TÜBİTAK

Turkish Journal of Mathematics

http://journals.tubitak.gov.tr/math/

Research Article

Turk J Math
(2017) 41: 80 – 95

doi:10.3906/mat-1509-19

New statistical randomness tests: 4-bit template matching tests

Fatih SULAK*
Department of Mathematics, Atılım University, Ankara, Turkey

Received: 04.09.2015 • **Accepted/Published Online:** 14.03.2016 • **Final Version:** 16.01.2017

Abstract: For cryptographic algorithms, secret keys should be generated randomly as the security of the system depends on the key and therefore generation of random sequences is vital. Randomness testing is done by means of statistical randomness tests. In this work, we show that the probabilities for the overlapping template matching test in the NIST test suite are only valid for a specific template and need to be recalculated for the other templates. We calculate the exact distribution for all 4-bit templates and propose new randomness tests, namely template matching tests. The new tests can be applied to any sequence of minimum length 5504 whereas the overlapping template matching test in the NIST test suite can only be applied to sequences of minimum length 10^6. Moreover, we apply the proposed tests to biased nonrandom data and observe that the new tests detect the nonrandom behavior of the generator even for a bias of 0.001, whereas the template matching tests in NIST cannot detect that bias.

Key words: Cryptography, overlapping template matching test, statistical randomness testing, NIST test suite

1. Introduction

Random sequences and random numbers are used in many fields, such as statistics, computer simulations, and cryptography. In cryptography, random sequences are needed for several applications, such as the generation of primes in RSA encryption, secret keys in symmetric encryption, challenges in challenge-response protocols, initialization vectors, or salts in hash functions, but the most common application is the generation of secret keys.

Secret keys should be generated randomly so that the best option of the attacker should not be better than trying all possible elements in the set from which the key was chosen. If an attacker narrows down the number of possible keys, then the protocol is assumed to be broken. In 1996, Goldberg and Wagner [8] showed that the "random numbers" used to generate the keys in the Netscape SSL protocol were based on the time of the processor and therefore predictable, which helped them to find a major weakness in the protocol. Thus, it is vital to use an algorithm that produces random numbers properly.

Ideally, random numbers should be produced by true random sources, like atmospheric noise, thermal noise, or noise in an electrical circuit. These generators are called true random number generators (TRNGs). However, producing random numbers by TRNGs is usually inefficient, and therefore deterministic algorithms are generally used to produce random numbers. These algorithms are called pseudorandom number generators (PRNGs).

An example of a PRNG is the linear congruential generator [18], which produces a pseudorandom sequence

*Correspondence: fatih.sulak@atilim.edu.tr

$x_1, x_2, x_3, \ldots$ using the linear recurrence

$$x_n = a \cdot x_{n-1} + b \ (mod\ m)$$

where x_0 is the seed and a, b, and m are parameters.

The output sequences of PRNGs should be statistically indistinguishable from truly random sequences; therefore, statistical analysis of PRNGs is crucial, and this analysis is performed by statistical randomness testing. In order to test a PRNG, first an output sample is produced, and then this sample is tested by various statistical randomness tests.

A test suite is a collection of statistical randomness test that are designed to tests the randomness properties of sequences. There are several test suites in the literature [14, 15, 20, 21]. Similarly there are several individual statistical randomness tests [1, 4, 6, 10–13, 17, 23].

The outputs of symmetric encryption algorithms should be indistinguishable from random sequences; that is, algorithms are expected to behave like PRNGs. Hence, their analysis from this point of view is crucial. Generally, a sample output set is taken from a symmetric encryption algorithm, and this set is evaluated in terms of randomness by a test suite.

The NIST test suite [21] is the most popular test suite for cryptographic applications. The statistical analysis of AES finalist algorithms was performed by Soto et al. using the NIST test suite [22]. Some tests in the suite require sequences of length 10^6, while the outputs of AES finalist algorithms are 128 bits. Soto et al. concatenated the outputs of the algorithms to obtain long sequences in order to apply all the tests. Recently, Sulak et al. proposed an alternative method where they computed and used the exact distributions instead of approximations or asymptotic distributions [24]. Having these exact probabilities, the necessity of long sequences was reduced, and they applied the randomness tests directly to the outputs of the algorithms instead of concatenating them.

There are several studies for the tests of the NIST test suite [5, 7, 9, 19, 25]. Okutomi et al. applied the tests in the NIST test suite to the random data taken from the cryptographic algorithms DES and SHA-1 [19]. They observed that Maurer's universal statistical test and the overlapping template matching test have problems with the ratio of the random data that pass the tests. Hamano et al. corrected the probabilities for the overlapping template matching test, where they took the template as $B = 111111111$ [9] and NIST updated the probabilities accordingly. However, as noted in [25], the probability of each pattern depends on the pattern itself. In this work, we set $m = 4$ and classify 16 possible patterns into four groups. Then, for each of the 16 patterns, we evaluate the exact probabilities using combinatorial approaches. Afterwards, we propose four new statistical randomness tests that can be applied to short sequences and long sequences. We observe that the probabilities are not the same for each overlapping template, which shows that the probabilities for overlapping template matching test in the NIST are valid only for $B = 111111111$. We apply the new tests to random data taken from various PRNGs and to nonrandom data to observe the power of the new tests, and we compare the results with the NIST test suite.

The organization of the paper is as follows. In Section 2, we give preliminaries. In Section 3, we obtain the exact distributions. In Section 4, we define new statistical randomness tests and state the corresponding bin probabilities. In Section 5, we apply the new tests to random and nonrandom data and observe the power of new tests. In Section 6, we conclude the paper by describing some future work.

2. Preliminaries

A statistical randomness test is a procedure that takes a binary sequence as an input and tests a null hypothesis (H_0) stating that the given input sequence is random. The test examines the input sequence, produces a real number between 0 and 1 that is called p-value, and accepts or rejects the hypothesis using a probabilistic approach. As it is probabilistic, the test may reject truly random sequences, and in that case, a type I error has occurred. The probability of such an error is called the level of significance of the test and denoted by α. If the p-value produced by the test is greater than α then H_0 is accepted; otherwise, it is rejected [18]. α is usually set to 0.01 for cryptographic applications [21].

The χ^2 distribution is used to compare how well the observed frequencies of events fit to the corresponding expected frequencies under the hypothesized distribution.

Definition 2.1 *[18] A random variable has a χ^2 distribution with degrees of freedom v if the corresponding probability density function $f(x) = 0$ for $x < 0$ and*

$$f(x) = \frac{1}{\Gamma(v/2)2^{v/2}} x^{\frac{v}{2}-1} e^{-\frac{x}{2}}, \quad x \geq 0$$

where v is a positive integer and Γ is the gamma function; that is, $\Gamma(t) = \int_0^\infty x^{t-1}e^{-x}dx$, for $t > 0$.

The χ^2 goodness of fit test is a statistical randomness test where the distribution of the test statistic follows χ^2 distribution, assuming H_0 is true. In other words, let E_i be the expected frequencies and F_i be the observed frequencies for $1 \leq i \leq k$. Then

$$\chi^2 = \sum_{i=1}^{k} \frac{(F_i - E_i)^2}{E_i} \quad \text{and} \quad p\text{-value} = \texttt{igamc}\left(2, \frac{\chi^2}{2}\right)$$

where igamc is the incomplete gamma function [21].

3. Four-bit template matching tests

The subject of the 4-bit template matching tests is the frequency of a prespecified template in a binary sequence. Similar tests are defined in the NIST test suite, namely the nonoverlapping template matching test and the overlapping template matching test. In both tests, first an m-bit template B is chosen, and the sequence subject to the test is divided into N subsequences of length M. An m-bit window is used to search the m-bit overlapping blocks of each subsequence. Then, for each block, the number of the template B in that subsequence is counted. Let W_i denote the number of B in the ith block. For the overlapping template matching test M is set to 1032. Let π_j denote the probability that $W_i = j$ for $0 \leq j \leq 4$ and π_5 denote the probability that $W_i \geq 5$. For $M = 1032$ and $B = 111111111$, the exact probabilities π_j are calculated in [9] using a recursion. A p-value is produced using the χ^2 goodness of fit test using those probabilities.

For the nonoverlapping template matching test, the prespecified template is chosen in such a manner that if the template is observed somewhere in the sequence, then it should not be seen before the template is completed. As noted in the NIST test suite, if the pattern is observed somewhere in the sequence, it cannot be observed again for the next $m - 1$ blocks, and hence the m-bit window slides m bits. This shows that the distribution is the same for all nonoverlapping templates. Using a similar idea, we have the following proposition.

Proposition 3.1 *The distribution of the frequency of a prespecified template depends only on the number of overlapping bits in the template.*

Proof Assume that the prespecified template B of length 4 has k overlapping bits. If B is observed somewhere in the sequence, the next $3-k$ blocks cannot be equal to B as B has k overlapping bits. The latter block may be equal to B with probability $\frac{1}{2^{3-k}}$. This shows that the distribution depends only on the number of overlapping bits in the block. □

Using this proposition, we classify the 4-bit templates according to their number of overlapping bits. There 4 types of blocks:

1. Nonoverlapping blocks: 0001, 0011, 0111, 1000, 1100, 1110
2. One-bit-overlapping blocks: 0010, 0100, 0110, 1001, 1011, 1101
3. Two-bit-overlapping blocks: 0101, 1010
4. Three-bit-overlapping blocks: 0000, 1111

We choose one representative block from each type and find the exact distributions. Different from the previous approaches, we assume that the bits are circular in each subsequence.

Example 3.2 *Let the subsequence be* 1000011000. *Then the number of* 001 *blocks is two, one starting from the fourth bit and one starting from the ninth bit, and the number of* 000 *blocks is three, starting from the second bit, the third bit, and the eighth bit.*

First we state some combinatorial formulae, which we will use in the calculation of probabilities.

Lemma 3.3 *[2] The number of nonnegative integer solutions of the equation* $x_1 + x_2 + \cdots + x_b = a$ *is* $\binom{a+b-1}{b-1}$.

Lemma 3.4 *[2] The number of integer solutions of the equation* $x_1+x_2+\cdots+x_b = a$ *with* $x_i \geq c$ *for* $1 \leq i \leq b$ *is* $\binom{a-b(c-1)-1}{b-1}$.

Proof With the substitution $x_i = y_i + c$, we get

$$\begin{aligned} (y_1 + c) + (y_2 + c) + \cdots + (y_b + c) &= a \\ y_1 + y_2 + \cdots + y_b &= a - bc. \end{aligned}$$

From Lemma 3.3 it follows that the number of solutions is:

$$\binom{(a-bc)+b-1}{b-1} = \binom{a-b(c-1)-1}{b-1}.$$

□

Lemma 3.5 *[2] [Inclusion - Exclusion Principle] The number of nonnegative integer solutions of the equation* $x_1 + x_2 + \cdots + x_b = a$ *with* $x_i \leq c$ *for* $1 \leq i \leq b$ *is* $\sum_{j=0}^{b} \binom{a+b-1-j(c+1)}{b-1} \binom{b}{j} (-1)^j$.

3.1. Nonoverlapping case

In order to define a randomness test, we need to find the probability that the prespecified template B occurs k times in the subsequence. For the nonoverlapping case, we choose $B = 0001$ and compute the probability accordingly. We assume that we know the weight W and the number of runs V of the sequence.

Theorem 3.6 *Let $\{a_1, a_2, \ldots, a_n\}$ be a binary sequence and $b_i = a_i a_{i+1} a_{i+2} a_{i+3}$ be blocks of length 4 for $1 \leq i \leq n$ with $a_{n+j} = a_j$ for $j = 1, 2, 3$, and let K denote the number of 0001 blocks among b_i for $1 \leq i \leq n$. Also let w be the weight of the sequence and $2r$ be the number of runs in the sequence. If the sequence is not all zero or all one then*

$$Pr(K = k) = \frac{n}{r \cdot 2^n} \binom{w-1}{r-1} \binom{r}{k} \sum_{a=0}^{r-k} \binom{r-k}{a} \binom{n-w-r-a-k-1}{k-1}.$$

Proof First note that the number of runs is even if the sequence is not all zero or all one. We assume the bits are arranged on a circle and we write 'one's and 'zero's consecutively to define $2r$ runs. As a result, $w - r$ 'one's and $n - w - r$ 'zero's remain.

As all the 0001 blocks contain 01 blocks, if a run of 'zero's has more than 2 'zero's, it produces exactly one 0001 block. Now we find the distribution of $w - r$ many 'one's and $n - w - r$ many 'zero's so that the number of 0001 blocks is k. The number of such arrangements is equal to the number of nonnegative integer solutions of the system

$$\begin{aligned} x_1 + x_2 + \cdots + x_r &= n - w - r \\ y_1 + y_2 + \cdots + y_r &= w - r \end{aligned}$$

with an additional condition that exactly k of x_is satisfy $x_i \geq 2$ for $1 \leq i \leq r$. This additional condition guarantees that there are exactly k many 0001 blocks. The second equation has $\binom{w-1}{r-1}$ solutions by Lemma 3.3.

$$\underbrace{x_1 + \cdots + x_k}_{\geq 2} + \underbrace{x_{k+1} + \cdots + x_{k+a}}_{=1} + \underbrace{x_{k+a+1} + \cdots + x_r}_{=0} = n - w - r$$

In order to find the number of solutions of the first equation, we may assume that $x_i \geq 2$ for $1 \leq i \leq k$ (with a factor $\binom{r}{k}$), $x_j = 1$ for $k+1 \leq j \leq k+a$ (with a factor of $\binom{r-k}{a}$), and $x_s = 0$ for $k+a+1 \leq s \leq r$; hence, the number of integer solutions of the first equation is

$$x_1 + x_2 + \cdots + x_k = n - w - r - a, \quad x_i \geq 2, \quad 1 \leq i \leq k,$$

which is $\binom{n-w-r-a-k-1}{k-1}$ by Lemma 3.4. Note that if $k = 0$, we cannot apply Lemma 3.4, and in that case we assume there is only one solution. We use this assumption throughout the paper.

Moreover, each arrangement on the circle gives n sequences. However, since there are r many 01 blocks, r of these sequences are identical. Thus, considering the circular symmetry, other than an all-zero or all-one

sequence, we have

$$Pr(K=k)=\frac{n}{r\cdot 2^n}\binom{w-1}{r-1}\binom{r}{k}\sum_{a=0}^{r-k}\binom{r-k}{a}\binom{n-w-r-a-k-1}{k-1}.$$

□

Example 3.7 *Assume that* $n=8$, $w=3$, $r=2$, $k=1$. *We need to find the number of integer solutions of the system*

$$\begin{aligned} x_1+x_2 &= 3\\ y_1+y_2 &= 1 \end{aligned}$$

where $x_1=3$, $x_2=0$, $y_1=0$, $y_2=1$ *is a solution. The corresponding sequence is obtained as:*

$$\underbrace{0000}_{x_1=3}\underbrace{1}_{y_1=0}\underbrace{0}_{x_2=0}\underbrace{11}_{y_2=1}.$$

Note that as $x_1\geq 2$, *it produces exactly one* 0001 *block. We show all the solutions and the corresponding sequences in Table 1.*

Table 1. An example for Theorem 3.6.

	$y_1=0,\ y_2=1$	$y_1=1,\ y_2=0$
$x_1=3,\ x_2=0$	00001011	00001101
$x_1=2,\ x_2=1$	00010011	00011001
$x_1=1,\ x_2=2$	00100011	00110001
$x_1=0,\ x_2=3$	01000011	01100001

Moreover, each arrangement gives 8 *sequences, and two sequences are always identical. Consider the solution* $x_1=3$, $x_2=0$, $y_1=0$, $y_2=1$ *and its corresponding sequence* 00001011. *It produces* 00010110, 00101100, 01011000, 10110000, 01100001, 11000010, *and* 10000101. *However, note that we also obtain* 01100001 *as the corresponding sequence of the solution* $x_1=0$, $x_2=3$, $y_1=0$, $y_2=1$. *As a result, there are* $\frac{8\cdot 8}{2}=32$ *sequences, which is consistent with Theorem 3.6.*

In the template matching test, we need to find the probabilities independent of weight and number of runs of the sequence. For this reason, we state the following corollary.

Corollary 3.8 *Let* $\{a_1,a_2,\ldots,a_n\}$ *be a binary sequence and* $b_i=a_ia_{i+1}a_{i+2}a_{i+3}$ *be blocks of length 4 for* $1\leq i\leq n$ *with* $a_{n+j}=a_j$ *for* $j=1,2,3$, *and let* K *denote the number of* 0001 *blocks among* b_i *for* $1\leq i\leq n$. *If the sequence is not all zero or all one then*

$$Pr(K=k)=\sum_{w=1}^{n-1}\sum_{r=1}^{\lfloor n/2\rfloor}\frac{n}{r\cdot 2^n}\binom{w-1}{r-1}\binom{r}{k}\sum_{a=0}^{r-k}\binom{r-k}{a}\binom{n-w-r-a-k-1}{k-1}.$$

Proof Since we compute $Pr(K=k|W=w,V=2r)$ in Theorem 3.6, by summing over all possible weights and runs, we obtain $Pr(K=k)$. □

3.2. One-bit-overlapping case

For a one-bit-overlapping case, we choose $B = 0110$ and obtain the probability accordingly.

Theorem 3.9 *Let $\{a_1, a_2, \ldots, a_n\}$ be a binary sequence and $b_i = a_i a_{i+1} a_{i+2} a_{i+3}$ be blocks of length 4 for $1 \leq i \leq n$ with $a_{n+j} = a_j$ for $j = 1, 2, 3$, and let K denote the number of 0110 blocks among b_i for $1 \leq i \leq n$. Also let w be the weight of the sequence and $2r$ be the number of runs in the sequence. If the sequence is not all zero or all one then*

$$Pr(K = k) = \frac{n}{r \cdot 2^n} \binom{n-w-1}{r-1} \binom{r}{k} \sum_{a=0}^{r-k} \binom{r-k}{a} \binom{w-2r+a-1}{r-k-a-1}.$$

Proof Using an idea similar to the proof of Theorem 3.6, we assume the bits are arranged on a circle and we write 'one's and 'zero's consecutively to define $2r$ runs. As a result, $w-r$ 'one's and $n-w-r$ 'zero's remain.

As all the 0110 blocks contain 01 blocks, if a run of 'one's contains exactly two 'one's, it produces exactly one 0110 block. Therefore, we need to find the distribution of $w-r$ many 'one's and $n-w-r$ many 'zero's so that the number of 0110 blocks is k. The number of such arrangements is equal to the number of nonnegative integer solutions of the system

$$\begin{aligned} x_1 + x_2 + \cdots + x_r &= n - w - r \\ y_1 + y_2 + \cdots + y_r &= w - r \end{aligned}$$

with an additional condition that exactly k of $y_i = 1$ for $1 \leq i \leq r$. This additional condition guarantees that there are exactly k many 0110 blocks. The first equation has $\binom{n-w-1}{r-1}$ solutions by Lemma 3.3.

$$\underbrace{y_1 + \cdots + y_k}_{=1} + \underbrace{y_{k+1} + \cdots + y_{k+a}}_{=0} + \underbrace{y_{k+a+1} + \cdots + y_r}_{\geq 2} = w - r$$

In order to find the number of solutions of the second equation, we assume that $y_i = 1$ for $1 \leq i \leq k$ (with a factor of $\binom{r}{k}$), $y_j = 0$ for $k+1 \leq j \leq k+a$ (with a factor of $\binom{r-k}{a}$), and $y_s \geq 2$ for $k+a+1 \leq s \leq r$. In other words, we need to find the integer solutions of the equation

$$y_{k+a+1} + y_{k+a+2} \cdots + y_r = w - r - k, \quad y_s \geq 2, \ k+a+1 \leq s \leq r$$

, which is $\binom{w-2r+a-1}{r-k-a-1}$ by Lemma 3.4. Similar to the proof of the previous theorem, considering the circular symmetry we obtain

$$Pr(K = k) = \frac{n}{r \cdot 2^n} \binom{n-w-1}{r-1} \binom{r}{k} \sum_{a=0}^{r-k} \binom{r-k}{a} \binom{w-2r+a-1}{r-k-a-1}.$$

□

Corollary 3.10 *Let* $\{a_1, a_2, \ldots, a_n\}$ *be a binary sequence and* $b_i = a_i a_{i+1} a_{i+2} a_{i+3}$ *be blocks of length 4 for* $1 \leq i \leq n$ *with* $a_{n+j} = a_j$ *for* $j = 1, 2, 3$, *and let* K *denote the number of* 0110 *blocks among* b_i *for* $1 \leq i \leq n$. *If the sequence is not all zero or all one then*

$$Pr(K = k) = \sum_{w=1}^{n-1} \sum_{r=1}^{\lfloor n/2 \rfloor} \frac{n}{r \cdot 2^n} \binom{n-w-1}{r-1} \binom{r}{k} \sum_{a=0}^{r-k} \binom{r-k}{a} \binom{w-2r+a-1}{r-k-a-1}.$$

Proof Since we compute $Pr(K = k | W = w, V = 2r)$ in Theorem 3.9, by summing over all possible weights and runs, we obtain $Pr(K = k)$. □

3.3. Two-bit-overlapping case

In this case we choose the prespecified block as 1010. Different from Theorem 3.6 and Theorem 3.9, to obtain the probabilities, we use another model where x_is are modeled as red boxes and y_is are modeled as white boxes.

Theorem 3.11 *Let* $\{a_1, a_2, \ldots, a_n\}$ *be a binary sequence and* $b_i = a_i a_{i+1} a_{i+2} a_{i+3}$ *be blocks of length 4 for* $1 \leq i \leq n$ *with* $a_{n+j} = a_j$ *for* $j = 1, 2, 3$, *and let* K *denote the number of* 1010 *blocks among* b_i *for* $1 \leq i \leq n$. *Also let* w *be the weight of the sequence and* $2r$ *be the number of runs in the sequence. If the sequence is not all zero or all one then*

$$Pr(K = k) = \frac{n}{r \cdot 2^n} \sum_{a=k}^{r-1} \binom{r}{a} \binom{a}{k} \binom{n-w-r-1}{r-a-1} \binom{w-a-1}{r-k-1}.$$

Proof Using an idea similar to the proof of Theorem 3.6, we assume the bits are arranged on a circle and we write 'one's and 'zero's consecutively to define $2r$ runs. As a result, $w-r$ 'one's and $n-w-r$ 'zero's remain.

Now we find the distribution of $n-w-r$ many 'zero's and $w-r$ many 'one's so that the number of 1010 blocks is k. We use another model to solve this problem. Assume that there are r red-white box pairs and we distribute $w-r$ balls into white boxes and $n-w-r$ balls into red boxes. $Pr(K = k)$ is the probability that exactly k pairs are empty.

In other words, we find the number of nonnegative integer solutions of the system

$$\begin{aligned} x_1 + x_2 + \cdots + x_r &= n - w - r \\ y_1 + y_2 + \cdots + y_r &= w - r \end{aligned}$$

with the condition that $x_i + y_i = 0$ is satisfied for exactly k different values of i where $1 \leq i \leq r$.

Let a of the red boxes x_i be empty. Assume that the empty boxes are the first a boxes (with a factor of $\binom{r}{a}$). If we consider the first a white boxes, k of them should be empty. Assume that the empty boxes are the first k boxes (with a factor of $\binom{a}{k}$). Now we need to find the number of integer solutions of the system

$$\begin{aligned} \underbrace{x_1 + \cdots + x_a}_{=0} + \underbrace{x_{a+1} + \cdots + x_r}_{\geq 1} &= n - w - r \\ \underbrace{y_1 + \cdots + y_k}_{=0} + \underbrace{y_{k+1} + \cdots + y_a}_{\geq 1} + \underbrace{y_{a+1} + \cdots + y_r}_{\geq 0} &= w - r. \end{aligned}$$

The first equation has $\binom{n-w-r-1}{r-a-1}$ solutions and the second equation has $\binom{w-a-1}{r-k-1}$ solutions by Lemma 3.4.

Considering the circular symmetry, other than an all-zero or all one-sequence, we have

$$Pr(K=k)=\frac{n}{r\cdot 2^n}\sum_{a=k}^{r-1}\binom{r}{a}\binom{a}{k}\binom{n-w-r-1}{r-a-1}\binom{w-a-1}{r-k-1}.$$

□

Corollary 3.12 *Let $\{a_1,a_2,\ldots,a_n\}$ be a binary sequence and $b_i=a_ia_{i+1}a_{i+2}a_{i+3}$ be blocks of length 4 for $1\le i\le n$ with $a_{n+j}=a_j$ for $j=1,2,3$, and let K denote the number of 1010 blocks among b_i for $1\le i\le n$. If the sequence is not all zero or all one then*

$$Pr(K=k)=\sum_{w=2}^{n-2}\sum_{r=2}^{\lfloor n/2\rfloor}\frac{n}{r\cdot 2^n}\sum_{a=k}^{r-1}\binom{r}{a}\binom{a}{k}\binom{n-w-r-1}{r-a-1}\binom{w-a-1}{r-k-1}.$$

Proof Since we compute $Pr(K=k|W=w,V=2r)$ in Theorem 3.11, by summing over all possible weights and runs, we obtain $Pr(K=k)$. □

3.4. Three-bit-overlapping case

We choose the prespecified block as 1111 for the three-bit-overlapping case. We apply the inclusion-exclusion principle to obtain the probability.

Theorem 3.13 *Let $\{a_1,a_2,\ldots,a_n\}$ be a binary sequence and $b_i=a_ia_{i+1}a_{i+2}a_{i+3}$ be blocks of length 4 for $1\le i\le n$ with $a_{n+j}=a_j$ for $j=1,2,3$, and let K denote the number of 1111 blocks among b_i for $1\le i\le n$. Also let w be the weight of the sequence and $2r$ be the number of runs in the sequence. If the sequence is not all zero or all one then*

$$\begin{aligned}Pr(K=k) &= \frac{n}{r\cdot 2^n}\binom{n-w-1}{r-1}\sum_{t=0}^{r}\binom{r}{t}\binom{k-1}{t-1}\\ &\cdot \sum_{i=0}^{r}\binom{w-k-3t-3i-1}{r-t-1}\binom{r-t}{i}(-1)^i.\end{aligned}$$

Proof Using an idea similar to the proof of Theorem 3.6, we assume the bits are arranged on a circle and we write 'one's and 'zero's consecutively to define $2r$ runs. As a result, $w-r$ 'one's and $n-w-r$ 'zero's remain. Now we find the distribution of $n-w-r$ many 'zero's and $w-r$ many 'one's so that the number of 1111 blocks is k and $V=2r$. Similarly, we should find the the number of integer solutions of the system

$$\begin{aligned}x_1+x_2+\cdots+x_r &= n-w-r,\quad x_i\ge 0,\ 1\le i\le r\\ y_1+y_2+\cdots+y_r &= w-r,\quad y_j\ge 0,\ 1\le j\le r,\end{aligned}$$

which should also satisfy that the number of 1111 blocks is k.

The first equation has $\binom{n-w-1}{r-1}$ solutions by Lemma 3.3. Without loss of generality (or multiplying by $\binom{r}{t}$), assume $y_j \geq 3$ for $1 \leq j \leq t$, and $0 \leq y_s \leq 2$ for $t+1 \leq s \leq r$. We find the number of solutions of the second equation in two parts:

Each run of 'one's with length $l \geq 4$ defines $l-3$ many 1111 blocks. Hence, in order to have k many 1111 blocks, we should have

$$\begin{aligned} y_1 - 2 + y_2 - 2 + \cdots + y_t - 2 &= k, \quad y_j \geq 3 \\ y_1 + y_2 + \cdots + y_t &= k + 2t, \quad y_j \geq 3 \end{aligned}$$

and, therefore, the number of solutions is $\binom{k-1}{t-1}$ (and $t = 0 \Leftrightarrow k = 0$) by Lemma 3.4.

As the weight of the sequence is w, we have:

$$y_{t+1} + y_{t+2} + \cdots + y_r = w - r - k - 2t, \quad 0 \leq y_j \leq 2, \ t+1 \leq j \leq r.$$

We apply the inclusion-exclusion principle to find the number of solutions and obtain:

$$\sum_{i=0}^{r} \binom{w-k-3t-3i-1}{r-t-1} \binom{r-t}{i} (-1)^i.$$

Considering the circular symmetry, other than an all-zero or all-one sequence, we have

$$\begin{aligned} Pr(K=k) &= \frac{n}{r \cdot 2^n} \binom{n-w-1}{r-1} \sum_{t=0}^{r} \binom{r}{t} \binom{k-1}{t-1} \\ &\cdot \sum_{i=0}^{r} \binom{w-k-3t-3i-1}{r-t-1} \binom{r-t}{i} (-1)^i. \end{aligned}$$

□

Note that we may use other models to prove the theorem, but the inclusion-exclusion principle can be generalized to other m values. We can still apply the inclusion-exclusion principle if we choose $m = 9$ and $B = 111111111$.

Corollary 3.14 *Let $\{a_1, a_2, \ldots, a_n\}$ be a binary sequence and $b_i = a_i a_{i+1} a_{i+2} a_{i+3}$ be blocks of length 4 for $1 \leq i \leq n$ with $a_{n+j} = a_j$ for $j = 1, 2, 3$, and let K denote the number of 1111 blocks among b_i for $1 \leq i \leq n$. If the sequence is not all zero or all one then*

$$\begin{aligned} Pr(K=k) &= \sum_{w=1}^{n-1} \sum_{r=1}^{\lfloor n/2 \rfloor} \frac{n}{r \cdot 2^n} \binom{n-w-1}{r-1} \sum_{t=0}^{r} \binom{r}{t} \binom{k-1}{t-1} \\ &\cdot \sum_{i=0}^{r} \binom{w-k-3t-3i-1}{r-t-1} \binom{r-t}{i} (-1)^i. \end{aligned}$$

Proof Since we compute $Pr(K = k | W = w, V = 2r)$ in Theorem 3.13, by summing over all possible weights and runs, we obtain $Pr(K = k)$. □

4. Test descriptions

The subject of the 4-bit template matching tests is the number of a prespecified template in a sequence. We apply the χ^2 goodness of fit test to measure how well the observed values fit the expected values. For this purpose, we divide the sequence into 128-bit blocks and find the number of occurrences of the template in each block. Afterwards, we apply the χ^2 test with 5 bins and produce the p-value using Table 2. The probabilities in Table 2 are evaluated using Corollaries 3.8, 3.10, 3.12, and 3.14.

Table 2. Bin probabilities for 4-bit template matching tests.

0-bit		1-bit		2-bit		3-bit	
0–6	0.24205627	0–5	0.16353105	0–5	0.19990529	0–4	0.21976555
7	0.17082354	6–7	0.27433485	6–7	0.25650023	5–6	0.18737326
8	0.18629989	8	0.15466485	8–9	0.25566342	7–8	0.18572664
9	0.16401892	9–10	0.24482145	10–11	0.16931656	9–10	0.15200493
10–128	0.23680138	11–128	0.16264780	12–128	0.11861450	11–128	0.25512962

Assume that we want to test a binary sequence of length n using the template matching test. We can summarize the procedure as follows:

- Choose a 4-bit template B.
- Divide the sequence into $M = \left\lfloor \frac{n}{128} \right\rfloor$ many 128-bit blocks.
- For each block, write the first 3 bits to the end of the sequence.
- Find the occurrence of B among the first block in an overlapping manner and increment the corresponding bin value, calling them F_i for $1 \leq i \leq 5$. Repeat the same procedure for all blocks.
- Apply the χ^2 goodness of fit test; that is, evaluate

$$\chi^2 = \sum_{i=1}^{5} \frac{(F_i - M \cdot p_i)^2}{M \cdot p_i} \quad \text{and} \quad p\text{-value} = \texttt{igamc}\left(2, \frac{\chi^2}{2}\right)$$

  where p_is are obtained from Table 2 according to the number of overlapping bits in template B.
- If the p-value is < 0.01, conclude as nonrandom, else conclude as random.

Let us demonstrate the template matching test using a simple example.

Example 4.1 *Assume the sequence subject to the template matching test is*

$$\{0,0,1,0,0,1,0,1,0,1,0,0,1,1,0,1,1,0,1,1,0,1,0,0\}$$

and assume that we choose a template $B = 0010$. *Note that* B *is a one-bit-overlapping block. We divide the sequence into three* 8*-bit blocks and extend the blocks.*

- *Block 1:* $\{0,0,1,0,0,1,0,1,0,0,1\}$, *there are two occurrences of* B, *the first one starts from the first bit and the second one starts from the fourth bit.*

- *Block 2:* $\{0,1,0,0,1,1,0,1,0,1,0\}$, B *does not occur in this block.*
- *Block 3:* $\{1,0,1,1,0,1,0,0,1,0,1\}$, *there is one occurrence of* B *starting from the seventh bit.*

As a result we find that $F_1 = 1$, $F_2 = 1$, *and* $F_3 = 1$.

In this example we divide the sequence into 8-bit blocks instead of 128-bit blocks. Note that we use different values for F_is and thus we cannot produce a p-value using Table 2, as the block size is not 128. The pseudocode of the test is stated in Algorithm 4.1. As the expected number of items in each bin should be at least 5, and the minimum probability in Table 2 is 0.1186145, the sequence subject to the test should be at least $128 \cdot \left\lfloor \dfrac{5}{0.1186145} \right\rfloor = 5504$ bits.

Algorithm 4.1: TEMPLATE MATCHING TEST$(\{a_1, a_2, \ldots, a_n\}, B)$

$F_1 = 0,\ F_2 = 0,\ F_3 = 0,\ F_4 = 0,\ F_5 = 0;$
$M = \left\lfloor \frac{n}{128} \right\rfloor;$
for $i \leftarrow 0$ **to** $M - 1$
  do
  for $j \leftarrow 1$ **to** 128
    do
    $b_j = a_{128i+j};$
  $b_{129} = a_{128i+1}, b_{130} = a_{128i+2}, b_{131} = a_{128i+3};$
  $count = 0;$
  for $j \leftarrow 1$ **to** 128
    do
    if $b_j b_{j+1} b_{j+2} b_{j+3} = B$
      then $count++;$
  Increment F_i *according to Table 2;*

Apply χ^2 *of goodness of fit test to* $F_1,\ F_2,\ F_3,\ F_4,\ F_5;$
return $(p - value)$

Example 4.2 *Assume the sequence subject to the template matching test is the first 5504 bits of the binary expansion of* π *and assume that we choose the template as* $B = 1111$. *Note that* B *is a three-bit-overlapping block. We divide the sequence into forty-three* 128*-bit blocks, and we find that* $F_1 = 11$, $F_2 = 7$, $F_3 = 10$, $F_4 = 6$, *and* $F_5 = 9$. *Using Table 2, we find the p-value as 0.861602.*

5. Simulation results

In this section, we apply the new statistical randomness tests to various sequences. We compare the new tests with the randomness tests in the NIST test suite [21]. There are 16 possible templates, and we choose a sample template from each class. For the nonoverlapping case we choose $B = 0001$, for one-bit-overlapping cases we choose $B = 0010$, for two-bit-overlapping cases we choose $B = 0101$, and for three-bit-overlapping cases we choose $B = 1111$. For the tests in the NIST test suite, we choose $M = 128$ for Frequency Test within a Block, $M = 10^4$ for Test for the Longest Run of Ones in a Block, $M = 32$ for Binary Matrix Rank Test, $m = 9$ and $B = 000000001$ for Nonoverlapping Template Matching Test, $m = 9$ and $B = 111111111$ for Overlapping

Template Matching Test, $L = 7$ and $L = 10$ for Maurer's Universal Statistical Test, $M = 500$ for Linear Complexity Test, $m = 16$ for Serial Test, $m = 14$ and $m = 14$ for Approximate Entropy Test, $state = 1$ for Random Excursions Test, and $state = -1$ for Random Excursions Variant Test. We produce two p-values for the Serial Test and the Cumulative Sums Test.

Table 3. Test results for the binary expansions of e, π, $\sqrt{2}$, and $\sqrt{3}$.

	e	π	$\sqrt{2}$	$\sqrt{3}$
Frequency	0.953749	0.578211	0.811881	0.610051
Block Freq	0.211072	0.380615	0.833222	0.473961
Runs	0.561917	0.419268	0.313427	0.261123
Long Run of Ones	0.718945	0.024390	0.012117	0.446726
Bin Matrix Rank	0.306156	0.083553	0.823810	0.314498
Nonover Temp	0.078790	0.165757	0.569461	0.532235
Over Temp	0.110434	0.296897	0.791982	0.082716
Maurer Univ	0.282568	0.669012	0.130805	0.165981
Linear Comp	0.826335	0.255475	0.317127	0.346469
Serial Test 1	0.766182	0.143005	0.861925	0.157500
Serial Test 2	0.462921	0.034354	0.629225	0.171100
App Entropy	0.700073	0.361595	0.884740	0.180481
CuSum Forw	0.669886	0.628308	0.879009	0.917121
CuSum Back	0.724265	0.663369	0.957206	0.689519
Rand Excur	0.786868	0.844143	0.216235	0.783283
Rand Excur Var	0.826009	0.760966	0.566118	0.798247
0-bit Temp (0001)	0.766497	0.975645	0.383993	0.884000
1-bit Temp (0010)	0.903124	0.717759	0.898930	0.849536
2-bit Temp (0101)	0.631473	0.981607	0.508969	0.236139
3-bit Temp (1111)	0.907699	0.294869	0.839803	0.600553

First, we apply the randomness tests to the binary expansions of e, π, $\sqrt{2}$, and $\sqrt{3}$. For this purpose, we produce approximately 10^6 bits ($7812 \times 128 = 999936$ bits) from the binary expansions of each number and apply the randomness tests. These four sequences are random according to all tests. The test results are presented in Table 3.

Second, we apply the randomness tests to four PRNGs. The random data are taken from the Mersenne Twister [16], the Random and RNGCryptoServiceProvider classes of $C\#$, and the outputs of AES [3]. The data from AES are produced using a fixed random key and low weight inputs. For both PRNGs $2^{17} \times 128 = 2^{24}$ bits are tested. Similar to the previous experiment, both generators pass all the statistical randomness tests. The test results are presented in Table 4.

Finally, in order to measure the power of the tests, we produce biased nonrandom data and observe which statistical randomness tests detect the bias. Using a random source, we produce sequences of length 2^{24} that satisfy $Pr(a_i = 1) = \frac{1}{2} + q$, for each i, where q is the bias. We then find that for which values of q the tests detect the nonrandom behavior of the generator. The results are presented in Table 5. We observe that three instances of our new randomness test can detect the nonrandom behavior of the generator even for $q = 0.001$, where the template matching tests in NIST cannot detect that bias.

Table 4. Test results for the four PRNGs.

	MerTwis	c# random	RNGCrypto	AES
Frequency	0.597616	0.906325	0.251393	0.156628
Block Freq	0.121173	0.958137	0.845587	0.482953
Runs	0.110322	0.686351	0.374532	0.348251
Long Run of Ones	0.191338	0.097943	0.213666	0.698123
Bin Matrix Rank	0.356745	0.269794	0.329875	0.698390
Nonover Temp	0.055189	0.804045	0.089171	0.276687
Over Temp	0.275223	0.024611	0.328769	0.249247
Maurer Univ	0.044998	0.854632	0.364266	0.423608
Linear Comp	0.693782	0.364427	0.105382	0.956454
Serial Test 1	0.147844	0.557261	0.126045	0.157290
Serial Test 2	0.382965	0.279386	0.129355	0.058659
App Entropy	0.252337	0.823391	0.037039	0.640333
CuSum Forw	0.715825	0.965085	0.062243	0.166657
CuSum Back	0.302646	0.906039	0.449610	0.089899
Rand Excur	0.589143	0.279374	0.421991	0.152072
Rand Excur Var	0.497518	0.483294	0.980570	0.450652
0-bit Temp (0001)	0.250408	0.762958	0.369484	0.804434
1-bit Temp (0010)	0.724940	0.961562	0.764521	0.416899
2-bit Temp (0101)	0.330876	0.418822	0.809574	0.241198
3-bit Temp (1111)	0.930115	0.194722	0.380861	0.865790

Table 5. Test results for the biased nonrandom data.

	0	0.001	0.002	0.003	0.004	0.005	0.01	0.02	0.03	0.04
Frequency	Ran	**Non**	Non	Non	Non	Non	Non	Non	Non	Non
Block Freq	Ran	Ran	Ran	Ran	Ran	**Non**	Non	Non	Non	Non
Runs	Ran	**Non**	Non	Non	Non	Non	Non	Non	Non	Non
Long Run of Ones	Ran	Ran	Ran	Ran	**Non**	Non	Non	Non	Non	Non
Bin Matrix Rank	Ran	Ran	Ran	Ran	Ran	Ran	Ran	Ran	Ran	Ran
Nonover Temp	Ran	Ran	**Non**	Non	Non	Non	Non	Non	Non	Non
Over Temp	Ran	Ran	**Non**	Non	Non	Non	Non	Non	Non	Non
Maurer Univ	Ran	Ran	Ran	Ran	Ran	Ran	**Non**	Non	Non	Non
Linear Comp	Ran	Ran	Ran	Ran	Ran	Ran	Ran	Ran	Ran	Ran
Serial Test 1	Ran	Ran	Ran	Ran	**Non**	Non	Non	Non	Non	Non
Serial Test 2	Ran	Ran	Ran	Ran	Ran	Ran	Ran	Ran	Ran	**Non**
App Entropy	Ran	Ran	**Non**	Non	Non	Non	Non	Non	Non	Non
CuSum Forw	Ran	**Non**	Non	Non	Non	Non	Non	Non	Non	Non
CuSum Back	Ran	**Non**	Non	Non	Non	Non	Non	Non	Non	Non
Ran Excur	Ran	**Non**	Non	Non	Non	Non	Non	Non	Non	Non
Ran Excur Var	Ran	**Non**	Non	Non	Non	Non	Non	Non	Non	Non
0-bit Temp (0001)	Ran	**Non**	Non	Non	Non	Non	Non	Non	Non	Non
1-bit Temp (0010)	Ran	**Non**	Non	Non	Non	Non	Non	Non	Non	Non
2-bit Temp (0101)	Ran	Ran	Ran	Ran	Ran	Ran	Ran	Ran	**Non**	Non
3-bit Temp (1111)	Ran	**Non**	Non	Non	Non	Non	Non	Non	Non	Non

6. Conclusion

Random sequences are used widely in cryptographic applications and it is vital to use a proper random number generator to produce keys. Randomness testing is done by statistical randomness tests, and the NIST test suite is the most popular suite for cryptographic applications. One of the randomness tests in this suite is the overlapping template matching test.

In this work, we classify all templates according to their number of overlapping bits and show that the probabilities used in NIST's overlapping template matching test are valid only for $B = 111111111$ and should be recalculated for different overlapping blocks. Moreover, we find the exact distributions for all 4-bit templates and propose new randomness tests, namely 4-bit template matching tests.

We apply the proposed tests to biased random data and observe that the new tests detect the nonrandom behavior of the generator even for $q = 0.001$, where the template matching tests in NIST cannot detect that bias. Moreover, NIST's overlapping template matching test can only be applied to long sequences, i.e. sequences of minimum length 10^6, whereas the new proposed tests can be applied to any sequence whose length is greater than 5504. Furthermore, for the new tests, it is also possible to change the subsequence length by calculating the bin probabilities for the new subsequence length.

As a future work, exact distributions can be obtained for all 5-bit templates. The probabilities for the other overlapping templates in NIST's overlapping template matching test can also be calculated.

References

[1] Alcover PM, Guillamón A, Ruiz MC. New randomness test for bit sequences. Informatica 2013; 24: 339-356.

[2] Charalambides AC. Enumerative Combinatorics. London, UK: CRC Press, 2002.

[3] Daeman J, Rijmen V. The Design of Rijndael: AES - The Advanced Encryption Standard. Berlin, Germany:Springer-Verlag, 2002.

[4] Doğanaksoy A, Çalık Ç, Sulak F, Turan MS. New randomness tests using random walk. In: 2nd National Conference Proceedings; 15–17 December 2006; Ankara, Turkey.

[5] Doğanaksoy A, Gologlu F. On Lempel-Ziv complexity of sequences. In: Guang G, editor. Sequences and Their Applications - SETA 2006 4th International Conference Proceedings; 24–28 September 2006; Beijing, China. Berlin, Germany: Springer-Verlag, 2006, pp. 180-189.

[6] Doğanaksoy A, Sulak F, Uğuz M, Şeker O, Akcengiz Z. New statistical randomness tests based on length of runs. Mathematical Problems in Engineering 2015; 2015: 626408.

[7] Doğanaksoy A, Tezcan C. An alternative approach to Maurer's universal statistical test. In: Information Security and Cryptology Conference ISC Turkey 2006 3rd International Conference Proceedings; 25–27 December 2008; Ankara, Turkey. pp. 187-189.

[8] Goldberg I, Wagner D. Randomness and the Netscape browser. Dr Dobb's Journal-Software Tools for the Professional Programmer 1996; 21: 66-70.

[9] Hamano K, Kaneko T. Correction of overlapping template matching test included in NIST randomness test suite. IEICE T Fund Electr 2007; 90: 1788-1792.

[10] Hamano K, Sato F, Yamamoto H. A new randomness test based on linear complexity profile. IEICE T Fund Electr 2009; 92: 166-172.

[11] Hamano K, Yamamoto H. A randomness test based on t-codes. In: International Symposium on Information Theory and Its Applications - ISITA 2008; 7–10 December 2008; Auckland, New Zealand. New York, NY, USA: IEEE, 2008, pp. 1-6.

[12] Hamano K, Yamamoto H. A randomness test based on t-complexity. IEICE T Fund Electr 2010; 93: 1346-1354.

[13] Katos V. A randomness test for block ciphers. Appl Math Comput 2005; 162: 29-35.

[14] Knuth DE. The Art of Computer Programming, Volume 2. Seminumerical Algorithms. 3rd ed. Boston, MA, USA: Addison-Wesley Longman Publishing, 1997.

[15] L'Ecuyer P, Simard R. Testu01: A C library for empirical testing of random number generators. ACM T Math Software 2007; 33: 22.

[16] Matsumoto M, Nishimura T. Mersenne Twister: a 623-dimensionally equidistributed uniform pseudo-random number generator. ACM T Model Comput S 1998; 8: 3-30.

[17] Maurer U. A universal statistical test for random bit generators. J Cryptol 1992; 5: 89-105.

[18] Menezes AJ, Vanstone SA, Van Oorschot PC. Handbook of Applied Cryptography. Boca Raton, FL, USA: CRC Press, 1996.

[19] Okutomi H, Kaneda M, Yamaguchi K, Nakamuro K. A study on the randomness evaluation method using NIST randomness test. In: Symposium on Cryptography and Information Security - International Conference Proceedings; 2006.

[20] Ruhkin A. Testing randomness: a suite of statistical procedures. Theor Probab Appl+ 2001; 45: 111-132.

[21] Rukhin AL, Soto J, Nechvatal J, Smid M, Barker E, Leigh S, Levenson M, Vangel M, Banks D, Heckert A et al. A Statistical Test Suite for Random and Pseudorandom Number Generators for Cryptographic Applications Sp 800-22 Rev. 1a. Gaithersburg, MD, USA: Booz-Allen and Hamilton, 2010.

[22] Soto J, Bassham L. Randomness Testing of the Advanced Encryption Standard Finalist Candidates. Gaithersburg, MD, USA: Booz-Allen and Hamilton, 1999.

[23] Sulak F. A new statistical randomness test: saturation point test. International Journal of Information Security Science 2013; 2: 81-85.

[24] Sulak F, Doğanaksoy A, Ege B, Koçak O. Evaluation of randomness test results for short sequences. In: Carlet C, Pott A, editors. Sequences and Their Applications - SETA10 6th International Conference Proceedings; 13–17 September 2010; Paris, France. Berlin, Germany: Springer-Verlag, 2010, pp. 309-319.

[25] Takeda Y, Huzii M, Kamakura T, Watanabe N, Sugiyama T. The Problem of Template Matching Test in the Testing Randomness by NIST. IEICE Technical Report. Tokyo, Japan: IEICE, 2005.