International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2015): 78.96 | Impact Factor (2015): 6.391

Searching and Analyzing for a Different Lengths of Microsatellite Repeats in the Completed Genome of Phoenix Dactylifera Chloroplast, by Using Regular Expression Builder-Language

Mohamed A. Ezz [1], Amal Mahmoud[2,3], Fawzy A. El-Feky[4], Alaa A. Hemeida[2]

[1]Central laboratory for research, October University for Modern Sciences and Arts

[2] Bioinformatics department, Genetic Engineering and Biotechnology Research Institute (GEBRI), University of Sadat City, Egypt

[3]Biology Department, Imam Abdulrahman Bin Faisal University, Saudi Arabia

[4]Biotechnology Department, Faculty of Agriculture, Al-Azhar University, Egypt

Abstract: ***Variations of Simple Sequence Repeats (SSRs) or microsatellites is mainly caused by slipped-strand impairing and resulting errors during DNA replication and recombination. In this study,A new tool called Repeater Finder Regular Expression (RFRE)was programmed by using visual basic language to detect different lengths of sequence repeats in the chloroplast of the completed genome of date palm(Phoenix dactylifera). The user could write many different probabilities of combined operator to retrieve more than one patterns and very specific patterns of repeats. Not only searching automatically for fixed di, tri, tetra penta or hexa repeats but finding the perfect repeat and imperfect repeat. The most abundant founding pattern was (AAAA) {2}. Three long repeats (40 nt) were detected, 2 repeats were located in the noncoding region, however, the 3rd repeat was partially located in the rp3 gene. The program was validated by designing specific primers to target the repeat(8675-8715). The genomic DNA was isolated from date palm, and the repeat 8675-8715 was amplified by PCR. The PCR product was sequenced and the repeat was submitted to DDBJ as microsatellite: EPDCO. In a comparison to date palm isolate and palm isolates from EPDCO was clustered with date palm isolates, EPDCO sequence was clustered to date palm isolates. The microsatelliteEPDCO is a unique marker for Date Palm. This study focused on analyzing the tandem repeats in date palm chloroplast genome by using the core power of the Regex.Engine.***

Keywords: *Phoenix dactylifera*, Regex language engine, SSR, Tandem repeats

1. Introduction

Simple sequence repeats (SSRs), also known as micro-satellites, include tandemly repeated genetic loci of 1 to 6 base pairs (bp) [1]. SSRs are highly abundant and displayvaried levels of polymorphisms in eukaryotic and prokaryotic genomes[2]. They are founding in coding and non-coding regions [2], with SSRs being more richer in noncoding regions than in expressed regions (exons) [3]. Studies have shown that certain trinucleotides are richer in coding regions than in noncoding regions of higher eukaryotic genomes [4]. During DNA replication or recombination, the variation of SSRs is caused by slipped-strand mispairing [5]. The repeat units of sequence polymorphisms have been found in a specific locus, whichresult from insertion or deletion mutation [6]. Regular expressions have studied and experienced a lot of success in both the bioinformatics and natural language processing specialists [7]. Regular expressions operators applied to matching repeats in DNA, which playvery important biological roles and can have a phenotypical effect and makes repeats important molecular markers[7].Also, microsatellites markerswere used to determine the sex of immature date palm [8].In this study, a new-programmed tool, Repeater Finder Regular Expression (RFRE Version 1.0) was created and used to analyze the repeated sequences of date palm (*Phoenix dactylifera*) chloroplast complete genome.

2. Material and Methods

2.1 Retrieving the date palm chloroplast, complete genome

The complete genome of date palm chloroplast (accession number GU811709.2), was retrieved from the core nucleotide database of GenBank (http://www.ncbi.nlm.nih.gov/nuccore) and downloaded in two formats: Fasta and GenBank format file. The GenBank format file converted to a map to annotate the location of the detected repeats by using Unipro UGENE version 1.16.0.

2.2 Finding the repeated sequences of date palmchloroplast by regular expression patterns

A new-programmed tool, Repeater Finder Regular Expression (RFRE Version 1.0)was created and used to analyze the repeated sequences of date palmchloroplast complete genome. The RFRE tool was used to find the repeat lengths, positions of repeat, frequencies of the repeat and the repeats density. Different regular expression patterns were used to find different lengths of repeats [9] (Table 1).

Table 1: Regular expression patterns which were used to find different lengths of repeats

Pattern (Di, Tri- Or Tetra nucleotide)	Pattern (penta nucleotide)	Pattern (hex nucleotides)
(AA){6}*	(AAAAA){2}	(AAAAAG){2}
(AA){7}	(AAAAA){3}	(TAAAAA){2}
(AT){6}	(AAAAT){2}	(GAAAAA){2}
(AT){7}	(AAAAG){2}	
(AT){8}	(TAAAA){2}	
(AT){9}	(GAAAA){2}	
(TA){6}	(CAAAA){2}	
(TA){7}	(ATAAA){2}	
(TA){9}	(AGAAA){2}	
(CC){6}	(ACAAA){2}	
(AAA){4}	(ATAAA){2}	
(AAA){5}	(AGAAA){2}	
(ATA){4}	(AATAA){2}	
(AAT){4}	(AAGAA){2}	
(AAAA){2}	(AACAA){2}	
(ATAA){2}	(AAATA){2}	
(ATAA){3}	(AAAGA){2}	
(ACAA){2}	(AAACA){2}	
(AGAA){2)	(ATCCG){2}	
(AAAC){2}	(AATCC){2}	
(AAAG){2}	(AGCTC){2}	
(AAAT){2}	(AATAT){2}	
(AAGA){2}	(AATAT){2}	
(AATA){2}	(AATAT){2}	
(AATA){3}	(AAGAA){2}	
(ATAA){3}	(AACAA){2}	
(AGAA){2)	(AAATA){2}	
	(AAAGA){2}	
	(AAACA){2}	
	(ATCCG){2}	
	(AATCC){2}	
	(AGCTC){2}	
	(AATAT){2}	
	(AATAT){2}	
	(AAAGG){2}	
	(ACCCG){2}	
	(AAATG){2}	

* (AA){6}match“Di” group (AA)and multiplied6 times

2.3 Validation of the extracted repeat

To validate the extracted date palm repeats, a specific primer was designedthen was tested by PCR. Specific primer was designed to target the repeat 8675-8715, which detected by RFRE tool Ver.1.0. The primerwas designed from the sequence of chloroplast genome (accession number GU811709.2).Leaves samples of date palm (*P. dactylifera*) were collected from 6th October City, Giza. The forward and reverse primers were designed by using Primer BLAST (https://www.ncbi.nlm.nih.gov/tools/primer-BLAST). The primers were designed to produce product of 582 nt where the forward primer was 5'actcagccatctctccccat3' and reverse primer was 5'cccggccagtacttaaacca3'. The PCR master mix Jena bioscience (PCR-101S) was used to make the total volume50 µl per reaction. The PCR cycle steps wererun to be 35xand designedto amplify the targeted repeat where the initial denaturation 94 °C for 1 min, 35 x (denaturation 94 °C for 30 sec, annealing 58°C for 30 sec, elongation 72 °Cfor 30 sec) and the final elongation was 1xat 72 °C for 7min.

2.4 Sequence analysis

The PCR product was cleaned up by gel elution kit of jena bioscience (PP-202S) then sequenced by sanger method. Sequenced repeat was analyzed by running BLAST 2.0 of NCBI to check the degree of the similarity between the subjected repeats and the retrieved repeat by the tool RFRE ver. 1.0. The fast minimum evolution methodwas used to build the tree, the algorithm used the score of pairwise alignment to construct the phylogenetictree.

3. Results and Discussion

3.1 Repeater Finder Regular Expression (RFRE) tool

To find the repeated sequences in chloroplast complete genomeof date palm, a new-programmed tool, Repeater Finder Regular Expression (RFRE Version 1.0) was created (Fig. 1) and a specific regular expression pattern,([agct]{20})\1 was used to detect long sequence repeats, this regular expression pattern is described in Table 2. The other regular expressing patterns are mentioned in Fig. 2. The outputs results can be saved in word file or excel sheet. The tool can be downloaded by sending an email (ezz111@yahoo.com).

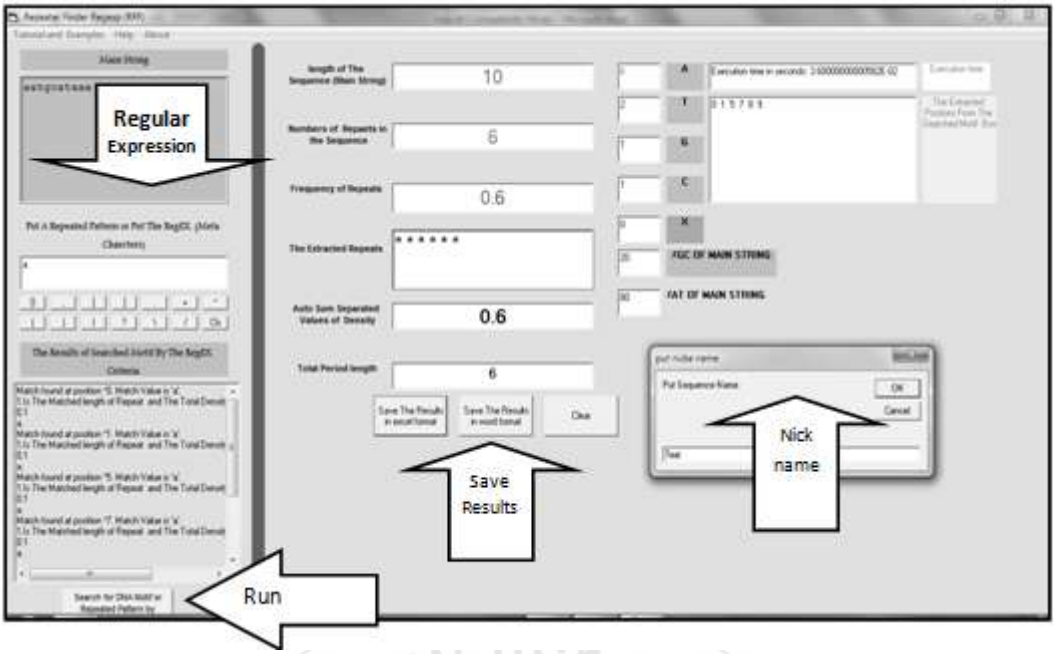

Figure 1: Graphical user interface (GUI) of RFRE tool for Regexp Engine.

Table 2: Description of regular expression search ([agct]{20})\1)

Regular Expression	What it matches
[agct]	A| G| C or T
([agct]{20})\1	A| G| C or T iteration length 20 ntand total repeat length 40nt
([agct]{20})\1	Metacharacter, backreference

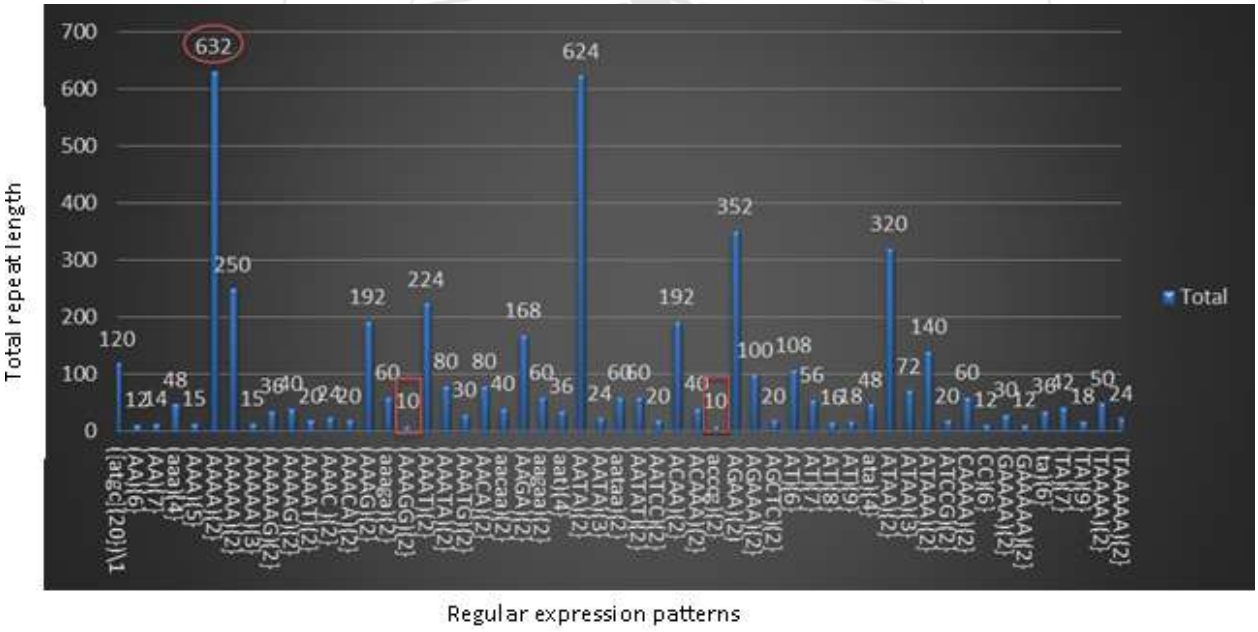

Figure 2: Distribution of repeats across the chloroplast date palm genome

Using the regular expression pattern ([agct]{20})\1), long sequence repeats(40 bp)were detected and summarized in Fig. 2 and Table 3.525short repeats were found,the highest frequent repeats length were found by the following pattern (AAAA){2},where the total length were632 bp. However, the lowest frequent repeats length were found by the patterns (ACCCG){2} and (AAAGG){2} where the total of repeats length were 10 bp for every pattern as shown in Fig. 2.Microsatellite survey of date palm whole nuclear genome shotgun sequences using the developed pipeline detected a total of 166,760 perfect repeats with an average of one SSR per 2.2kb [10]. The microsatellite density profiles in the Arecaceae family showed the predominant occurrence of dinucleotide repeats in the expressed genes of palm members [10].

Table 3: Long repeats extracted and retrieved by using the specific regular expression pattern ([agct]{20})\1

Repeat	Iterations	Start	End	Repeat length / iteration	Extracted Repeat	Regualr expression pattern	Repeat Length
AAAGATATAA GATTATATAA	2	8675	8715	20	AAAGATATAAGATTATATAAA AAGATATAAGATTATATAA	([ATGC]{20})\1	40
TTCATTGCTAC AAATATGGA	2	78458	78498	20	TTCATTGCTACAAATATGGAT TCATTGCTACAAATATGGA	([ATGC]{20})\2	40
CTCGTTTACAA ATATCCAAA	2	85021	85061	20	CTCGTTTACAAATATCCAAAC TCGTTTACAAATATCCAAA	([ATGC]{20})\3	40

3.2 Alignment of the detected sequence repeats against date palm chloroplast genome

The long sequence repeats which were found by the regular expression([agct]{20})\1) were aligned against the annotated date palm genome accession number (GU811709.2) using the program Unipro UGENE ver. 1.16.1. The longest repeats aligned against date palm genome are shown in Figs 3-5. The first 2 repeats8675-8715 and 78458-78498 were located in the non-coding region of the date palm genomeas shown in Fig. 3 and 4, respectively. The third long repeat85021-85061 was partially located in rps3gene (Fig. 5). Rps3 gene is one of three co-transcribed gene clusters–*18S-5S rRNA*, *rps3-rpl16* and *nad3-rps12*–in *P. dactylifera*[11].

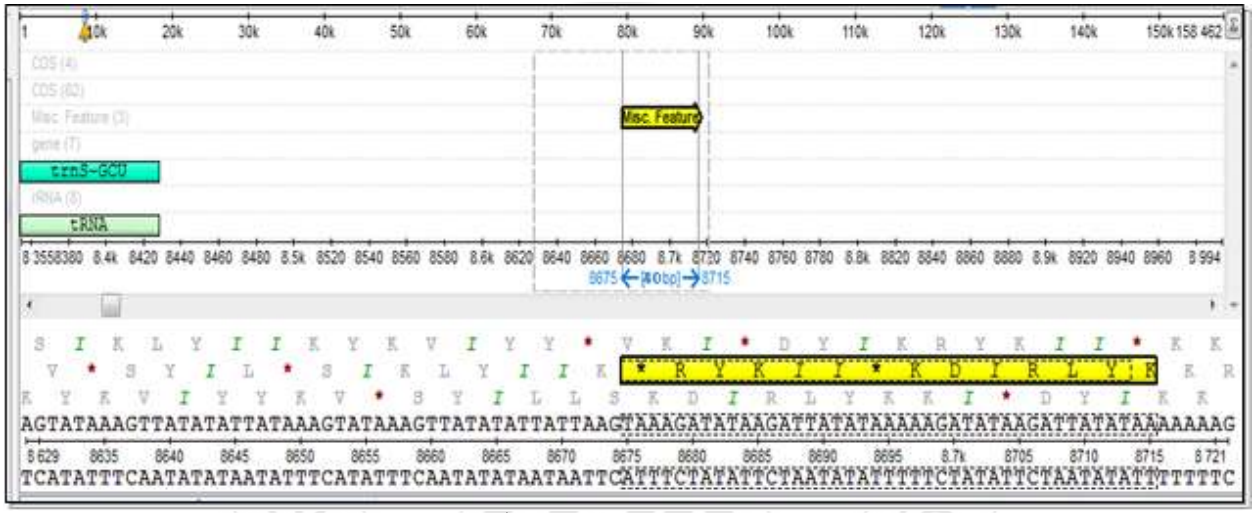

Figure 3: Sequence alignment of the 1st long repeat 8675-8715 against date palm chloroplast genome

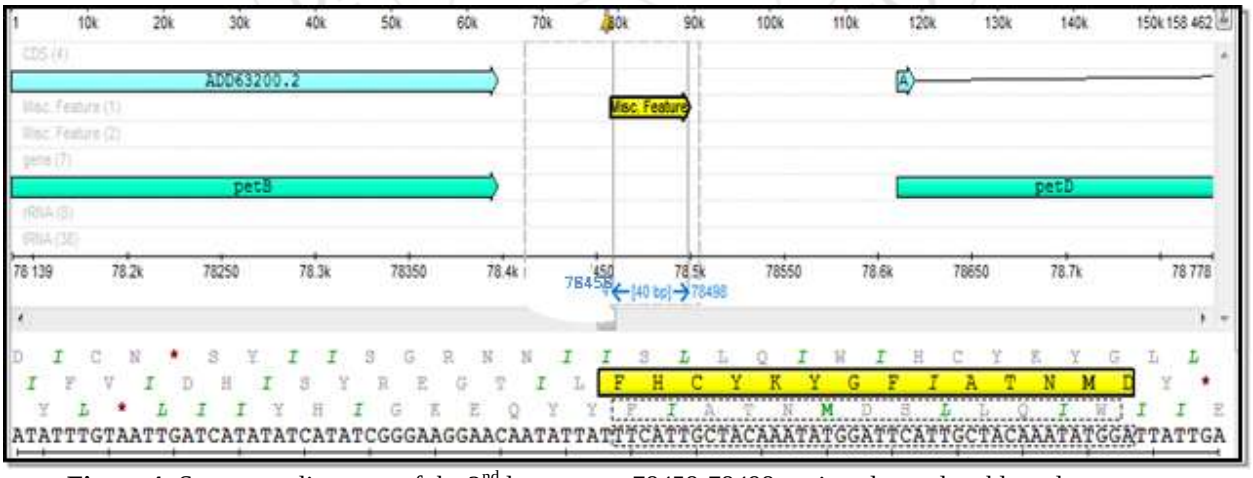

Figure 4: Sequence alignment of the 2nd long repeat 78458-78498 against date palm chloroplast genome

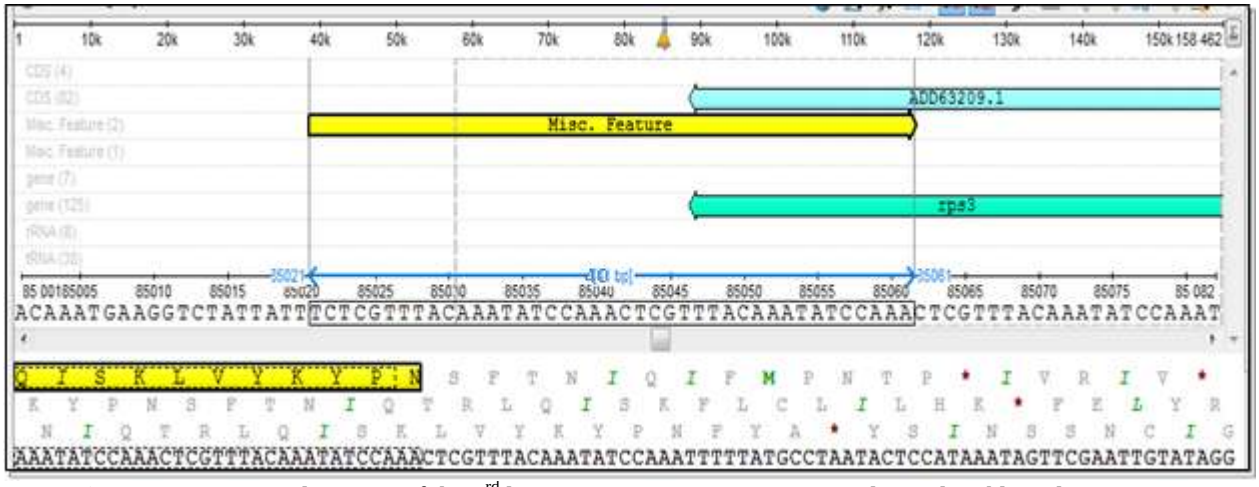

Figure 5: Sequence alignment of the 3rd long repeat 85021-85061 against date palm chloroplast genome

3.3 Validation of the extracted repeats

To validate the extracted repeat 8675-8715, primerswere designed using the retrieved genome sequence (GU811709.1). PCR product of 580bp was produced (Fig. 6),the product was sequenced and submitted to the GenBank of DDBJ (LC202941.1) and annotated as repeat type (rpt_type=tandem), repeat unit rpt_unit_seq="(aaagatataagattatataa)2". The name of the repeat was annotated as a microsatellite: EPDCO". Local alignment was done using EBI (http://www.ebi.ac.uk/Tools/services/rest/emboss_water).The sequence repeat 8675-8715 was 100% similar to date palm genome(Fig. 7).

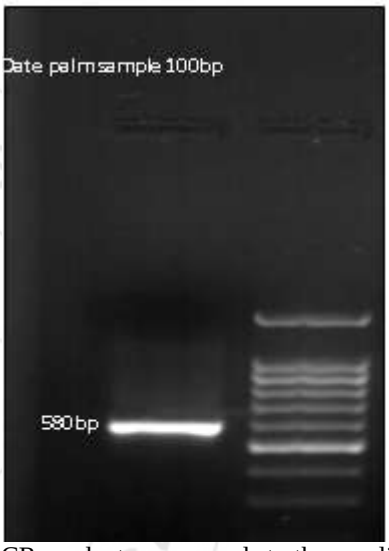

Figure 6: PCR product corresponds to the amplified region include the sequence repeat 8675-8715.

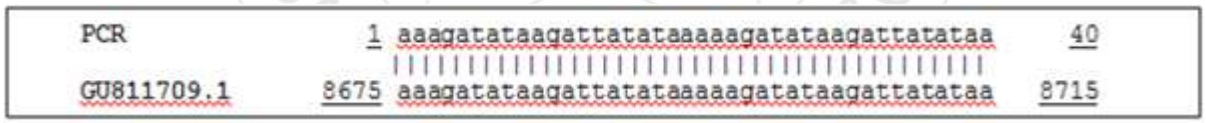

Figure 7: Local alignment between "EPDCO" repeat designed using RFRE tool and that of the recorded chloroplast genome (GU811709.1).

3.4 BLAST tree construction

The sequence repeat (microsatellite "EPDCO") was compared to date palm and palm isolates from subfamily *Coryphoideae*, nucleotide sequence retrieved from GenBank database Two date palm isolates: GU811709.2 (cultivar Khalas: Al-Hssa Oasis, Saudi Arabia) and FJ212316.3 (specimen from Herbarium, Department of Botany, University of Karachi, Pakistan)were 100% identical to the microsatellite marker "EPCO". EPDCO marker was clustered with the date palm isolates but not with those of other palm isolates from subfamily*Coryphoideae* as shown in (Fig.8).

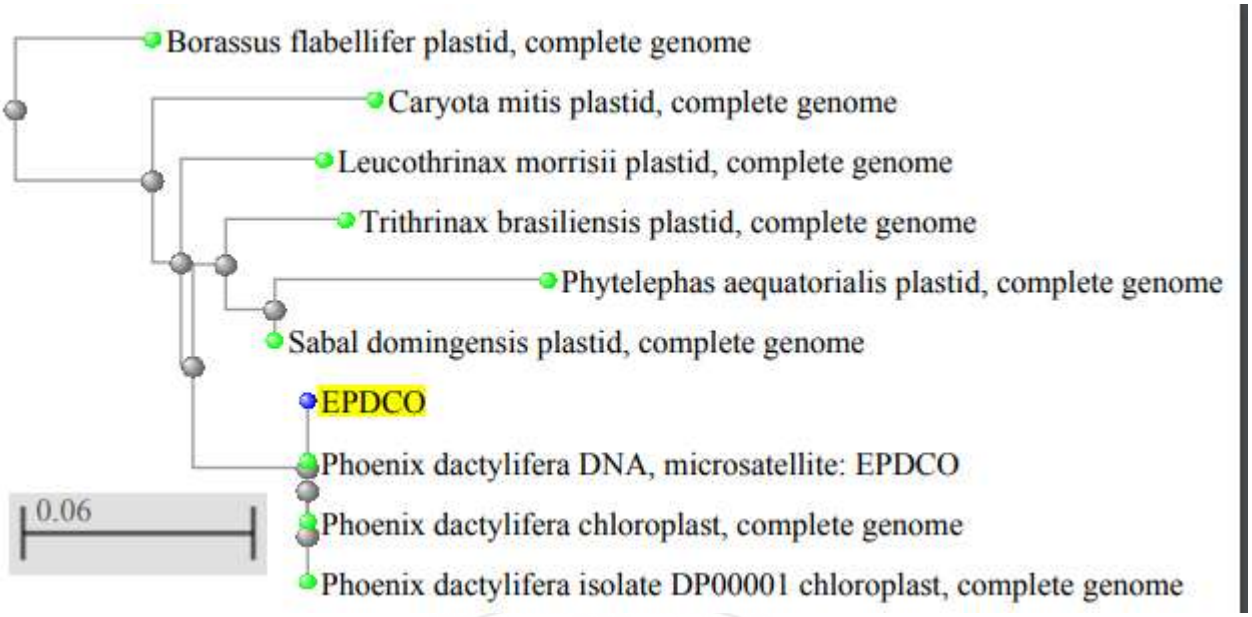

Figure 8: BLAST tree view of sequence repeats: The "EPDCO" sequence repeat, sequence repeats of date palm chloroplast isolates (*P. dactylifera*) and other palm isolates from subfamily *Coryphoideae*.

4. Conclusion

This study focused on the importance of characterized and identifying the types of tandem repeats in Phoenix dactylifera (cultivar, Khalas- female) chloroplast genome by using the power of regular expression language which was entered in the main form of RFRE Ver.1.0. The highest total length of repeats which were founded by the RFRE tool had the following pattern (AAAA){2} and the smallest length of repeats where were founded by the RFRE tool had the following patterns (ACCCG){2} & (AAAGG){2}. There were three different longest repeat in the chloroplast genome with 40 characters in length as shown previously in (Table 2.0). To validate the in silicon analysis a specific primer was designed to amplify the repeated region (8675..8715) which located in the accession number (GU811709.2) and compare it to the repeated sequence of chloroplast genome after it was sequenced where the similarity is 100%. Microsatellite EPDCO which detected and isolated from chloroplast genome of phoenix dactylifera was unique for phoenix dactylifera (date palm). The program RFRE tool had a lot of features, The size of the programmed tool had a small size and had an [.exe] extension. Using the controller of VB and the visual form object did visualization for the statistical output after the targeted patterns were retrieved. The user could write many different probabilities of combined operator to retrieve more than one patterns and very specific patterns of repeats. User can write a different flexible Regexp operator to retrieve more than one patterns not only searching automatically for fixed Mono, Di repeat or Hexa repeats (perfect repeat and imperfect repeat). Data could be exported to excel sheet file and word file. The programmer could embed inMs-Access (2003) by using VB editors.

References

[1] Weber, J. L. 1990. In formativeness of human poly(GT) polymorphisms. Genomics 7:524–530.
[2] Jurka, J., and C. Pethiyagoda. 1995. Simple repetitive DNA sequences from primates: compilation and analysis. J. Mol. Evol. 40:120–126.FANG
[3] Hancock, J. M. 1995. The contribution of slippage-like processes to genome evolution. J. Mol. Evol. 41:1038–1047.
[4] Borstnik, B., and D. Pumpernik. 2002. Tandem repeats in protein coding regions of primate genes. Genome Res. 12:909–915.
[5] Tautz, D., and C. Schlo¨ tterer. 1994. Simple sequences. Curr. Opin. Genet Dev. 4:832–837.
[6] Tautz, D., and M. Renz. 1984. Simple sequences are ubiquitous repetitive components of eukaryotic genomes. Nucleic Acids Res. 12:4127–4138.
[7] Have, C.T. and Christiansen, H. 2011. Modeling repeats in DNA using extended probabilistic regular expressions. Proc. 1st International Work-Conference on Linguistics, Biology and Computer Science: Interplays. IOS Press (to appear 2011). Tarragona, Spain, March 14-18, 2011.
[8] Elmeer, Khaled, and ImeneMattat. "Marker-assisted sex differentiation in date palm using simple sequence repeats." 3 Biotech 2.3 (2012): 241-247.
[9] Tóth, Gábor, ZoltánGáspári, and Jerzy Jurka. "Microsatellites in different eukaryotic genomes: survey and analysis." Genome research 10.7 (2000): 967-981.
[10] Palliyarakkal, ManjuKalathil, ManimekalaiRamaswamy, and ArunachalamVadivel. "Microsatellites in palm (Arecaceae) sequences." Bioinformation 7.7 (2011): 347.
[11] Fang, Yongjun, *et al*. "A complete sequence and transcriptomic analyses of date palm (Phoenix dactylifera L.) mitochondrial genome." *PloS one* 7.5 (2012): e37164.