Sigma J Eng & Nat Sci 35 (4), 2017, 763-770

Sigma Journal of Engineering and Natural Sciences
Sigma Mühendislik ve Fen Bilimleri Dergisi

Research Article

THE DETECTION OF SSRs FROM OLIVE (Olea europaea L.) EST COLLECTION BY COMPARING 4 DIFFERENT BIOINFORMATICS SOFTWARES

Nehir ÖZDEMİR ÖZGENTÜRK[1]*, Gamze TURAN[2], Nur Hatice AYDIN[3], Salih ULU[4], Zehra ÖMEROĞLU ULU[5]

[1] *Department of Molecular Biology and Genetics, Faculty of Science and Art, Yıldız Technical University, Esenler-ISTANBUL;* ORCID:0000-0003-3809-6303
[2] *Department of Molecular Biology-Genetics and Biotechnology, Faculty of Science and Art, Istanbul Technical University, Sarıyer-ISTANBUL;* ORCID:0000-0003-4563-2792
[3] *Department of Molecular Biology and Genetics, Faculty of Science and Art, Yıldız Technical University, Esenler-ISTANBUL;* ORCID:0000-0001-9213-285X
[4] *Department of Molecular Biology and Genetics, Faculty of Science and Art, Yıldız Technical University, Esenler-ISTANBUL;* ORCID:0000-0002-4505-0197
[5] *Department of Molecular Biology and Genetics, Faculty of Science and Art, Yıldız Technical University, Esenler-ISTANBUL;* ORCID:0000-0002-8884-4683

Received: 28.11.2016 Revised: 18.04.2017 Accepted: 23.10.2017

ABSTRACT

Bioinformatics is an interdisciplinary science that is formed by the combination of computer science and biology. Over the years, accumulating data were stored carefully, organized, incorporated, provided classification and accessed easily by bioinformatics. Experimental studies take a long time and are expensive but by means of bioinformatics these studies take a short time and are cheaper.

There are two approaches to determine SSRs (Simple Sequence Repeat/Microsatellite) in plants which are analysis of DNA libraries and analysis of EST (Expressed Sequence Tag) collections. Like many studies these two approaches also benefit from bioinformatics. Thus SSRs, which were determined, can be used as markers in genetic diversity studies.

In this study, to determine SSRs in olive EST collection that has 3734 EST; SSRIT, SSR Finder, WebSat and IMEx Softwares, which have web-based versions and are easily accessible and utilized, were used. SSR motifs and positions of the EST collection have been determined by means of these four softwares. Each of the four softwares found 2, 3, 4 and 6-nucleotide SSRs. "GA" pattern repeated 6 times, is the most abundant SSR with 324 occurrences. The advantages and disadvantages of the softwares used were determined by comparing the results.

Keywords: Bioinformatics, EST, SSR , Olive (*Olea europaea* L.).

* Corresponding Author: e-mail: nozdemir@yildiz.edu.tr, tel: (212) 383 44 77

1. INTRODUCTION

Bioinformatics is a new and interdisciplinary science that is born by the combination of biology and computer science, as well as biochemistry, chemistry and medicine and mathematics and statistics. It is based on the usage of computer technology in biological problem solving [1]. The publication of the first article in Scientific American Magazine in 1966, about drawing of molecular graphs by computers, is considered as the start of bioinformatics [2]. The term bioinformatics had begun to be used in the late 1980s and the Human Genome Project (HGP) studies launched in 1987 has been a crucial driving force in the development of bioinformatics [3]. Techniques developed in the last 20-25 years have led to many developments in medicine, genetics, biology and molecular biology. Bioinformatics creates databases and ensure that the data obtained as a result of these developments are carefully stored, organized, merged, classified and easily accessed in databases [4]. Through bioinformatics analysis, researchers can analyze large-scale information in their hands. In this study, identification of SSRs from olive (*Olea europaea* L.) EST collection was performed by bioinformatics analysis.

Olive (*Olea europaea* L.) belongs to the genus Olea involved in the Oleaceae family. It is a plant whose homeland is upper Mesopotamia that started to be cultivated 3000 BC. Olive is a temperate climate plant, whose production shows prevalence in the countries of the Mediterranean climate zone [5]. The major olive producing countries in the Mediterranean are Spain, Italy, Greece, Turkey, Tunisia, Syria and Morocco [6]. While meeting 10% of the world's olive production Turkey is the first in black olive production, the second in table olive production and the fourth in olive oil production. In Turkey, as the most produced variety of black table olive, Gemlik variety ranks first (Olive Culture Research Institute) [7]. For the Gemlik variety, 3734 EST collection was prepared from two cDNA libraries prepared for leaf and olive fruit [8]. ESTs are short, unedited, and single-pass sequence readings from randomly selected cDNA libraries. They are segments of functional genes but are not functional in protein coding. In 1991, ESTs were as the primary source for the discovery of human genes. Later, for numerous organisms, EST production in databases and data accumulation has increased exponentially. Currently ESTs are used to facilitate gene discovery, help identification of genomes and define gene structure, identify alternative transcripts, guide single-nucleotide polymorphism (SNP) characterization, and facilitate proteomic analysis. In these studies, ESTs offer a fast and low cost way [9]. Since ESTs are gene discovery tools, the EST database (dbEST) is easily accessible from NCBI [10].

SSRs are short repeat sequences (GA, GCT, AGAT, TATACA, etc.) 2-6 bases in length and show genome specificity in higher organisms [11]. They are found in high amounts in the eukaryotic genomes; whereas in prokaryotic genomes they have low frequency and are distributed throughout the entire genome. Although the functions of SSRs are not fully known, they are thought to have coding and regulatory functions within the genome [12].

SSRs are classified according to the sequence of the motif sequence in the genome. There are 4 types of SSRs (GAGAGAGAGAGA - Perfect Repeats, GAGAGA**T**GAGAGA - Imperfect Repeats, TCTCTCGAGAGA - Compound Repeats TGTGTCTCTGTA - Region of Cryptic Simplicity) [13].

Because of their ease in identifying the differences between individuals, SSRs have become one of the most widely used genetic methods, in recent years. The difference in the number of repeats between the two alleles gives the difference between individuals. Generally, SSRs are used as molecular markers, since they vary widely in the number of repeat regions in most of the studied loci [14]. SSRs have been widely used as molecular markers, which are widely and abundantly dispersed in genomes. Compared with other molecular markers, expressed sequence tag-based SSR (EST-SSR) markers have the advantages of co-dominant inheritance, highly polymorphic loci, rich information content, good transferability between species, and they are easily visualized and stable [15]. EST-SSR molecular markers can be used to assess genetic diversity, the evolution of specie s, and in comparative genomics research. Because EST-SSRs are derived directly from expressed gene coding sequence, they can be used to screen for differences in the EST-SSRs that are associated with different phenotypes from similar species [16].

SSRs are highly polymorphic DNA markers and have a wide range of applications in many species [17]. Being distributed almost evenly to the whole genome makes them practical for genomic mapping projects. The high diversity they have makes them a genetic marker for population genetics and paternity and kinship detection. Since SSRs are codominant and polymorphic, they are becoming increasingly important in the research of the structures of natural populations [18]. Because of their codominant inheritance, they are used in the detection of homozygous-heterozygote allelic variations, in gene duplication and deletion studies, in criminological studies, in the extraction of genomic maps, in the prediction of genetic parameters of populations, in determining population differences [19].

Many software have been developed to find SSRs in EST collections. SPUTNIK (1994) is the first SSR detection software [20]. Then many softwares such as RPT (Repeat Pattern Toolkit) [21], REPuter [22], TRF (Tandem Repeat Finder) [23], SSRIT (Simple Sequence Repeat Identification Tool) [24], TROLL (Tandem Repeat Occurrence Locator) [25], SSR Finder [26], WebSat (A Web Software for MicroSatellite Marker Development) [27], IMEx (Imperfect Microsatellite Extractor) [28] have been developed. It was aimed to determine the correct SSRs by comparing four different common used (SSR Finder, SSRIT, IMEx, WebSat) the software results in order to prevent mistakes caused by the weaknesses of the software.

2. MATERIALS AND METHODS

2.1. ESTs: 3734 EST, sequenced from fruit and leaf cDNA libraries prepared from *Olea europaea* L. by N. Ozdemir Ozgenturk et al. were used. Accession numbers found in GenBank (DBEST) of the ESTs used: GO242703-GO246436 [8].

2.2. Softwares of SSRs Detection:

Simple Sequence Repeat Identification Tool (SSRIT), was developed in 2001 by S. Temnykh, G. DeClerk, A. Lukashova and colleagues to detect microsatellites (SSRs) of rice (Oryza sativa L.) in a set of 57.8 Mb size. The SSRIT Software uses Perl script to detect regular SSRs in the sequence and accepts the data in FASTA format. By eliminating the single nucleotide motifs, the software finds motifs of 2 to 10 base lengths and the minimum number of repetitions can be adjusted. The SSR data can be obtained in table format as output [24].

SSR Finder; SSR Finder is a software developed by the California State University in 2009.This software, in which the motif length and the minimum repeat amount can be adjusted, accepts only FASTA format data and gives results in a very short time in tabular format[26].

WebSat (A Web Software for Microsatellite Marker Development); is a microsatellite finding software developed by Martins WS, Lucas DCS, Neves KFS and Bertioli DJ in 2009. The data to be analyzed must be in FASTA format and maximum 150.000 characters. The data can be uploaded as a file.Finds motifs from 1 to 6 bases in length and the minimum number of repetitions can be adjusted. The software also shows overlapping SSRs and is capable of primer design [27].

IMEx (Imperfect Microsatellite Extractor); was developed by Suresh B. Mudunuri and Hampapathalu. A. Nagarajaram in 2007. The software finds regular and irregular SSRs separately and isavailable in desktop and web-based versions. Finds motifs of different lengths from 1 to 6 and the minimum number of repetitions can be adjusted. The data to be uploaded must be in FASTA or Plain format. Gives results in tabular format and is capable of primer design. SSRs of bacterial and viral genomes found in the database of the software can be found without loading data by selecting the species [28].

2.3. Detection of SSRs: All the softwares used in this study are web-based versions of the softwares listed in 2.2.The softwares used accept the data in FASTA format and the EST collection including 3734 EST that we used in this study is available in FASTA format.

SSR's were found by using the SSRIT Software. Parameters from the main page of the software were set to maximum 6 nucleotide motifs and minimum 5 repetitions. As the ESTs collection to be searched for the including SSRs, was a large piece of data, it was divided into 10 parts. Each piece was pasted to the search window individually and output was generated by pressing the "FIND SSRs" button. This process has been repeated 10 times. The resulting output was combined and tabulated.

SSRs were found using SSR Finder Software. Parameters from the "Options" section of the main page of the software were set 2-6 nucleotides and a minimum of 5 repetitions. The EST collection is divided into 10 parts, as in the SSRIT Software, to obtain faster results. Each piece was pasted to the search window individually and output was generated by pressing the "FIND" button. This process has been repeated 10 times. The resulting output was combined and tabulated.

SSRs were found using WebSat Software. The main page of the software has been accessed. As motif length of SSRs, 2-6 options were selected and minimum repetition number was entered as 5. Since the software searches for a maximum of 150,000 characters, the data was divided into 30 parts. Each piece was pasted to the search window individually and output was generated by pressing the “Submit It!” button. This process has been repeated 30 times. The resulting output was combined and tabulated.

SSRs were found by using IMEx Software. Basic search mode has been selected from the main page of the software. The parameters were set to "perfect" as the repeat type.2-6 nucleotides as motif lengths and 5 as minimum repetition number were entered. SSRs were searched by pasting ESTs into the search window, and "Cut and Paste Your Sequence" was selected. EST collection was divided into 30 parts. Each piece was pasted to the search window individually and output was generated by pressing the “EXTRACT MICROSATELLITES” button. This process has been repeated 30 times. The resulting output was combined and tabulated.

3. RESULTS

In all four softwares used the 2, 3, 4 and 6 nucleotide motifs were found whereas 5 nucleotide motifs were not found. Table 3.1, Table 3.2, Table 3.3 and Table 3.4 show the SSR results of each sofware separately in tabular form. As the number of nucleotides in the motif increased, the repeat number of motifs found in the EST collection decreased and the total number of motifs decreased. There are some common results with the four softwares. The most common motif in 2 nucleotides motifs and with the longest repeat number is also the most common motif in whole EST collection and has the longest repetition motif. These; 324 of them are 6 repeats sequence of GA and 34 repeats of AT motif. The most common motif in 3 nucleotide motifs is the 5 repeats of TCT and the motifs with the longest repeat sequence are 9 repeats of CTT and TAT. The most abundant and the longest repeat motif in 4 nucleotides motif is same and that is 5 repetitions of ATTT. In motifs with 6 nucleotides the most abundant and the motifs with longest repeats are the same and are 5 repetitions of AGCACA and TATACA.

3.1. SSRIT Software

Table 3.1. Nucleotide motifs detected by SSRIT software

Motif Length	Amount of Motif Types	Total Amount of Motifs	The Most Common Motifs	Amount of The Most Common Motifs
2 nt.	40	539	$(GA)_6$	324
3 nt.	27	50	$(TCT)_5$	6
4 nt.	2	2	$(AGAT)_5$ $(ATTT)_5$	1
6 nt.	2	2	$(AGCACA)_5(TATACA)_5$	1

2, 3, 4 and 6 nucleotide (nt) long motifs were counted by SSRIT software and total number of motifs and most common motif were reported.

3.2. SSR Finder Software

Table 3.2. Nucleotide motifs detected by SSR Finder software

Motif Length	Amount of Motif Types	Total Amount of Motifs	The Most Common Motifs	Amount of The Most Common Motifs
2 nt.	42	536	$(GA)_6$	324
3 nt.	21	41	$(TCT)_5$ $(GCT)_5$	3
4 nt.	1	1	$(ATTT)_5$	1
6 nt.	2	2	$(AGCACA)_5(TATACA)_5$	1

2, 3, 4 and 6 nucleotide (nt) long motifs were counted by SSR Finder software and total number of motifs and most common motif were reported.

3.3. WebSat Software

Table 3.3. Nucleotide motifs detected by WebSat software

Motif Length	Amount of Motif Types	Total Amount of Motifs	The Most Common Motifs	Amount of The Most Common Motifs
2 nt.	40	538	$(GA)_6$	324
3 nt.	27	50	$(TCT)_5$	6
4 nt.	2	2	$(AGAT)_5$ $(ATTT)_5$	1
6 nt.	2	2	$(AGCACA)_5(TATACA)_5$	1

2, 3, 4 and 6 nucleotide (nt) long motifs were counted by WebSat software and total number of motifs and most common motif were reported.

3.4. IMEx Software

Table 3.4. Nucleotide motifs detected by IMEx software

Motif Length	Amount of Motif Types	Total Amount of Motifs	The Most Common Motifs	Amount of The Most Common Motifs
2 nt.	40	539	$(GA)_6$	324
3 nt.	26	49	$(TCT)_5$	6
4 nt.	2	2	$(AGAT)_5$ $(ATTT)_5$	1
6 nt.	2	2	$(AGCACA)_5(TATACA)_5$	1

*$(\text{Motif})_{\text{repeat number}}$ *nt: nucleotide

2, 3, 4 and 6 nucleotide (nt) long motifs were counted by IMEx software and total number of motifs and most common motif were reported.

4. DISCUSSION AND CONCLUSION

The amount of SSRs detected by all the softwares used was close to each other but different. In 3734 ESTs, with an average of 41 types of motifs and with a number of average 538, 2 nucleotide motifs were found the most. Then with 25 types of motif and with a number of 48, 3 nucleotide motifs were found and with average 2 types of motifs and with a number of average 2, 4 and 6 nucleotide motifs were found the least.

The softwares used for SSR detection have their own advantages and disadvantages:

SSRIT Software; receives large data and outputs it in tabular form. The software informs about the length of the uploaded EST and the SSR's motif, the number of motifs, and the location at the EST in the output, but it is observed that when manually checked the location of the detected SSRs is incorrect.

SSR Finder Software; Although the software analyzes the uploaded data quickly and unlike other softwares it finds multiple SSRs, but has detected fewer SSRs in the ESTs than other softwares and was unable to find some SSR motifs

WebSat Software; Unlike other softwares, the software output is not in tabular format, but it is observed that the output is clear and useful. This software, which detects overlapping SSRs, can also be used for primer design. The biggest disadvantage of the WebSat software is that it can analyze data smaller than 150,000 characters at a time.

IMEx Software; Unlike other softwares, it accepts plain format data. If parameters are set, it finds irregular SSRs and performs primer design, but takes the uploaded data as a whole and does not output data according to the directory information. Therefore, the location of the SSR in a single loaded EST sequence is readily available, and simple mathematical calculations are needed to find the exact location of SSRs in multiple loaded ESTs.

Even though differences arise from the functioning of the softwares, it was easy to identify SSRs from the ESTs by all the softwares used. Despite the fact that the programs did not see some SSRs, the fact that there were no major differences between the results of these four programs, proving the correctness of the SSRs found, supporting each other's results.

REFERENCES

[1] Luscombe NM, Greenbaum D, Gerstein M. What is bioinformatics? An introduction and overview. Yearbook of Medical Informatics 2001; s:83-85.

[2] Polat M., Karahan A.G. (2009) "Multidisipliner Yeni Bir Bilim Dalı: Biyoinformatik ve Tıpta Uygulamaları" S.D.Ü. TIp Fak. Derg. 16(3)/ 41-50

[3] Collins F.S., Morgan M. and Patrinos A. The Human Genome Project: Lessons from Large- Scale Biology Science 2003; 300, 286-290.

[4] Altschul, S.F., Boguski, M.S., Gish, W. & Wootton, J.C. (1994) "Issues in Searching Molecular Sequence Databases" Nature Genet. 6:119-129

[5] Yurtsever M. (2011) "Zeytin (*Olea Europaea* L.) Bitkisinden Karbonik Anhidraz Enzimini Kodlayan Genin Klonlanması" s. 20.

[6] Rugini E., Biasi R. ve Muleo R., (2000). "Olive (*Olea europaea* var. *sativa*) Transformation; Mohan J.S., Minocha S.C., Molecular Biology of Woody Plants, 2, Kluwer Academic Publishers, Dordrecht.

[7] http://arastirma.tarim.gov.tr/izmirzae.

[8] Ozdemir Ozgentürk N., Oruc F., Sezerman U., Kuçukkural A., Vural Korkut Ş., Toksoz F., Un C. Generation and Analysis of Expressed Sequence Tags from Olea europaea L.. Hindawi Publishing Corporation Comparative and Functional Genomics Volume 2010, Article ID 757512.

[9] Nagaraj, S.H., Gasser, R.B., Ranganathan S. 2006 A hitchhiker's guide to expressed sequence tag (EST) analysis. Brıefıngs In Bıoınformatıcs. Vol 8. No 1. 6-21.

[10] Matukumalli L.K., Grefenstette J.J. , Sonstegard T.S. , Van Tassell C.P. EST-PAGE—managing and analyzing EST data, Bioinformatic Applications Note Vol. 20 no. 2 2004, pages 286–288 DOI: 10.1093/bioinformatics/btg411.

[11] Litt, M. and Luty, J.A. 1989. A hypervariable microsatelllite revealed by in vitro amplification of a dinucleotide repeat within the cardiac muscle actin gene. Am J Hum Genet. 44: 397-401.

[12] Devrim K.A ve Kaya N. 2004. Genetik Polimorfizm ve Mikrosatelitler.Kafkas Üniversitesi Veterinerlik Fakültesi Dergisi. 10(2):215-220.

[13] Kibar U. 2012. "Ankara Üniversitesi Biyoteknoloji Enstitüsü Est (Expressed Sequence Tag) Koleksiyonlarından Üzüm Mikrosatellit Lokuslarının Tanımlanması" Ankara. s. 3-4

[14] Aktaş P. ve Sönmez Z. "Çiftlik Hayvanlarının Islahında İmleç Yardımlı Seleksiyon (MAS)" 7. Ulusal Zooteknoloji Öğrenci Kongresi 20-22 Mayıs 2011. Aydın.

[15] Castillo, A., Budak,H., Varshney, R.K., Dorado, G., Graner, A., Hernandez, P. Transferability and polymorphism of barley EST-SSR markers used for phylogenetic analysis in *Hordeum chilense,* BMC Plant Biol., 8 (1) (2008), p. 97.

[16] Yuan, Y. Long, P., Jiang, C., Li, M., Huang, L., Development and characterization of simple sequence repeat (SSR) markers based on a full-length cDNA library of *Scutellaria baicalensis* Genomics, 105 (1) (2015), pp. 61–67.

[17] Karsi, A., Patterson, A., Feng, J., Liu, Z.J. 2002. Translational machinery of channel catfish: I. A transcriptomic approach to the analysis of 32 40S ribosomal protein genes and their expression, Gene, 291: 177–186.

[18] Çiftçi Y.2004. YUNUS Araştırma Bülteni-yıl :4, sayı:2, Haziran 2004.

[19] Wenz HM, Robertson JM, Menchen S, Oaks F, Demorest DM, Scheibler D, Rosenblum BB, Wike C, Gilbert DA, Efcavitch JW (1998) High-precision genotyping by denaturing capillary electrophoresis. Genome Res s:69–80.

[20] Abajian C., SPUTNIK, 1994, Espresso Software Development http://espressosoftware.com/pages/ sputnik.jsp.

[21] Agarwal, P. and States, D.J. 1994. The repeat pattern toolkit (RPT): analyzing the structure and evolution of C. elegans. Proc. Int. Conf. Intell. Sys. Mol. Biol. 2: 1–9.

[22] Kurtz, S., Choudhuri, J.V., Ohlebusch E., Schleiermacher, C., Stoye, J., Giegerich R. 2001. REPuter: the manifold application of repeats analysis on a genomic scale. Nucleic Acids Res. 29: 4633–4642.

[23] Benson B., “Tandem repeats finder: a software to analyze DNA sequences” Nucleic Acids Research (1999) Vol. 27, No. 2, pp. 573-580.

[24] Temnykh, S., DeClerk, G., Lukashova, A. *et al.* 2001. Computational and experimental analysis of microsatellites in rice (*Oryza sativa* L.): Frequency, length variation, transposon associations, and genetic marker potential. Genome Res. 11: 1441–1452.

[25] Castelo, A.T., Martins W., Gao, G.R. 2002. TROLL – tandem repeat occurrence locator. Bioinformatics 18: 634–636.

[26] California State University. 2009. SSR Finder http://fresnostate.edu/csm/faculty-research/ssrfinder/).

[27] Martins WS, Lucas DCS, Neves KFS, Bertioli DJ, WebSat - A Web Software for MicroSatellite Marker Development, Bioinformation 2009, 3(6):282-283.

[28] Suresh B. Mudunuri and Hampapathalu. A. Nagarajaram (2007) IMEx: Imperfect Microsatellite Extractor. *Bioinformatics* 23(10):1181-1187.