 4.

From the Fig. 4, the following observations can be made.

- AGAA tandem repeat has maximum of 54 successive base pairs,
- CTTT tandem repeat has maximum of 42 successive base pairs,
- TCTA tandem repeat has maximum of 20 successive base pairs twice,
- The remaining tandem repeats have successive base pairs from a minimum of 14 to a maximum of 20,
- All the above observations have a significant role in the bio-informatics studies.

3.4. G. gorilla

The western gorilla is a great ape, the type species as well as the most populous species of the genus *Gorilla*.

In the paper, the proposed string matching algorithm is used to retrieve the perfect successive tandem repeats from *G. gorilla* genomes which consists *1, 2A, 2B, 3 to 22, MT, X and Un* chromosomes as shown in Table 2. A total of *11,63,843* perfect successive repeats are extracted from the above chromosomes, which are stored in the *gorilla_gorilla* table.

Table 6 gives the summary of the extracted MAX number of successive occurrences from the *gorilla_gorilla* table for TAGA, TCAT, GAAT, AGAT, AGAA, GATA, TATC, CTTT, TCTG and TCTA repeats. The *TAGA* results are extracted from the table by executing $Query_1$ and $Query_2$.

$$Query_1 = \pi_{\max(noofoccurences)}(\sigma_{codi=TAGA} \quad (gorilla\_gorilla))$$

$$Query_2 = \pi_{\max(noofoccurences)}(\sigma_{codi=TAGA \wedge noofoccurences=Query1} \quad (gorilla\_gorilla))$$

$Query_1$ and $Query_2$ are executed for remaining repeats TCAT, GAAT, AGAT, AGAA, GATA, TATC, CTTT, TCTG and TCTA.

The extracted MAX number of successive occurrences from the *gorilla_gorilla* table for TAGA, TCAT, GAAT, AGAT, AGAA, GATA, TATC, CTTT, TCTG, and TCTA repeats is graphically shown in Fig. 5.

From the Fig. 5, the following observations can be made:

- CTTT tandem repeat has maximum of 66 successive base pairs,
- AGAA tandem repeat has maximum of 41 successive base pairs,
- GAAT tandem repeat has maximum of 12 successive base pairs ten times,
- The remaining tandem repeats have successive base pairs from a minimum of 14 to a maximum of 26,
- All the above observations have a significant role in the bio-informatics studies.

3.5. M. fascicularis

The crab-eating macaque, also known as the long-tailed macaque, is a cercopithecine primate native to Southeast Asia. It is referred to as the cynomolgus monkey in laboratories

In the paper, the proposed string matching algorithm is used to retrieve the perfect successive tandem repeats from *M. fascicularis* genomes which consists *1 to 20, MT, X and Un* chromosomes as shown in Table 2. A total of *12,31,029* perfect successive repeats are extracted from the above chromosomes, which are stored in the *macaca_fascicular* table.

Table 5
Tandem repeat successive occurrences for all chromosomes of *C.sabaeus*.

codi/Repeat name	MAX number of codi in the successive occurrences	Number of times the MAX number appeared
TAGA	19	Twice
AGAA	54	Once
GATA	20	Once
TCTA	20	Twice
TCAT	14	Once
GAAT	14	Once
AGAT	20	Once
CTTT	42	Once
TATC	20	Once
TCTG	14	Once

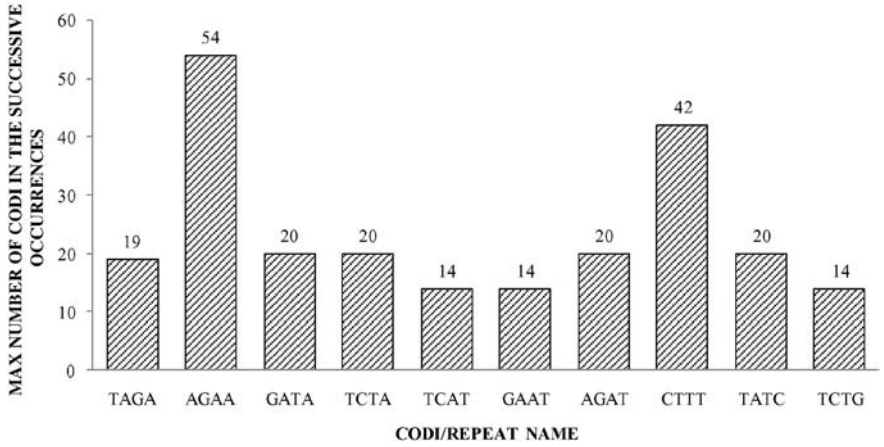

Fig. 4. Max number of successive occurrences of all repeats for all chromosomes of *C.sabaeus*.

Table 7 gives the summary of the extracted MAX number of successive occurrences from the *macaca_fascicularis* table for TAGA, TCAT, GAAT, AGAT, AGAA, GATA, TATC, CTTT, TCTG and TCTA repeats. The *TAGA* results are extracted from the table by executing $Query_1$ and $Query_2$

$$Query_1 = \pi_{\max(noofoccurences)}(\sigma_{codi=\text{TAGA}} \quad (macaca\_fascicularis))$$

$$Query_2 = \pi_{\max(noofoccurences)}(\sigma_{codi=\text{TAGA} \wedge \text{noofoccurences}=Query1} \quad (macaca\_fascicularis))$$

$Query_1$ and $Query_2$ are executed for remaining repeats TCAT, GAAT, AGAT, AGAA, GATA, TATC, CTTT, TCTG and TCTA.

The extracted MAX number of successive occurrences from the *macaca_fascicularis* table for TAGA, TCAT, GAAT, AGAT, AGAA, GATA, TATC, CTTT, TCTG, and TCTA repeats is graphically shown in Fig. 6.

From the Fig. 6, the following observations can be made:

- CTTT tandem repeat has maximum of 221 successive base pairs,
- AGAA tandem repeat has maximum of 218 successive base pairs ,
- GAAT tandem repeat has maximum of 14 successive base pairs thrice,
- The remaining tandem repeats have successive base pairs from a minimum of 16 to a maximum of 33,
- All the above observations have a significant role in the bio-informatics studies.

3.6. M. mulatta

The rhesus macaque (*M. mulatta*), is one of the best-known species of Old World monkeys. It is listed as the least concern in the IUCN Red List of Threatened Species in view of its wide distribution, presumed large population, and its tolerance of a broad range of habitats.

In the paper, the proposed string matching algorithm is used to retrieve the perfect successive tandem repeats from *M. mulatta* genomes which consists *1 to 20, MT, X and Un* chromosomes as shown in Table 2. A total of *12,74,556* perfect successive repeats are extracted from the above chromosomes, which are stored in the *macaca_mulatta* table.

Table 8 gives the summary of the extracted MAX number of successive occurrences from the *macaca_mulatta* table for TAGA, TCAT, GAAT, AGAT, AGAA, GATA, TATC, CTTT, TCTG and TCTA repeats.

The *TAGA* results are extracted from the table by executing $Query_1$ and $Query_2$

$$Query_1 = \pi_{\max(noofoccurences)}(\sigma_{codi=\text{TAGA}} \quad (macaca\_mulatta))$$

$$Query_2 = \pi_{\max(noofoccurences)}(\sigma_{codi=\text{TAGA} \wedge \text{noofoccurences}=Query1} \quad (macaca\_mulatta))$$

$Query_1$ and $Query_2$ are executed for remaining repeats TCAT, GAAT, AGAT, AGAA, GATA, TATC, CTTT, TCTG and TCTA.

The extracted MAX number of successive occurrences from the *macaca_mulatta* table for TAGA, TCAT, GAAT, AGAT, AGAA, GATA, TATC, CTTT, TCTG, and TCTA repeats is graphically shown in Fig. 7.

From the Fig. 7, the following observations can be made:

- AGAA tandem repeat has maximum of 84 successive base pairs,
- CTTT tandem repeat has maximum of 79 successive base pairs,
- TCTA tandem repeat has maximum of 21 successive base pairs twice,
- The remaining tandem repeats have successive base pairs from a minimum of 12 to a maximum of 31,
- All the above observations have a significant role in the bio-informatics studies.

3.7. N. leucogenys

The northern white-cheeked gibbon is a species of gibbon native to South East Asia. It is closely related to the southern white-cheeked gibbon, with which it was previously considered conspecific.

In the paper, the proposed string matching algorithm is used to retrieve the perfect successive tandem repeats from *N. leucogenys* genomes which consists *1 to 6, 7b, 8 to 21, 22a, 23 to 25, X and Un* chromosomes as shown in Table 2. A total of *11,71,594* perfect successive repeats are extracted from the above chromosomes, which are stored in the *nomascus_leucogenys* table.

Table 9 gives the summary of the extracted MAX number of successive occurrences from the *nomascus_leucogenys* table for TAGA, TCAT, GAAT, AGAT, AGAA, GATA, TATC, CTTT, TCTG and TCTA repeats.

Table 6
Tandem repeat successive occurrences for all chromosomes of *G.gorilla*.

codi/Repeat name	MAX number of codi in the successive occurrences	Number of times the MAX number appeared
TAGA	19	Thrice
AGAA	41	Once
GATA	20	Once
TCTA	26	Once
TCAT	14	Once
GAAT	12	Ten times
AGAT	20	Twice
CTTT	66	Once
TATC	26	Once
TCTG	16	Once

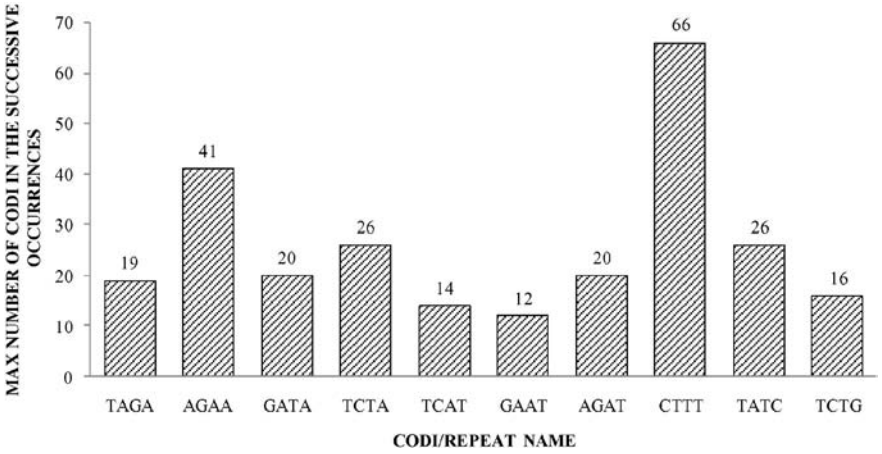

Fig. 5. Max number of successive occurrences of all repeats for all chromosomes of *G.gorilla*.

Table 7
Tandem repeat successive occurrences for all chromosomes of *M.fascicularis*.

codi/Repeat name	MAX number of codi in the successive occurrences	Number of times the MAX number appeared
TAGA	29	Once
AGAA	218	Once
GATA	29	Once
TCTA	33	Once
TCAT	19	Once
GAAT	14	Thrice
AGAT	28	Once
CTTT	221	Once
TATC	33	Once
TCTG	16	Once

The *TAGA* results are extracted from the table by executing $Query_1$ and $Query_2$

$$Query_1 = \pi_{\max(noofoccurences)}(\sigma_{codi=\text{TAGA}} \quad (nomascus\_leucogenys))$$

$$Query_2 = \pi_{\max(noofoccurences)}(\sigma_{codi=\text{TAGA} \wedge \text{noofoccurences}=Query1} \quad (nomascus\_leucogenys))$$

$Query_1$ and $Query_2$ are executed for remaining repeats TCAT, GAAT, AGAT, AGAA, GATA, TATC, CTTT, TCTG and TCTA.

The extracted MAX number of successive occurrences from the *nomascus_leucogenys* table for TAGA, TCAT, GAAT, AGAT, AGAA, GATA, TATC, CTTT, TCTG, and TCTA repeats is graphically shown in Fig. 8.

From the Fig. 8, the following observations can be made:

- AGAA tandem repeat has maximum of 52 successive base pairs,
- CTTT tandem repeat has maximum of 33 successive base pairs,
- TCTG tandem repeat has maximum of 17 successive base pairs thrice,
- The remaining tandem repeats have successive base pairs from a minimum of 11 to a maximum of 23,
- All the above observations have a significant role in the bio-informatics studies.

3.8. P. troglodytes

The common chimpanzee (*P. troglodytes*), also known as the robust chimpanzee, is a species of great apes. Colloquially, the common chimpanzee is often called the chimpanzee.

In the paper, the proposed string matching algorithm is used to retrieve the perfect successive tandem repeats from *P. troglodytes* genomes which consists *1, 2A, 2B, 3 to 22, MT, X, Y and Un* chromosomes as shown in Table 2. A total of *12,76,766* perfect successive repeats are extracted from the above chromosomes, which are stored in the *pan_troglodytes* table.

Table 10 gives the summary of the extracted MAX number of successive occurrences from the *pan_troglodytes* table for TAGA, TCAT, GAAT,

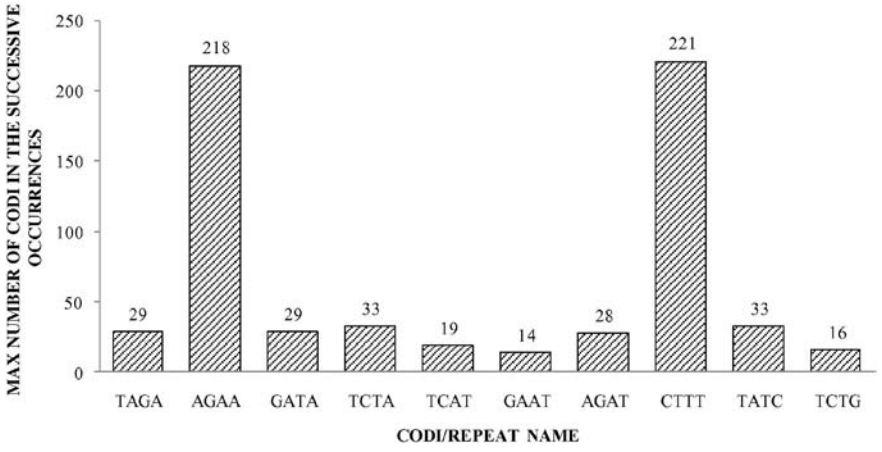

Fig. 6. Max number of successive occurrences of all repeats for all chromosomes of *M.fascicularis*.

Table 8
Tandem repeat successive occurrences for all chromosomes of *M.mulatta*.

codi/Repeat name	MAX number of codi in the successive occurrences	Number of times the MAX number appeared
TAGA	31	Once
AGAA	84	Once
GATA	31	Once
TCTA	21	Twice
TCAT	19	Once
GAAT	15	Once
AGAT	31	Once
CTTT	79	Once
TATC	21	Once
TCTG	12	Twice

Table 9
Tandem repeat successive occurrences for all chromosomes of *N.leucogenys*.

codi/Repeat name	MAX number of codi in the successive occurrences	Number of times the MAX number appeared
TAGA	17	Thrice
AGAA	52	Once
GATA	17	Once
TCTA	22	Once
TCAT	12	Once
GAAT	11	Once
AGAT	17	Once
CTTT	33	Once
TATC	23	Once
TCTG	13	Once

AGAT, AGAA, GATA, TATC, CTTT, TCTG and TCTA repeats. The *TAGA* results are extracted from the table by executing $Query_1$ and $Query_2$

$$Query_1 = \pi_{\max(noofoccurences)}(\sigma_{codi=\text{TAGA}} \quad (pan\_troglodytes))$$

$$Query_2 = \pi_{\max(noofoccurences)}(\sigma_{codi=\text{TAGA}\ \Lambda\ \text{noofoccurences}=\text{Query1}} \quad (pan\_troglodytes))$$

$Query_1$ and $Query_2$ are executed for remaining repeats TCAT, GAAT, AGAT, AGAA, GATA, TATC, CTTT, TCTG and TCTA.

The extracted MAX number of successive occurrences from the pan_troglodytes table for TAGA, TCAT, GAAT, AGAT, AGAA, GATA, TATC, CTTT, TCTG, and TCTA repeats is graphically shown in Fig. 9.

From the Fig. 9, the following observations can be made.

- AGAA tandem repeat has maximum of 43 successive base pairs,
- CTTT tandem repeat has maximum of 30 successive base pairs,
- TCAT tandem repeat has maximum of 10 successive base pairs five times,

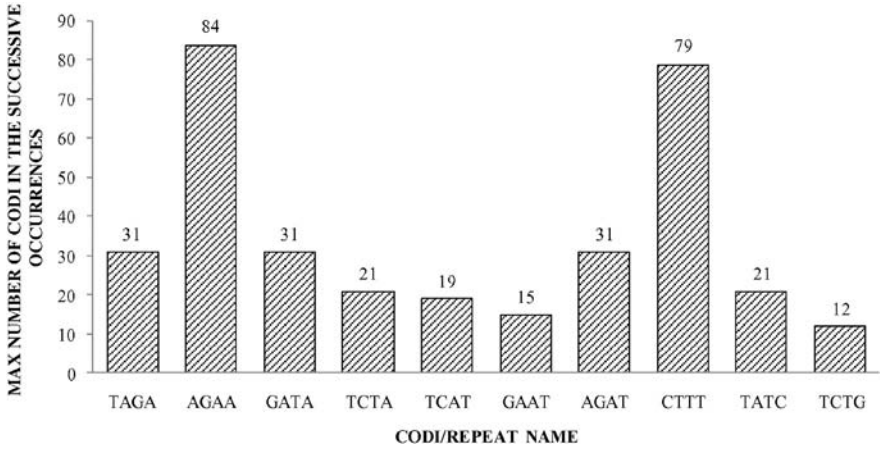

Fig. 7. Max number of successive occurrences of all repeats for all chromosomes of *M.mulatta*.

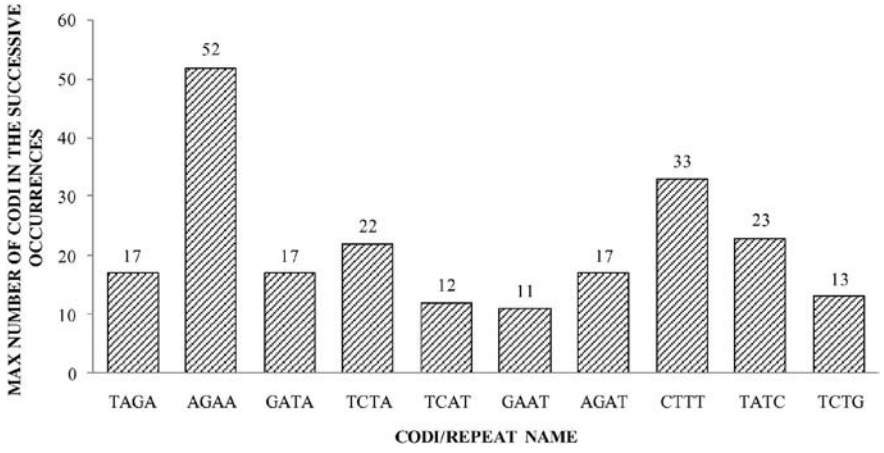

Fig. 8. Max number of successive occurrences of all repeats for all chromosomes of *N.leucogenys*.

Table 10
Tandem repeat successive occurrences for all chromosomes of *P.troglodytes*.

codi/Repeat name	MAX number of codi in the successive occurrences	Number of times the MAX number appeared
TAGA	17	Four Times
AGAA	43	Once
GATA	18	Twice
TCTA	18	Once
TCAT	10	Five Times
GAAT	11	Once
AGAT	18	Once
CTTT	30	Once
TATC	19	Once
TCTG	13	Once

Table 11
Tandem repeat successive occurrences for all chromosomes of *P.anubis*.

codi/Repeat name	MAX number of codi in the successive occurrences	Number of times the MAX number appeared
TAGA	31	Once
AGAA	54	Once
GATA	31	Once
TCTA	22	Once
TCAT	15	Once
GAAT	14	Once
AGAT	32	Once
CTTT	47	Twice
TATC	21	Twice
TCTG	15	Once

- The remaining tandem repeats have successive base pairs from a minimum of 11 to a maximum of 19,
- All the above observations have a significant role in the bio-informatics studies.

3.9. P. anubis

The olive baboon, also called the Anubis baboon, is a member of the family Cercopithecidae. The species is the most widely ranging of all baboons.

In the paper, the proposed string matching algorithm is used to retrieve the perfect successive tandem repeats from *P. anubis* genomes which consists *1 to 20, MT, X and Un* chromosomes as shown in Table 2. A total of *13,51,393* perfect successive repeats are extracted from the above chromosomes, which are stored in the *papio_anubis* table.

Table 11 gives the summary of the extracted MAX number of successive occurrences from the *papio_anubis* table for TAGA, TCAT, GAAT, AGAT, AGAA, GATA, TATC, CTTT, TCTG and TCTA repeats. The *TAGA* results are extracted from the table by executing $Query_1$ and $Query_2$

$$Query_1 = \pi_{\max(noofoccurences)}(\sigma_{codi=TAGA} \quad (papio\_anubis))$$

$$Query_2 = \pi_{\max(noofoccurences)}(\sigma_{codi=TAGA \wedge noofoccurences=Query1} \quad (papio\_anubis))$$

$Query_1$ and $Query_2$ are executed for remaining repeats TCAT, GAAT, AGAT, AGAA, GATA, TATC, CTTT, TCTG and TCTA.

The extracted MAX number of successive occurrences from the *papio_anubis* table for TAGA, TCAT, GAAT, AGAT, AGAA, GATA, TATC, CTTT, TCTG, and TCTA repeats is graphically shown in Fig. 10.

From the Fig. 10, the following observations can be made.

- AGAA tandem repeat has maximum of 54 successive base pairs,
- CTTT tandem repeat has maximum of 47 successive base pairs twice,
- TCTG tandem repeat has maximum of 21 successive base pairs twice,
- The remaining tandem repeats have successive base pairs from a minimum of 14 to a maximum of 32,
- All the above observations have a significant role in the bio-informatics studies.

3.10. P. abelii

The Sumatran orangutan is one of the two species of orangutans. Found only in the island of Sumatra, in Indonesia, it is rarer than the Bornean orangutan. In the paper, the proposed string matching algorithm is used to retrieve the perfect successive tandem repeats from *P. abelii* genomes which consists *1, 2A, 2B, 3 to 22, MT, X and Un* chromosomes as shown in Table 2. A total of *13,81,887* perfect successive repeats are extracted from the above chromosomes, which are stored in the *pongo_abelii* table.

Table 12 gives the summary of extracted MAX number of successive occurrences from the *pongo_abelii* table for TAGA, TCAT, GAAT, AGAT, AGAA, GATA, TATC, CTTT, TCTG and TCTA repeats. The *TAGA* results are extracted from the table by executing $Query_1$ and $Query_2$

$$Query_1 = \pi_{\max(noofoccurences)}(\sigma_{codi=TAGA} \quad (pongo\_abelii))$$

$$Query_2 = \pi_{\max(noofoccurences)}(\sigma_{codi=TAGA \wedge noofoccurences=Query1} \quad (pongo\_abelii))$$

$Query_1$ and $Query_2$ are executed for remaining repeats TCAT, GAAT, AGAT, AGAA, GATA, TATC, CTTT, TCTG and TCTA.

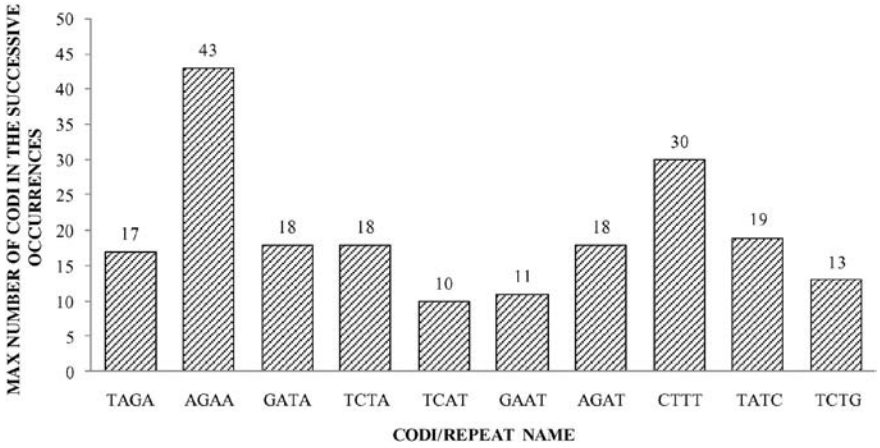

Fig. 9. Max number of successive occurrences of all repeats for all chromosomes of *P.troglodytes*.

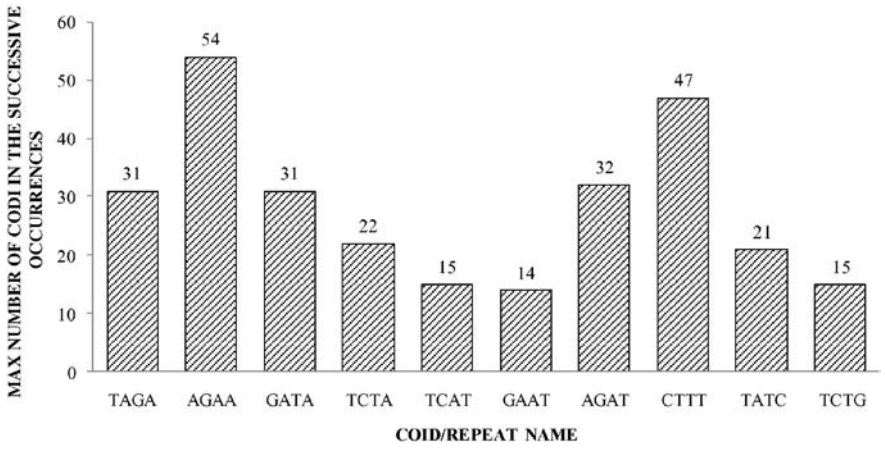

Fig. 10. Max number of successive occurrences of all repeats for all chromosomes of *P.anubis*.

The extracted MAX number of successive occurrences from the *pongo_abelii* table for TAGA, TCAT, GAAT, AGAT, AGAA, GATA, TATC, CTTT, TCTG, and TCTA repeats is graphically shown in Fig. 11.

From the Fig. 11, the following observations can be made.

- CTTT tandem repeat has maximum of 63 successive base pairs,
- AGAA tandem repeat has maximum of 37 successive base pairs,
- TCTA tandem repeat has maximum of 18 successive base pairs twice,
- The remaining tandem repeats have successive base pairs from a minimum of 11 to a maximum of 20,
- All the above observations have a significant role in the bio-informatics studies.

Table 12
Tandem repeat successive occurrences for all chromosomes of *P.abelli*.

codi/Repeat name	MAX number of codi in the successive occurrences	Number of times the MAX number appeared
TAGA	20	Once
AGAA	37	Once
GATA	19	Once
TCTA	18	Twice
TCAT	12	Once
GAAT	13	Once
AGAT	20	Once
CTTT	63	Once
TATC	19	Once
TCTG	11	Twice

4. Conclusions

In this paper we developed the TandemRepeatDB that provides a single portal access to perfect successive repeats in genomes of *H. sapiens*, *C. jacchus*, *C. sabaeus*, *G. gorilla*, *M. fascicularis*, *M. mulatta*, *N. leucogenys*, *P. troglodytes*, *P. anubis* and *P. abelii*. The database is known to be the first of its kind to host all types of perfect successive tandem repeats for the considered genomes. From the analysis of all the records existing in the TandemRepeatDB, it is observed that CTTT tandem repeat and AGAA tandem repeat occupy the major role. This TandemRepeatDB will be a very valuable resource for researchers studying repeats in the above mentioned genomes.

Conflict of interests

Authors did not have any conflict of interests.

References

[1] Gabor Toth, Zoltan Gaspari, Jerzy Jurka, Microsatellites in different eukaryotic genomes: survey and analysis. Genome Res. (7) (2000) 967–981.
[2] Riva Gur-Arie, Cyril J. Cohen, Yuval Eitan, Leora Shelef, Eric M. Hallerman, Yechezkel Kashi, Simple sequence repeats in *Escherichia coli*: abundance, distribution, composition, and polymorphism. Genome Res. (1) (2000) 62–71.
[3] Colette Dib, Sabine Faure, Cecile Fizames, Delphine Samson, Nathalie Drouot, Alain Vignal, Philippe Millasseau, et al., A comprehensive genetic map of the human genome based on 5,264 microsatellites. Nature (6570) (1996) 152–154.
[4] Yechezkel Kashi, David King, Morris Soller, Simple sequence repeats as a source of quantitative genetic variation. Trends Genet. 13 (2) (1997) 74–78.

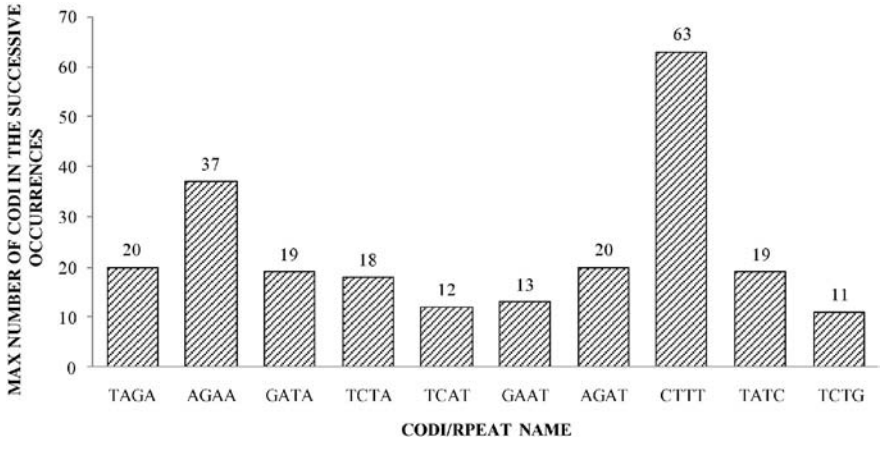

Fig. 11. *Max number of successive occurrences of all repeats for all chromosomes of P.abelii.*

[5] Eric S. Lander, Lauren M. Linton, Bruce Birren, Chad Nusbaum, Michael C. Zody, Jennifer Baldwin, Keri Devon, et al., Initial sequencing and analysis of the human genome. Nature 409 (6822) (2001) 860–921.
[6] P. Kunzler, Koichi Matsuo, Walter Schaffner, Pathological, physiological, and evolutionary aspects of short unstable DNA repeats in the human genome. Biol. Chem. Hoppe Seyler (1995) 201.
[7] Moxon E. Richard, Christopher Wills, DNA microsatellites: agents of evolution? Sci. Am. (1) (1999) 94–99.
[8] http://www.ncbi.nlm.nih.gov.
[9] Karen Norrgard, Forensics, DNA fingerprinting, and CODIS. Nat. Educ. (1) (2008).
[10] Mark A. Jobling, Peter Gill, Encoded evidence: DNA in forensic analysis. Nat. Rev. Genet. (10) (2004) 739–751.
[11] Subbaya Subramanian, Rakesh K. Mishra, Lalji Singh, Genome-wide analysis of microsatellite repeats in humans: their abundance and density in specific genomic regions. Genome Biol. 4 (2) (2013).
[12] Subbaya Subramanian, Vamsi M. Madgula, Ranjan George, Satish Kumar, Madhusudhan W. Pandit, Lalji Singh, SSRD: simple sequence repeats database of the human genome. Comp. Funct. Genomics 4 (3) (2003) 342–345.
[13] T. Boby, A.-M. Patch, S.J. Aves, TRbase: a database relating tandem repeats to disease genes for the human genome. Bioinformatics (6) (2005) 811–816.
[14] Christian M. Ruitberg, Dennis J. Reeder, John M. Butler, STRBase: a short tandem repeat DNA database for the human identity testing community. Nucleic Acids Res. (1) (2001) 320–322.
[15] Richard R. Sinden, Neurodegenerative diseases: origins of instability. Nature (6839) (2001) 757–758.
[16] Christopher J. Cummings, Huda Y. Zoghbi, Fourteen and counting: unraveling trinucleotide repeat diseases. Hum. Mol. Genet. (6) (2000) 909–916.
[17] V.S. Raju, M. Mrudula, Backend engine for parallel string matching using boolean matrix. International Symposium on Parallel Computing in Electrical Engineering 2006, pp. 281–283.
[18] V.S. Raju, K.K.V.V.V.S. Reddy, Recent advancements in parallel algorithms for string matching on computing models—a survey and experimental results. Advanced Computing, Networking and Security, Springer, Berlin Heidelberg 2012, pp. 270–278.
[19] V.S. Raju, A. Vinayababu, Optimal parallel algorithm for string matching on mesh network structure. Int. J. Appl. Math. Sci. (2) (2006) 167–175.
[20] V.S. Raju, A. Vinayababu, Parallel algorithms for string matching problem on single and two-dimensional reconfigurable pipelined bus systems. J. Comput. Sci. (9) (2007) 754–759.