Journal of
Data Mining in Genomics & Proteomics

Ophir, J Data Mining Genomics Proteomics 2013, 4:4
DOI: 10.4172/2153-0602.1000140

Research Article Open Access

An Analysis of Palindromes and n-nary Tract Frequencies found in a Genomic Sequence

Dan Ophir*

Department of Computer Science and Mathematics, Ariel University, Israel

Abstract

The motivation to investigate n-nary and palindrome tracts arose following the discovery by Chargaff and coworkers of over-representation of certain DNA binary tracts in genomes. They investigated the frequencies of various ternary tracts in diverse locations in genes of various species.

The current research further examines ternary tracts and the palindromes will hereafter be called designated tracts. Does a designated tract have any extraordinary frequencies of length and location? A theoretical mathematical analysis has been performed to analyze the amount of designated tracts according to the frequencies of its single elements. The designated tracts are categorized according to those that have mixed elements from a subset of a set composing the sequence, and according to which consist of a long tract of lower n-nary order.

For example, tract analysis investigates whether the special phenomena are due to the ternary tract or due to a long binary tract that is included in it. The maximal n-nary tract order of interest in the genome is of three (ternary); four is the whole gene itself. However, the higher order of n-nary tracts is of interest in other areas like “Reliability theory”.

Therefore, the general formulation and treatment of designated tracts is presented here and is demonstrated for the genomic aspects, which were thoroughly investigated in the two past decades.

Keywords: Palindrome; Unary tract; Binary tract; Ternary tract; n-nary tract

Introduction

The intriguing question of whether ternary over-representation is due to included binary tracts is answered in the present work. As will be shown, the binary tracts are over represented to the same extent as the ternary tracts; therefore, it can be concluded that the binary tracts dominate because they have biological impacts, but the ternary tracts do not contribute to biological impacts.

The genomic sequence may be compared to the reliability sequence in the mass production of S (successes) and F (failures), which may be designated according to their severity. Thus, it is possible to receive a sequence such as $SF_1F_2F_1F_2SSF_1$. The searching of some patterns might result in interesting answers regarding the line production behavior.

Palindromes are another investigated pattern. There were no notable special extraordinary frequencies associated with their appearances.

Materials

As previously mentioned, the current work is based on previous research [1-6]. The data used in the investigation concerned the human gene p53, as given in genebank.

Methods

A special technique has been derived that is supported by further formalism and was validated by a program [7].

Definitions

Tract: A tract is a sub-sequence within a sequence of events having a mutual property.

Examples:

a. A sequence of events, for example, tossing a coin. Each event can have one of two values S (the coin falling on heads) or F (the coin falling on tails) [8].

b. A genomic sub-sequence consisting of nucleotides having special properties. An event is associated with one of the nucleotides from a set of nucleotides.

c. sub-sequence of a wide spectrum of qualities of a product in a serial mass production.

Source-set: Source-Set is a set of elements that builds the sequence.

Kernel set: Kernel-Set is a set of elements. The tract is built from elements taken from the Kernel-Set. The Kernel-Set is a subset of a Source-Set.

Unary tract: A tract that is composed of one type of elements, i.e., a set of elements from which the tract elements are composed, and which contain one element only. The cardinality of the Kernel-Set of the unary tract is one.

Exact unary tract: Exact Unary Tract is a unary tract that cannot be enlarged by exactly one element (one of its neighbors’ elements) and remain a unary-tract. Namely, the exact-unary-tract cannot be contained in another unary-tract.

Binary tract: Binary Tract is a tract whose Kernel-Set has a cardinality of two.

***Corresponding author:** Dan Ophir Department of Computer Science and Mathematics, Ariel University, Israel, E-mail: dano@ariel.ac.il

Received July 24, 2013; **Accepted** August 22, 2013; **Published** August 29, 2013

Citation: Ophir D (2013) An Analysis of Palindromes and n-nary Tract Frequencies found in a Genomic Sequence. J Data Mining Genomics Proteomics 4: 140. doi:10.4172/2153-0602.1000140

Copyright: © 2013 Ophir D, et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Exact-binary-tract: Exact-Binary-Tract is a binary tract that cannot be enlarged by one neighboring element and it remains a binary tract.

Pure-binary-tract: Pure-binary-tract is a tract whose elements are mixed enough, i.e. it does not contain a dominant - long enough (longer than some critical length) unary-tract.

This restriction is required in order to exclude the possibility that the properties that are assumed to belong to a binary tract are not due to some sub-unary tract that is included in the binary tract.

Exact-pure-binary-tract: Exact-Pure-Binary-Tract is a binary tract that is both: a pure binary tract and an exact binary tract.

Ternary and n-nary tracts: Similarly to the unary tract of all types, there are ternary tracts and n-nary tracts of the same types. Ternary's kernel set possesses three elements; an n-nary tract kernel-set has n elements, i.e. n-nary tract is of order n.

According to these definitions, n-nary tract of order n-1 is also an n-nary tract. Namely, each binary tract is also a ternary tract. In the context of genomics, the highest order of n-nary tracts in the area of interest is three because the source set contains four elements and therefore there is justification for treating the quarto nary tracts in which the kernel set and the source set are equal.

Probability computations

Let p be the probability of an element to be from the kernel-set and q=1-p the probability of not being from the kernel set, i.e. the probability of belonging to the complementary set, namely to the

complementary-set=source-set–kernel-set.

Random variable $X_i^{(1)}$

$X_i^{(1)}$ is a random variable representing the length of an Exact Unary Tract starting at the i'th position of the sequence; the superscript (1) denotes the "unarity" of the tract. The general case is Xi(n) for any named "exact n-narity tract".

Pr{$X_i^{(1)}$ =k (1)

the exact unary tract starting at the i' th position (i=0,…,L) is of the length k}

The random variable $X_i^{(1)}$ covers the whole domain as k→ ∞:

$$\sum_{k=0}^{\infty}\Pr\{X_i^{(1)}=k)=\sum_{k=0}^{\infty}q^2p^k=q^2\frac{1}{1-p}=q \quad < \quad 1 \tag{2}$$

The sum (equation 2) does not equal 1. Otherwise, each sequence should have at least one exact unary tract. This is not true; as a counterexample, see the sequence 'FS' (failure followed by a success), which has no exact tracts, versus the sequence 'FF', which has a tract (however of length 0).

Namely, the probability of the exact-unary-tract (even of length 0) converges to q. The probability of a tract not being an exact-unary-tract is 1-q=p. Such a tract should start with an element not belonging to the kernel-set and therefore its probability is p.

Exact unary tracts

The expected number of exact unary tracts of length k in a sequence of length L (assumed k<<L) can be found in [1]. Owing to the assumption about the long L compared to k, the extreme cases (i=0 and i=L – the first position is designated by 0) may be neglected and treated as the regular one in the additional formulas in this article.

The random variable Yi ={0, 1} is introduced:

$$Y_i=\begin{cases}1 & X_i^{(1)}=k\\ 0 & X_i^{(1)}\neq k\end{cases} \tag{3}$$

Using the formula for expectation, the following relations are derived (equation 5- equation 6):

$$E(Y_i)=\sum_{j=0}^{1}j\Pr\{Y_i=j\}=\Pr\{X_i^{(1)}=k\} \tag{4}$$

$$Exact(1,k,L,p)=E\left(\sum_{i=0}^{i=L}Y_i\right)=\sum_{i=0}^{L}E(Y_i)=\sum_{i=0}^{L}\Pr\{X_i^{(1)}=k\} \tag{5}$$

$$Exact(1,k,L,p)=\sum_{i=0}^{L}\Pr\{X_i^{(1)}=k)\approx\sum_{i=0}^{L}q^2p^k=Lq^2p^k=L(1-p)^2p^k \tag{6}$$

The term Exact (n,k,L,p) is a function whose name and parameters can be explained as follows: Exact – Exact tract, n- nary order: (1 stands for Unary, 2 stands for Binary, 3 stands for ternary), k-tract length, L-Sequence length, and

p is the probability of being an element from the kernel set of the unary tract (Figure 4).

p_n–The probability of an element taken from the kernel-set of the nary tract of the order n.

$$p_1=p \tag{7}$$

$$p_2=2p_1=2p \tag{8}$$

$$p_n=np_1 \tag{9}$$

L designates the length of a sequence, i.e., the number of elements in a sequence.

It is assumed that k << L (k of an order of ten and L at least of an order of ten thousand) and therefore the events Xi(1)=k and Xj(1)=k with i+k < j are disjoint.

In (equation 5), if Pr{$X_i^{(1)}$=k} is given at each i, it means that the expected number of exact unary tracts of length k at i is its probability. It follows that the sum over the all is gives the total number of expected exact unary tracts of length k in the sequence, as is supported by a MATLAB program [7].

The following equality takes place:

Exact (2,k,L,p)= Exact (1,k,L,2p) (10)

Exact binary tracts

$X_i^{(2)}$ is a random variable representing the length of the Exact Binary Tract starting at the i'th position of the sequence; the superscript (2) designates the "binarity" of the tract.

Pr{Xi(2)=k (11)

which represents the exact binary tract, starting at the i th position, is of the length k}.

As in the unary tracts, also in the binary tracts, k<<L is assumed. This assumption enables treating the extreme cases in which i=0 or i=L as regular ones without loss of much accuracy.

$$\begin{aligned}Exact(2,k,L,p)&=\sum_{i=0}^{L}\Pr\{X_i^{(2)}=k)\approx\sum_{i=0}^{L}(1-p_2)^2p_2{}^k=\sum_{i=0}^{L}(1-2p)^2(2p)^k=\\ &L(1-2p)^2(2p)^k=(1-2p)^2 2^k Lp^k=2^k\left(\frac{1-2p}{1-p}\right)^2 Exact(1,k,L,p)\end{aligned} \tag{12}$$

Citation: Ophir D (2013) An Analysis of Palindromes and n-nary Tract Frequencies found in a Genomic Sequence. J Data Mining Genomics Proteomics 4: 140. doi:10.4172/2153-0602.1000140

Exact n-nary tracts

$X_i^{(n)}$ is a random variable representing the length of the Exact Binary Tract starting at the i'th position of the sequence.

Equation (12) Pr{Xi(n)=k | , representing the exact n-nary tract, starting at the i'th position, is of length k}.

$$Exact(n,k,L,p)=\sum_{i=0}^{L}\Pr\{X_i^{(n)}=k)\approx\sum_{i=0}^{L}(1-p_n)^2 p_n{}^k=\sum_{i=0}^{L}(1-np)^2(np)^k=$$
$$L(1-np)^2(np)^k=(1-np)^2n^kLp^k=n^k\left(\frac{1-np}{1-p}\right)^2 Exact(1,k,L,p) \qquad (13)$$

Pure -exact n-nary tracts

$Z_i^{(n)}$ is a random variable representing the length of the Pure Exact n-nary Tract at the i'th position of the sequence.

Pr{Zi(n)=k (14)

representing the pure exact n-nary tract starting, at the i'th (i=0,…,L) position, is of the length k}.

$$P_E(n,k,L,p)=\sum_{i=0}^{L}\Pr\{Z_i^{(n)}=k)=Exact(n,k,L,p)\left(1-n\sum_{j=l\min}^{k}\sum_{\mathrm{i=1}}^{\mathrm{k\text{-}lmin}}\Pr\{\mathrm{X_i^{(n\text{-}1)}=j}\ \}\right) \qquad (15)$$

The developed formula (equation 15) is an approximated formula. The closed form of the definition 6.1 for small k values is under development [9]. The idea is to eliminate from the 'average' pure–exact tracts those exact n-nary tracts that are inside the original tract, but are of lower order and above some threshold length (l min). The expression is multiplied by n, because there are n possible combinations of sub-tracts.

The general form $\varphi(\alpha,\beta,\gamma,\delta)$ represents the following terms:

Φ-the expected number of tract functions of PE or E

α-the tract order

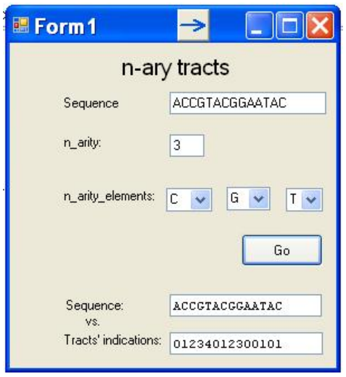
β-tract length

γ-sequence length

δ-probability of one element in the tract.

n-ary algorithm

The GUI (Figure 1) of the n-ary algorithm (Figure 2) is demonstrated. The GUI requires the nucleotides to be found in an n-ary tract. Its output is displayed in the GUI showing the found tracts, indicating the position of the nucleotides in their tract.



Sequence ACCGTACGGAATAC

n_arity: 3

n_arity_elements: C G T

Sequence: ACCGTACGGAATAC
vs.
Tracts' indications: 01234012300101

Figure 1: GUI of the n-ary algorithm (its Pseudo code, see Figure 2).

```
For i = 1 To sequenceLength
        myChar=sequenceString (i)
        myInt=Integer(myChar)
        mapString_nary(i)=codes(myInt)
        num 'codes(j)is 1 if j is the ASCII of one of  the tract characters
        If mapString_nary(i-1) = 0 Then
           'mapString_nary(i) contains the number of the position in
           'the tract of the i'th character of the sequence
           If codes(myInt)> 0 Then
            'If the i'th character of the sequence is a character belonging to
            'the 'tracts kernel (set of legal characters of the tract)
              mapString_nary(i)=1
                  Numerate the position of the character in the tract is 1
           Else
              mapString_nary(i)=0
                  Numerate the position of the character in the tract is 0,
                  'i.e. The character is not a tract character
           End If
        Else
           If codes(myInt)> 0 Then
              mapString_nary (i)=mapString_nary(i-1) + 1
                        Number the position of the character in the tract
                        'greater by 1 than the previous character
           Else
              mapString_nary(i)=0
                     'The Number the position of the character in the tract
is 0,
                     'i.e. The character is not a tract character
           End If
        End If
     Next          i -End of the loop with the index i.
           Display the array: mapString_nary(i)(i=1,...,sequenceLength)
```

Figure 2: Pseudo-code (for its GUI, see Figure 1) of an algorithm for finding tracts and numeration of the corresponding nucleotides in the tracts.

Figure 2 shows the pseudo-code of an algorithm for finding tracts and numerating their corresponding nucleotides.

The complexity of an algorithm is O(n), where n is the length of the sequence in which the tracts are searched. The specific nucleotides to be checked in the n-ary tracts are treated as ASCII characters. The values of the elements of an array named codes (i) are 0, 1, corresponding to the element number; however, if its position in the array equals the value of one of the specific nucleotides, then the value of array element is 1; otherwise it is 0. Using the codes (i) array significantly simplifies the algorithm.

Palindrome

A palindrome is a symmetric tract. In order to find the distribution of palindromes in the p53 sequence, a more rigorous palindrome identification approach is required.

Definitions: There are two distinguishable types of palindromes, depending on the parity of the number of their elements: even or odd.

$$Even: a_{-n-1}a_{-n}a_{-n+1}a_{-n+2}....a_{-3}a_{-2}a_{-1}a_1a_2a_3a_4....a_{n-2}a_{n-1}a_na_{n+1}$$

$$Odd: a_{-n-1}a_{-n}a_{-n+1}a_{-n+2}....a_{-3}a_{-2}a_{-1}a_0a_1a_2a_3a_4....a_{n-2}a_{n-1}a_na_{n+1}$$

An Exact-palindrome is a symmetric tract that cannot be extended any more. The following tract is an Exact-even-palindrome:

$$a_{-n-1}a_{-n}a_{-n+1}a_{-n+2}....a_{-3}a_{-2}a_{-1}a_1a_2a_3a_4....a_{n-2}a_{n-1}a_na_{n+1}$$

And it has a length of 2n if the following condition holds:

$$a_{-i}=a_i, i=1,...n \text{ and } a_{-n-1}\neq a_{n+1}$$

The following tract is an Exact-odd-palindrome:

$$a_{-n-1}a_{-n}a_{-n+1}a_{-n+2}....a_{-3}a_{-2}a_{-1}a_0a_1a_2a_3a_4....a_{n-2}a_{n-1}a_na_{n+1}$$

And has a length of 2n+1 if the following condition holds:

$a_{-i} = a_i, i = 1,...n$ and $a_{-n-1} \neq a_{n+1}$

Exact-pure-palindrome is an exact-palindrome that does not include any sub-palindrome.

For example, ACGCA is an exact-pure-palindrome,

whereas AGCGTGCGA is not an Exact-pure-palindrome, because it contains the GCG palindrome.

An algorithm for finding the palindromes

The palindrome algorithm given in Figure 3 above is of O(n) complexity.

Random sequence: A random sequence is generated for comparison with the p53 sequence to search for extraordinary properties of the gene sequence. For a comparison, the term uniformity is defined and used.

Uniformity is an index to measure the randomness of the distribution of the nucleotides in the sequence. Nucleotide uniformity of the nucleotide at the i'th position in the tract receives the following values: 1 if two of its neighbors are different and differ from the nucleotide at the i'th be consistent with I'th position; 0.5 if two of its neighbors of the i'th nucleotide differ from the i'th nucleotide but they are equal themselves; 0.25 if one of the neighbors of the i'th nucleotide and only one is the same as the i'th nucleotide; 0 for the other cases.

Tract-uniformity is the average uniformity of all its internal nucleotides.

Results

Analysis of the frequencies of the exact tracts

The question is as follows: do the n-nary exact tracts in the genomes have exceptional frequency? For the binary (n=2) tracts the answer [10,11] is positive, namely, the frequencies of binary tracts are over-represented in the genomes.

In the following section the analysis of exact tracts will be described, performed on the human oncostatic gene, p53 (entry HSP53G, accession number X54156, 20303 bases) and on a randomly generated sequence of bases.

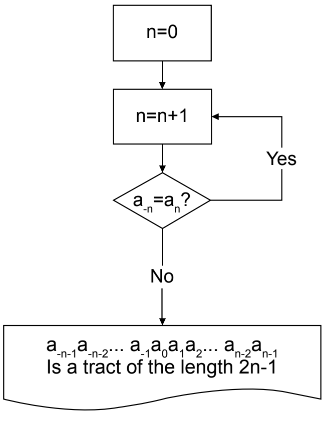

n=0
n=n+1
Yes
$a_{-n}=a_n$?
No
$a_{-n-1}a_{-n-2}... a_{-1}a_0a_1a_2... a_{n-2}a_{n-1}$
Is a tract of the length 2n-1

Figure 3: An algorithm for constructing an exact-palindrome.

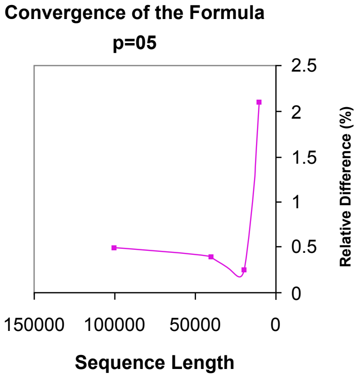


Figure 4: The graph that describes the convergence of the formula 5.4.4 (for p=0.5) for the values received numerically from pseudo-random sequences.

Unary exact tracts: As mentioned, relation 5.4.4 is an approximate one. Since that formula is the basis of deriving the other formulas, verification was made in order to check how close this formula describes the actual situation.

This verification is shown (Table 1 with specific columns: A, B, C, ...) by checking the E(1,k,L,p) (p was taken 0.25, 0.5, 0.75– corresponding to the probability of the nucleotide appearing in the unary, binary, and the ternary tracts) convergence for growing L (the length of the sequence); (Column B) min_tract size is the limit for minimal tracts allowed.

(Columns C, F, I) values received their sequence generated randomly; (Columns D, G, J) values of the expected number of exact unary tracts were obtained by formula 5.4.4.

It was deduced from Table 1 that the error received using the formula compared to the values received using a random sequence was for greater L (20, 000 <L values, i.e. the sequence-length) of about 0.03%.

Another approach for treating this issue (statistics of unary exact tracts) is demonstrated in Gera and Ophir [9].

Binary Exact tracts: The exceptions for the frequencies of binary exact tracts were introduced before [10,11]. In the current paper the results of over representing the binary exact tracts in the p53 gene are validated and compared to the derived general formula dealing with n-nary tracts of any order.

Comparing the two tables, Table 2 and 3, we can see that the binary exact tracts found in the p53 gene do not always decrease as do their counterparts (the distribution of similar tracts in the randomized sequence). The results received from the randomized sequence show that the exact tracts are shorter in general than in the p53 gene sequence; however, the frequencies of the nucleotides in the gene are almost the same as the theoretical one (Table 2).

Ternary exact tracts: Figures 5 and 6 describe graphs of the number of tracts for each length. They compare the difference between the exact ternary tracts' distributions of a randomly generated sequence and a p53 gene sequence of the same length. The ternary exact tracts in the randomly generated sequence are uniformly generated, similarly to a graph of an exponential function, where as they appear in the p53 gene, which is not a monotonically descending function, and consequently, there is a peak with the value 12 (Figures 7 and 8).

Table 4 presents several indications of the frequencies of ternary

Lengths Data		p=0.25			p=0.5			p=0.75		
		Expected Lengths		Result	Expected Lengths		Result	Expected Lengths		Result
Sequence length	Min-tract length	Random Sequence	Formula	Relative Difference (%)	Random Sequence	Formula	Relative Difference (%)	Random Sequence	Formula	Relative Difference (%)
10	1	1	1.33	24.8120301	1	1.3	23.0769231	3	3.4033	11.85026298
20	1	1.25	1.33	6.01503759	1.4	2	30	5.333	3.9964	33.44510059
50	1	1.33	1.33	0	2.09	2	4.5	3.0909	4	22.7275
20	5	0	5.333	100	0	5.999	100	10	7.838	27.58356724
50	5	0	5.333	100	0	6	100	9	7.9999	12.50140627
100	5	0	5.3333	100	5.5	6	8.33333333	8.1667	8	2.08375
1000	5	7	5.3333	31.2508203	6.31	6	5.16666667	9.1765	8	14.70625
10000	5	5.28	5.3333	0.99938125	6.13	6	2.16666667	8.0543	8	0.67875
20000	5	5.3	5.3333	0.6243789	6.015	6	0.25	8.0666	8	0.8325
40000	5	5.249	5.3333	1.58063488	6.0235	6	0.39166667	7.8913	8	1.35875
100000	5	5.117	5.3333	4.05565035	5.9705	6	0.49166667	8.0096	8	0.12

Table 1: Comparison of the expected tract lengths under various conditions.

Nucleotides Frequencies in p53			
Nucleotide	Frequency	Theoretical Probability	Difference
A	0.2584	0.25	0.0084
C	0.2448	0.25	-0.0052
G	0.2515	0.25	0.0015
T	0.2453	0.25	-0.0047
		σ	0.006375474

Table 2: Frequencies of the nucleotides in the given gene p53.

Binary Exact Tracts				
Exact Tract	Expected Length			Results
	Sequence Origin			
Symbol	Name	p53 Gene	Random	Diff
AC	Keto	6.6049	5.9385	0.6664
AG	Pyramidin	6.8898	6.0997	0.7901
AT	Strong	7.6508	6.009	1.6418
CG	Weak	6.0468	5.8696	0.1772
CT	Purines	6.5942	6.0031	0.5911
GT	Imino	6.4223	6.0228	0.3995
			σ	0.50418

Table 3: Comparing the Exact Tract Frequencies in the p53 gene to a randomly generated sequence (σ is 0.50418).

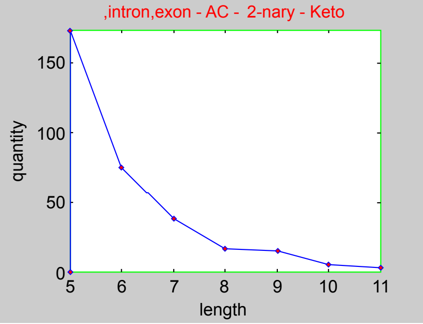


Figure 5: Exponentially distributed lengths of the binary (AC–Keto) Exact tracts of a Random sequence.

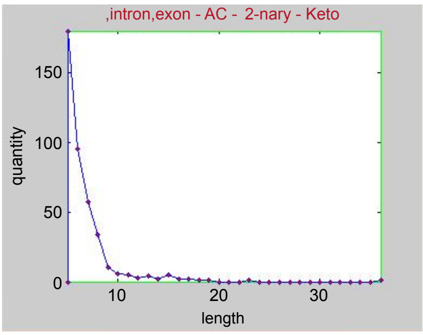


Figure 6: Distribution of the binary (AC–Keto) Exact tracts of p53.

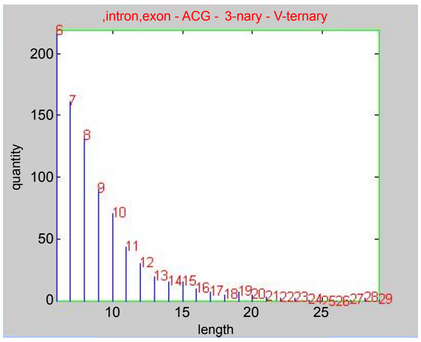


Figure 7: V-Ternary Exact Tracts' Distribution in a Random Sequence.

tracts in the chosen gene in comparison to the frequencies of ternary tracts in a randomly generated sequence.

The ternary tracts are over represented and especially long tracts appear more frequently in the gene than in the random sequence. In the random sequence, long ternaries (more than 20 nucleotides) are not seen in the random sequence at all.

Each ternary tract is composed of binary subtracts. Each binary tract is (by definition) a special case of a ternary tract. The decomposition of the ternary tract into its components, the binary tracts, is shown (Table 5), where the ternary tracts used for the demonstration are of the V type, longer than 20 nucleotides, which appear in the p53 gene. The longest one, whose length is 40 nucleotides, starts at nucleotide 11806 in the gene. The longest sub-tract in each tract is denoted by a pale (red) color and an italic font.

The question being investigated here is whether the properties of ternary tracts are due to their binary sub-tracts or are independent of the binary sub-tracts included in them.

Ternary Exact Tracts										
		Min. Tract Length-5			Min. Tract Length-10			Min. Tract Length-15		
Tract Designation		Expected Length (theoretical 8)		Results	Expected Length (theoretical 13)		Results	Expected Length (theoretical 18)		Results
		Sequence Origin			Sequence Origin			Sequence Origin		
Symbol	Name	p53 Gene	Random	Diff	p53 Gene	Random	Diff	p53 Gene	Random	Diff
ACG	V	8.5262	7.8877	0.6385	14.318	13.244	1.074	19.6635	18.5441	1.1194
ACT	H	9.0367	7.9486	1.0881	16.327	13.0667	3.2603	22.744	18.724	4.02
AGT	C	8.9774	8.074	0.9034	15.9188	13.0207	2.8981	22.7025	17.6264	5.0761
CGT	B	8.2962	8.0421	0.2541	14.1567	12.9412	1.2155	19.444	18.333	1.111
			σ	0.36186		σ	1.12808		σ	2.02831

Table 4: Comparison of a ternary exact tract frequency in a gene and in a random sequence.

No.	Begin	End	Length	Type	Tract
8	7031	7051	21	intron	CAGGAGGCAGAGGCAGGAGAA
9	13913	13933	21	intron	CAAAAAAAAAAAAAAAAGGCC
10	2003	2024	22	intron	CCCGGAGAAAAAAAAAAAAGAA
11	12636	12657	22	intron	CAAAGAGGCCAAGGCAGGCAGA
12	14591	14612	22	intron	AAGCAAGCAGGACAAGAAGCGG
13	14707	14728	22	exon	CCCCAGCCAAAGAAGAAACCAC
14	18529	18550	22	Intron	CAGGGAAAAGGGGCACAGACCC
15	1567	1589	23	Intron	GCCCGCCAGGCCGAGGAGGACCG
16	1954	1976	23	Intron	GCAGAAGGCAAGCCCGGAGGCAC
17	5376	5398	23	Intron	CAAAAAAAGAAAAAGAAAAAGGA
18	12882	12904	23	Intron	CAAAAAAAAAAAAAAAGAAAAGC
19	16914	16936	23	Intron	GCAGGGAGCCAAGACGGCGCCAC
20	6517	6540	24	Intron	CAAAAAAAAAAAAAGAAAAAGAAA
21	9377	9400	24	Intron	CAAAAAAAAAAAAAAACAGAAAAG
22	13125	13148	24	exon	CCACACCCCCGCCCGGCACCCGCG
23	1706	1730	25	Intron	GAGAGGGGAGGAGAGAGAGAGAAAA
24	16684	16708	25	Intron	CAAAAAAAGAAAAGGCCAGGCGCAC
25	17644	17669	26	exon	GGGAAGGAGCCAGGGGGGAGCAGGGC
26	2157	2183	27	Intron	GAAGCGGAAGGGGCGGGCCCGCAGGCG
27	4603	4629	27	Intron	CAGAAAAAAAAAGAAAGAAAGAAAAAA
28	7746	7773	28	Intron	GCACACCACGCCGGGCAACAGAGCGAGA
29	9844	9871	28	Intron	CCAAAAAAAAAAAGAAAAAGAAAAAGAC
30	10018	10046	29	Intron	CAAAAGAAAAAAGAAAGAAAGAAAGAACA
31	14504	14532	29	exon	GGGAGAGACCGGCGCACAGAGGAAGAGAA
32	16974	17003	30	Intron	CAGAAAAAAAAGAAAAGAAACGAGGCACAG
33	9478	9508	31	Intron	GAAAAAAAAAAAAAGAAAAAGAAAGAGAGCA
34	7139	7171	33	Intron	CAAAAAAAAAAAAAAAAAAGGAAAGAAAA AAAA
35	6100	6133	34	Intron	CAAAAAAAAAAAAAAAAAAAAAAGAAAAG AAAAC
36	10332	10367	36	Intron	CAAAAAACAAACAAAAAAACAAAA-CAAAAAAAAACA
37	11806	11845	40	Intron	GGAAGGGCAGGCCACCACCCC-GACCCCAACCCCAGCCCCC

Table 5: The ternary tracts of the V type, their location, and their decomposition.

In order to answer this question, the following examination was performed. The pure exact tract was defined and a formula (5.7.2) was derived to approximate the frequencies of the pure exact tracts. In addition, an algorithm was developed to filter only those ternary exact tracts that are pure, i.e., they do not include binary sub-tracts longer than some specified length.

A question that arises here is: are the ternary tracts that appear in the gene dominant? As mentioned before, it is obvious that each ternary tract is composed of binary subtracts, so maybe from a biological point of view, the binary subtracts included in the ternary tracts are responsible for the biological phenomena in the unwinding process [12] and the ternary tracts that compose the binary one have no special purpose.

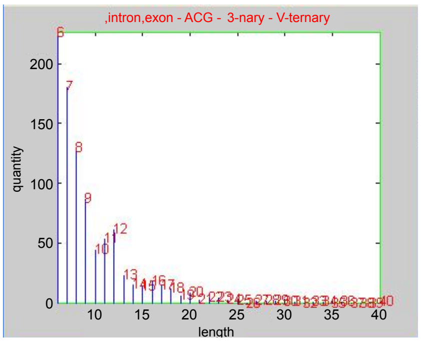


Figure 8: V-Ternary Exact Tracts' Distribution in a p53 Gene DNA Sequence.

The length of the binary subtracts included in the ternary tracts was investigated (Table 5). The chosen represented tract is the one whose number in the ordered list of V tracts and whose length is greater than 4 is 1196. The longest subtract in this tract is the M-sub-tract (Table 6) having a length of 12 nucleotides, which is 32.5% of the whole tract.

Analysis of the frequencies of palindromes

The expected palindrome length of the p53 gene is compared with the random sequence of nucleotides (Table 7).

The comparison demonstrates that the difference between the quantities of palindromes of the two sequences of the same length, the p53 gene, and the random sequence are of the same order of magnitude. The classification of the palindromes to types is performed according to the types of the nucleotides composing the palindromes. The type case in which A (adenine) is missing means that the palindrome consists only of the nucleotides {C, G, T}.

Programs

MATFREQ Program

The MATFREQ [7] (Figure 9) program for tract frequency analysis was developed especially for parsing the genes, searching for the tracts, investigating them, generating pseudo-random sequences, and comparing the results with the given gene sequence and with the developed formula. The results shown here are produced by MATFREQ.

Palindrome program

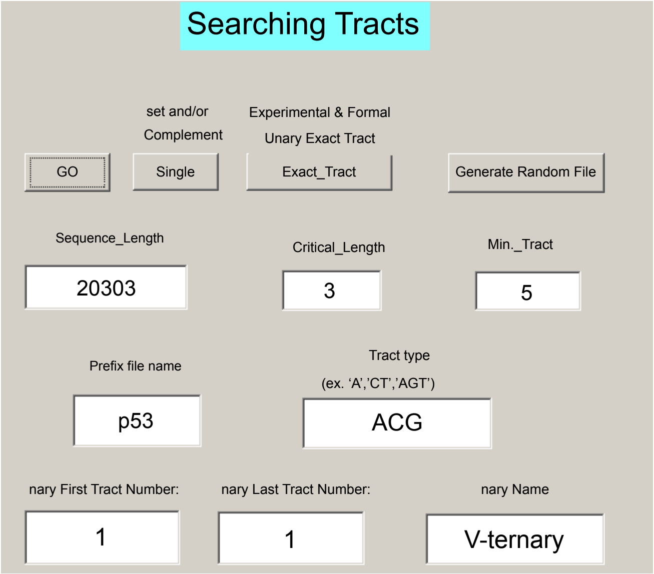
The GUI of the programs is presented in Figures 10a and 10b.

No.	Begin	End	Length	Type	Code	Binary subtract
1	1	7	7	AG	R	GGAAGGG
2	5	8	4	CG	S	GGGC
3	5	7	3	G	H	GGG
4	9	11	3	AG	R	AGG
5	10	13	4	CG	S	GGCC
6	12	21	10	AC	M	CCACCACCCC
7	18	22	5	CG	S	CCCCG
8	18	21	4	C	D	CCCC
9	23	34	12	AC	M	ACCCCAACCCCA
10	24	27	4	C	D	CCCC
11	30	33	4	C	D	CCCC
12	35	40	6	CG	S	GCCCCC
13	36	40	5	C	D	CCCCC

Table 6: The decomposition of the longest V-ternary tract into its exact binary sub-tracts. The tract is shown at the top of the table.

p53 palindromes' quantities		
Type: Missing A		
Length	Number	Number
	p53	random
1	3576	3666
2	1329	1258
3	1602	1573
4	410	426
5	384	371
6	74	103
7	77	75
8	12	13
9	15	19
10	6	4
11	3	5
12	0	1
13	0	1
14	0	0
Expected length	2.124332	2.128809

Table 7: Comparison of the number of palindrome tracks (type A) in the p53 gene and in a random sequence of the same length.



Searching Tracts

set and/or Complement

Experimental & Formal Unary Exact Tract

GO

Single

Exact_Tract

Generate Random File

Sequence_Length: 20303

Critical_Length: 3

Min._Tract: 5

Prefix file name: p53

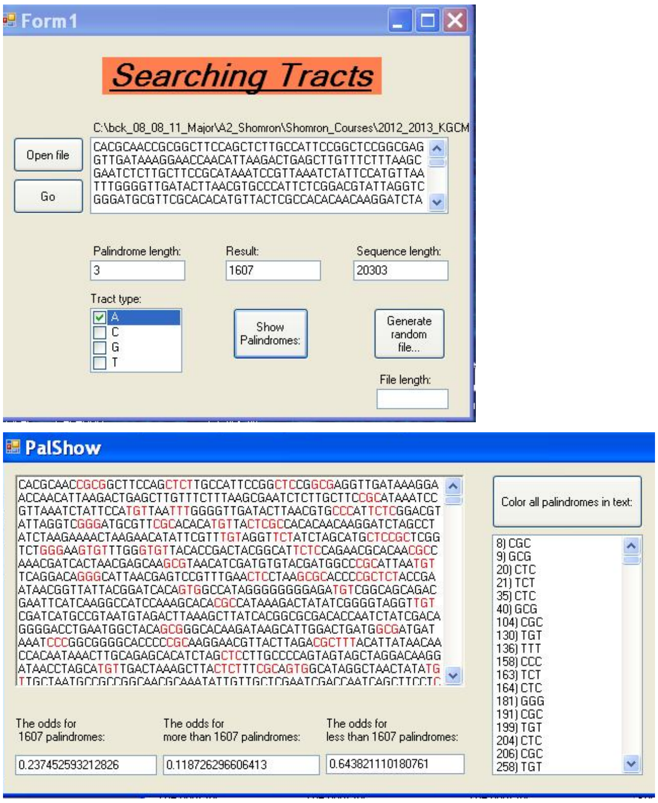
Tract type (ex. 'A','CT','AGT'): ACG

nary First Tract Number: 1

nary Last Tract Number: 1

nary Name: V-ternary

Figure 9: The GUI of MATFREQ.



Form1

Searching Tracts

C:\bck_08_08_11_Major\A2_Shomron\Shomron_Courses\2012_2013_KGCM

Open file

Go

```
CACGCAACCGCGGCTTCCAGCTCTTGCCATTCCGGCTCCGGCGAG
GTTGATAAAGGAACCAACATTAAGACTGAGCTTGTTTCTTTAAGC
GAATCTCTTGCTTCCGCATAAATCCGTTAAATCTATTCCATGTTAA
TTTGGGGTTGATACTTAACGTGCCCATTCTCGGACGTATTAGGTC
GGGATGCGTTCGCACACATGTTACTCGCCACACAACAAGGATCTA
```

Palindrome length: 3

Result: 1607

Sequence length: 20303

Tract type:

☑ A
□ C
□ G
□ T

Show Palindromes:

Generate random file...

File length:

PalShow

```
CACGCAACCGCGGCTTCCAGCTCTTGCCATTCCGGCTCCGGCGAGGTTGATAAAGGA
ACCAACATTAAGACTGAGCTTGTTTCTTTAAGCGAATCTCTTGCTTCCGCATAAATCC
GTTAAATCTATTCCATGTTAATTTGGGGTTGATACTTAACGTGCCCATTCTCGGACGT
ATTAGGTCGGGATGCGTTCGCACACATGTTACTCGCCACACAACAAGGATCTAGCCT
ATCTAAGAAAACTAAGAACATATTCGTTTGTAGGTTCTATCTAGCATGCTCCGCTCGG
TCTGGGAAGTGTTTGGGTGTTACACCGACTACGGCATTCTCCAGAACGCACAACGCC
AAACGATCACTAACGAGCAAGCGTAACATCGATGTGTACGATGGCCCGCATTAATGT
TCAGGACAGGGCATTAACGAGTCCGTTTGAACTCCTAAGCGCACCCCGCTCTACCGA
ATAACGGTTATTACGGATCACAGTGGCCATAGGGGGGGGAGATGTCGGCAGCAGAC
GAATTCATCAAGGCCATCCAAAGCACACGCCATAAAGACTATATCGGGGTAGGTTGT
CGATCATGCCGTAATGTAGACTTAAAGCTTATCACGGCGCGACACCAATCTATCGACA
GGGGACCTGAATGGCTACAGCGGGCACAAGATAAGCATTGGACTGATGGCGATGAT
AAATCCCGGCGGGGCACCCCCGCAAGGAACGTTACTTAGACGCTTTACATTATAACAA
CCACAATAAACTTGCAGAGCACATCTAGCTCCTTGCCCCAGTAGTAGCTAGGACAAGG
ATAACCTAGCATGTTGACTAAAGCTTACTCTTTCGCAGTGGCATAGGCTAACTATATG
TTGCTAATGCCGCCGGCAACGCAAATATTGTTGCTCGAATCGACCAATCAGCTTCCTC
```

Color all palindromes in text:

```
8) CGC
9) GCG
20) CTC
21) TCT
35) CTC
40) GCG
104) CGC
130) TGT
136) TTT
158) CCC
163) TCT
164) CTC
181) GGG
191) CGC
199) TGT
204) CTC
206) CGC
258) TGT
```

The odds for 1607 palindromes: 0.237452593212826

The odds for more than 1607 palindromes: 0.118726296606413

The odds for less than 1607 palindromes: 0.643821110180761

Figure 10: A program counting the number of palindromes of different types and of various lengths. (a) Input, (b) Output.

Discussion

The binary subtracts dominate the ternary tracts (Table 5), i.e. their length is more than 50% of the ternary tracts in which they reside (ternary tracts longer than 20 bases). The average uniformity of ternary tracts longer than 20 for the p53 gene is 0.24334, whereas the average uniformity of such tracts in a pseudo-randomly generated tract is 0.3924, namely, the uniformity in the p53 gene is 62.5% more than in the random sequence. This confirms the assumption that the ternary tracts' proprieties are received from the binary subtracts included in them. This means that general ternary tracts do not play any significant role in the gene. However, special ternary tracts, including a binary subtract sufficiently long, play a significant role.

Analysis and comparison of the palindrome track frequencies of various types in the p53 gene and in a random sequence resulted in not finding any significant differences.

References

1. Tamm C, Shapiro HS, Lipshitz R, and Chargaff E (1952) Distribution density of nucleotides within a deoxyribonucleic acid chain. J Bioi Chem 203: 673-698.
2. Chargaff E (1963) Essays in Nucleic Acids. Elsevier Publishing Corporation, Amsterdam.
3. Chao MT, Fu JC, Koutras MV (1995) Survey of reliability studies of consecutive-k-out—of-n: F and related systems. IEEE Transactions of Reliability 44: 120-127.
4. Bernardi G, Mouchiroud D, Gautier C (1988) Compositional patterns in vertebrate genomes, conservation and change in evolution. J Mol Evol 28: 7-18.
5. Burge C, Campbell AM, Karlin S (1993) Over and under representation of short oligonucleotides in DNA sequences. Proc Natl Acad Sci USA, 89: 1358-1362.

6. Gal M, Katz T, Ovadia A, Yagil G (2003) TRACTS: a program to map oligopurine, oligopyrimidine and other binary tracts. Nucleic Acids Res 31: 3682-3685.

7. Ophir D (2005) MATFREQ Exact_Tracts_Computation Frequencies–a MATLAB Program, Ariel–Academic College in Yudea and Samaria 4.

8. Feller W (1967) An Introduction to the probability theory and its Applications (3rd edn), John Wiley & Sons, USA.

9. Gera A, Ophir D (2013) A set of tests involving strings of successes and/or failures.

10. Yagil G (2004) The over-representation of binary DNA tracts in seven sequenced chromosomes. BMC Genomics 5: 19.

11. Shomer B, Yagil G (1999) Long W tracts are over-represented in the E. coli and H. influenza genomes. Nucleic Acid Res. 27: 4491-4480.

12. Yagil G, Shirnron F, Tal M (1998) DNA unwinding in the CYC1 and DEDI yeast promoters. Gene 225: 152-163.