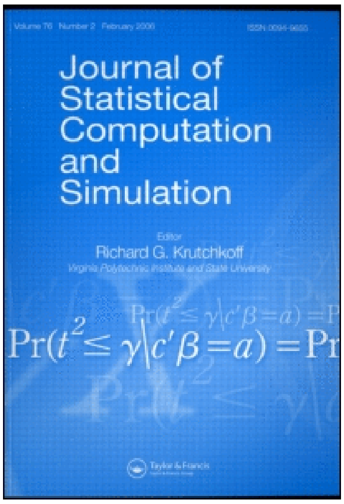

Journal of Statistical Computation and Simulation
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713650378

Testing randomness via aperiodic words

Andrew L. Rukhin [a,b]; Zeev Volkovich [c]
[a] Department of Mathematics and Statistics, UMBC, Baltimore, MD, USA [b] Statistical Engineering Division, National Institute of Standards and Technology, Gaithersburg, MD, USA [c] Software Engineering Department, ORT Braude College, Karmiel, Israel

Online Publication Date: 01 January 2008

To cite this Article Rukhin, Andrew L. and Volkovich, Zeev(2008)'Testing randomness via aperiodic words',Journal of Statistical Computation and Simulation,78:12,1131 — 1142

To link to this Article: DOI: 10.1080/10629360600864142

URL: http://dx.doi.org/10.1080/10629360600864142

Journal of Statistical Computation and Simulation
Vol. 78, No. 12, December 2008, 1131–1142

Taylor & Francis
Taylor & Francis Group

Testing randomness via aperiodic words

ANDREW L. RUKHIN*†‡ and ZEEV VOLKOVICH§

†Department of Mathematics and Statistics, UMBC, 1000 Hilltop Circle, Baltimore, MD 21250, USA
‡Statistical Engineering Division, National Institute of Standards and Technology, Building 820, Gaithersburg, MD 20899-0001, USA
§Software Engineering Department, ORT Braude College 78, Karmiel 21982, Israel

(Received 29 January 2004; final version received 16 June 2006)

The properties of statistical procedures based on occurrences of aperiodic patterns in a random text are summarized. Accurate asymptotic formulas for the expected value of the number of aperiodic words occurring a given number of times and for the covariance matrix are given. The form of the optimal linear test based on these statistics is established. These procedures are applied to testing for the randomness of a string of binary digits originating from block ciphers, US government-approved random number generators or classical transcendental numbers.

Keywords: Asymptotic normality; Block ciphers; χ^2 test; Efficacy; Optimal linear test; Patterns; Random number generators

MSC 2000 Subject Classifications: Primary: 60E05; Secondary: 60F99, 62E20, 62F03

1. Introduction

Consider a random text formed by realizations of letters chosen from a finite alphabet. For a given word (pattern), it is of interest to determine the distribution of the number of (overlapping) occurrences of this pattern in the text. This problem appears in different areas of information theory such as source coding and code synchronization. It is also important in molecular biology in DNA analysis and for gene recognition.

One of the most important applications of this distribution is in testing for randomness of the underlying text. A number of classic tests of randomness are reviewed in ref. [1]. However, some of these tests turn out to be rather weak as they pass patently non-random sequences (see discussion in [2]). Most conventional pseudo-random number generators show patterning because of their deterministic recursive algorithms. Because of this fact, it is natural to employ statistical tests based on the occurrences of words of a given length, say m. The counts of appearances of the patterns in a random text have been used in a battery of statistical tests to assess the quality of different random number generators (RNGs) [3].

*Corresponding author. Email: rukhin@email.nist.gov

Journal of Statistical Computation and Simulation
ISSN 0094-9655 print/ISSN 1563-5163 online © 2008 Taylor & Francis
http://www.tandf.co.uk/journals
DOI: 10.1080/10629360600864142

The tests discussed here utilize the observed frequencies of aperiodic words which appear in a random text a prescribed number of times (*i.e.* which are missing, appear exactly once, exactly twice and so on). In practice, these statistics are easier to evaluate than the entire empirical distribution of occurrences of all m-words. A mathematical advantage of aperiodic words is that a Poisson limit theorem for the number of occurrences of such words holds [4]. Also, the normal approximation discussed in section 3 is more accurate for aperiodic words.

Denote by $Y = (Y_1, \ldots, Y_n)$ a sequence of i.i.d. discrete random variables each taking values in the finite set $\{1, \ldots, q\}$ such that $P(Y_i = k) = p_k, k = 1, \ldots, q$. Thus, the probability of the word $\iota = (i_1 \cdots i_m)$ is $P(\iota) = p_{i_1} \cdots p_{i_m}$. The situation when $p_k \equiv q^{-1}$ corresponds to the randomness hypothesis. The word $\iota = (i_1 \cdots i_m)$ is *aperiodic* if for every k, $1 \le k \le m$, $(i_{m-k+1} \cdots i_m) \neq (i_1 \cdots i_k)$. Thus, when $m = 2$, aperiodic patterns are merely formed by two different letters, but their number grows as $q^m - q^{m-1}$ as m increases (see section 2).

To find words with unexpected frequencies, one can use asymptotically normal estimates of word probabilities or the exact distributions obtained from generating functions (see, for example, refs. [5, 6, section 7.6]). These results suggest that, under the condition of i.i.d. sequence, the probability for a given word ι to appear exactly r times in the string of length n can be approximated by the Poisson probability of the value r, when the Poisson parameter is $nP(\iota)$. Thus, the distribution of the number of words with prescribed r must be approximately equal to that of the sum of Bernoulli random variables whose success probability is this Poisson probability. However, further information about this distribution particularly the covariance structure for several random variables corresponding to different patterns needed in the study of large sample efficiency is not clear.

The approximate Poisson distribution for the number of missing words is alluded to in ref. [7]. It forms the basis of the so-called OPSO test of randomness in the Diehard Battery [8]. Rukhin [9] developed asymptotic formulas for the expected number of words and for the covariance of words with given occurrences. We show here that these formulas are applicable (and, in fact, are more exact) when only aperiodic words are considered.

Section 2 deals with the expected value of the number of words occurring a given number of times and the covariance structure of corresponding random variables. In section 3, asymptotic normality of these variables is stated and the form of the optimal linear test based on such statistics is established. These results are applied to a practical problem of testing block ciphers in section 4. The example of two advanced encryption standard (AES) competitors is examined there along with the results of numerical experiments on *unirnd* Matlab function on files generated with a HG400 RNG and on a physical random bit generator. The National Institute of Standards and Technology (NIST)-recommended RNGs are also discussed. In addition, we study randomness of binary digits in expansions of classical numbers e, π, $\sqrt{2}$ and $\sqrt{3}$ by evaluating the P-values of a test statistic.

2. Asymptotic formulas for the expected number and the covariance of aperiodic words with given occurrences

We will need formulas for the probabilities that a given m-pattern appears a prescribed number of times in a series of length n formed by q-valued independent bits. Assume that both $n \to \infty$ and $q \to \infty$, so that $n/q^m \to \alpha$ with a fixed positive α. To implement this setting in the case of binary alphabet, take non-overlapping substrings formed by zeros and ones of given length p to represent the letters of the new alphabet, so that there are $q = 2^p$ new letters. Then, the number of m-letter patterns (the original substrings of length mp) with a given number of occurrences is evaluated. (In the Diehard test $m = 2$, $p = 10$, $q = 2^{10}$.) Of course, then $n = n'/p$, where n' is the length of the original binary string.

In the study of asymptotic efficiency of tests for randomness, the distribution of the alphabet letters under the alternative hypothesis is commonly supposed to be close to the uniform. Typically, for any letter k, $p_k - q^{-1} \sim q^{-s}$ with $s > 1$. It is known that a judicious choice of s may depend on m. For example, for the efficient test based on the number of missing patterns, $s = 1 + m/4$. Similar conditions are required in the Poisson approximation of the probability that given patterns are missing [10, Chapter 3, section 1]. To determine efficient tests, we assume that

$$p_k = \frac{1}{q} + \frac{\eta_k}{q^{3/2}}, \quad k = 1, \ldots, q, \tag{1}$$

$\sum_{k=1}^{q} \eta_k = 0$, so that as $n \to \infty$ and $q \to \infty$

$$\frac{1}{q} \sum_k \eta_k^2 \longrightarrow \mathbf{B},$$

with uniformly bounded sequences η_k, $k = 1, \ldots, q$. Then, $nP(\iota) \to \alpha$.

Denote by $\pi_\iota^r(n)$ the probability that a word ι appears exactly r times in a string of size n and by $p_r(\alpha) = \alpha^r e^{-\alpha}/r!$, $r = 0, 1, \ldots,$ the Poisson probabilities. According to Rukin [4], for $r = 0, 1, \ldots,$

$$\pi_\iota^r(n) = p_r(\alpha) \left[1 - \frac{(\alpha - r) \sum_k \eta_{i_k}}{q^{1/2}} \right.$$
$$\left. + \frac{((\alpha - r)^2 - r)(\sum_k \eta_{i_k})^2 - 2(\alpha - r) \sum_{1 \le k < j \le m} \eta_{i_k} \eta_{i_j}}{2q} + \mathrm{O}\left(\frac{1}{q^{3/2}} \right) \right]. \tag{2}$$

The form of the probabilities (2) leads to the formula for the expected value of the number of aperiodic m-words, which occur r times in a sequence of i.i.d. random bits of size, n, $X^r = X_n^r$. Indeed, the number L_m of aperiodic words of length m satisfies the recurrent relation,

$$L_m + \sum_{k=\mathrm{o}}^{\lfloor m/2 \rfloor} L_k q^{m-2k} = 2q^m, \quad L_0 = 1,$$

which follows from [11, Theorem 7.1, p 31]. According to this formula, $L_1 = q$, $L_2 = q^2 - q$, $L_3 = q^3 - q^2$, $L_4 = q^4 - q^3 - q^2 + q$. For $m > 4$,

$$L_m = q^m - q^{m-1} + \mathrm{O}(q^{m-2}).$$

As $\Sigma_\iota \Sigma_{k<j} \eta_{i_k} \eta_{i_j} = 0$, one has $\Sigma_\iota (\Sigma_k \eta_{i_k})^2 = \Sigma_\iota \Sigma_k \eta_{i_k}^2 m q^{m-1} \Sigma_{\ell=1}^{q} \eta_k^2 = m q^m \mathbf{B}$. Therefore, with $\pi_\iota^r(n)$ determined from equation (2) for $r = 0, 1 \ldots$

$$\begin{aligned} \mathbf{E} X^r &= L_m \pi_\iota^r(n) \\ &= \frac{\alpha^r \mathrm{e}^{-\alpha}}{r!} q^m \left[1 + \frac{m\mathbf{B}}{2q} (\alpha^2 - (2\alpha + 1) r + r^2) - \frac{1}{q} + \mathrm{O}\left(\frac{1}{q^{3/2}} \right) \right]. \end{aligned} \tag{3}$$

Observe that this formula is different from formula (4.3) for the expected number of all m-words, which occur r times in [9].

The formula for the covariance can be obtained from the fact that $X^r = \Sigma_j x_j^r$, where x_j^r is 0 or 1 according to the occurrence of the word j exactly r times in the string of length n. As

$$\mathbf{E}x_\iota^r x_j^t = \pi_{\iota j}^{rt}(n) = P(\iota \text{ appears } r \text{ times}, \ j \text{ appears } t \text{ times}),$$

one gets

$$\begin{aligned}\mathrm{Var}(X^r) &= \sum_\iota \mathrm{Var}(x_\iota) + \sum_{\iota \neq j} \mathrm{Cov}(x_\iota, x_j) \\ &= \sum_\iota \pi_\iota^r(n)\left[1 - \pi_\iota^r(n)\right] + \sum_{\iota \neq j}\left[\pi_{\iota j}^{rr}(n) - \pi_\iota^r(n)\pi_j^r(n)\right]. \qquad (4)\end{aligned}$$

For $r \neq t$,

$$\mathrm{Cov}(X^r, X^t) = \sum_{\iota = j}\left[\pi_{\iota j}^{rt}(n) - \pi_\iota^r(n)\pi_j^t(n)\right] - \sum_\iota \pi_\iota^r(n)\pi_\iota^t(n). \qquad (5)$$

The probabilities $\pi_\iota^r(n)$ have been determined in equation (2).

The formulas for the probabilities $\pi_{\iota j}^{rt}(n)$ are given in ref. [9],

$$\begin{aligned}\pi_{\iota j}^{rt}(n) - \pi_\iota^r(n)\pi_j^t(n) = &-\frac{\mathrm{e}^{-2n[P(\iota)+P(j)]}[nP(\iota)]^r[nP(j)]^t(\alpha - r)(\alpha - t)}{\alpha r! t!} \\ &\times \left[\frac{2m-1}{q^m} + \mathrm{O}\left(\frac{1}{q^{m+1}}\right)\right]. \qquad (6)\end{aligned}$$

It has been noticed there that the main contribution to the sums in equations (4) and (5) (of order q^m) is due to the pairs of uncorrelated aperiodic words. It follows now from equation (6) that, for $r \neq t$,

$$\begin{aligned}\mathrm{Cov}(X^r, X^t) &= -q^m p_r(\alpha)p_t(\alpha)\left[(2m-1)\left(\alpha - r - t + \frac{rt}{\alpha}\right) - 2(m-1)\right. \\ &\quad \left.\times \frac{(\alpha - r)(\alpha - t)}{\alpha} + 1\right] + \mathrm{O}(q^{m-1}) \\ &= -q^m p_r(\alpha)p_t(\alpha)\left[\frac{(\alpha - r)(\alpha - t)}{\alpha} + 1\right] + \mathrm{O}(q^{m-1}). \qquad (7)\end{aligned}$$

Similarly,

$$\mathrm{Var}(X^r) = -q^m p_r(\alpha)\left[1 - \frac{\mathrm{e}^{-\alpha}\alpha^r}{r!}\left(\frac{(\alpha - r)^2}{\alpha} + 1\right)\right] + \mathrm{O}(q^{m-1}). \qquad (8)$$

We summarize the results of this section.

THEOREM 2.1 *Assume that the q-valued random variables $Y_1, \ldots, Y_n$ are independent with probabilities satisfying equation* (1). *Let, for $n \to \infty$, $n/q^m \to \alpha$ with a fixed positive α. Then, the probability $\pi_\iota^r(n)$ admits the asymptotic representations equation* (2). *If X^r denotes the number of aperiodic words appearing r times in the sequence $Y_1, \ldots, Y_n$, then the expected value $\mathbf{E}X^r$ has the asymptotic expression* (3). *The covariance between the number of such words appearing exactly r and t times, $Cov(X^r, X^t)$, is of the form* (7) *and the variance of the number of aperiodic words appearing exactly r times*, $\mathrm{Var}(X^r)$, *has the form* (8).

Kolchin *et al.* [10, Chapter 3, Theorem 6] gave the formulas for the first two moments of the joint distribution of the words appearing a prescribed number of times when their frequencies are independent, *i.e.* when the occurrences of words are counted in the non-overlapping m-blocks. A rather surprising fact is that the asymptotic behavior of the expected value and of the covariance matrix is the same for overlapping and non-overlapping occurrences. Therefore, the form of the optimal linear test discussed in the next section, which is determined by these characteristics, coincides with that in [10, Chapter V, Theorem 2].

3. Asymptotic normality and the optimal linear test

The theoretical justification for approximate normality of the distribution of X^r when $n \to \infty, n/q^m \sim \alpha$, is provided by a result of Mikhailov [12]. According to Theorem 2.1, $\mathrm{Var}(X^r) \to \infty$, so that the crucial condition in Mikhailov's theorem is satisfied.

For a fixed positive integer R, denote by Σ the covariance matrix of the limiting distribution of the random variables $X^0, X^1, \ldots, X^R$. The elements of matrix Σ have the form

$$\sigma_{rr} = p_r(\alpha)\left[1 - p_r(\alpha)\left(\frac{(\alpha - r^2)}{\alpha} + 1\right)\right], \tag{9}$$

and for $r \neq t$,

$$\sigma_{rt} = -p_r(\alpha)p_t(\alpha)\left[\frac{(\alpha - r)(\alpha - t)}{\alpha} + 1\right]. \tag{10}$$

THEOREM 3.1 *Under conditions of Theorem* 2.1, *the random number of m-letter aperiodic words,* $X^r = X^r_n$, *which appears exactly r times in a string of length n, is asymptotically normal with the asymptotic mean given by equation* (3) *and the variance determined by equation* (8). *The asymptotic joint distribution of the random variables* $X^0, X^1, \ldots, X^R$ *is normal with the covariance matrix* Σ *determined by equations* (9) *and* (10).

Thus, the vector $q^{-m/2}[(X^0, X^1, \ldots, X^R) - \mathbf{E}(X^0, X^1, \ldots, X^R)]$ has approximate multivariate normal distribution with mean 0 and the covariance matrix Σ. We use Theorem 3.1 to derive the optimal test of the null hypothesis H_0: $\eta \equiv 0$ within the class of linear test statistics of the form

$$S = \sum_{r=0}^{R} w_r(X^r - \mathbf{E}X^r).$$

Indeed, this theorem can be used to find the Pitman efficiency of this statistic, as it is asymptotically normal both under the null hypothesis and the alternative H_1: $\mathbf{B} > 0$. The efficacy of the corresponding statistical test is determined by the normalized distance between the means under the null hypothesis and under the alternative, divided by the standard deviation (which is common to the null hypothesis and the alternative),

$$\mathrm{eff}(S) = \frac{|\sum_{r=0}^{R} w_r p_r(\alpha)[(\alpha - r)^2 - r]|}{(\sum_{r,t} \sigma_{rt} w_r w_t)^{1/2}} = \frac{|\mathbf{w}\mathbf{b}^{\mathrm{T}}|}{\sqrt{\mathbf{w}^{\mathrm{T}}\Sigma\mathbf{w}}}.$$

Here, $(R+1)$-dimensional vector $\mathbf{w}$ has coordinates $w_0, \ldots, w_R$ and $\mathbf{b}$ has coordinates $p_r(\alpha)(\alpha^2 - 2\alpha r + r(r-1)) = \alpha^2[p_r(\alpha) - 2p_{r-1}(\alpha) + p_{r-2}(\alpha)]$, $r = 0, 1, \ldots, R$.

Maximization of this ratio gives the formula for the coordinates of $\mathbf{w}$,

$$\mathbf{w}_r = \alpha^2 - 2\alpha r + r(r-1) + (\alpha - r)\theta \frac{\left[(\alpha - R)^2 + \alpha\theta(\alpha - R) + R\right]}{d} + \frac{(\alpha - R + \alpha\theta)\alpha\theta}{d}, \tag{11}$$

where $\theta = p_R(\alpha)\left[\sum_{r=R+1}^{\infty} p_r(\alpha)\right]^{-1}$ and $d = 1 + (R - \alpha + 1)\theta - \alpha\theta^2$, so that

$$\mathbf{b}^{\mathrm{T}}\Sigma^{-1}\mathbf{b} = \mathbf{b}^{\mathrm{T}}\mathbf{w} = 2\alpha^2 \sum_{0}^{R-1} p_r(\alpha) + \alpha p_R(\alpha) \frac{(\alpha - R + \alpha\theta)[(\alpha - R)^2 + \alpha\theta(\alpha - R) + R]}{d}. \tag{12}$$

THEOREM 3.2 *The weights* $\mathbf{w}_r$ *of the optimal linear test statistic*

$$\mathbf{S} = \sum_{r=0}^{R} \mathbf{w}_r (X^r - \mathbf{E}X^r) \tag{13}$$

of H_0: $p_k \equiv 1/q$ *are given by equation* (11) *with the corresponding efficacy determined by equation* (12).

Table 1 gives the value of $\alpha = \alpha^*$ for $R = 0, \ldots, 8$, which maximizes the efficacy and the corresponding optimal weights normalized so that their sum is equal to one.

For moderate values of $R(\leq 100)$, the optimal value α^* admits a remarkably accurate linear approximation $\alpha^* = 3.60 + 1.09R$ (figure 1). However, $\alpha^*/R \to 1$.

To implement this test on the basis of a string of binary bits for a fixed R, choose a positive integer p, such that $n \approx 2^{mp}\alpha^*$, and take all strings of length p formed by zeros and ones to represent the letters of the new alphabet of the size $q = 2^p$. The numbers X^r of aperiodic m-letter patterns (the original non-overlapping consecutive substrings of length pm), which occurred r times are combined with the weights from the table leading to the asymptotically optimal test. Actually, this test is asymptotically optimal not only within the class of linear functions but also in the class of all statistics based on $X^0, \ldots, X^R$.

In particular, the most efficient test based on the number of missing aperiodic words arises when $\alpha^* = 3.594\ldots$, which means that the best relationship between q and n is $n \approx 3.6q^2$. This formula is used in section 4 to determine the size 231 K of the data array when $m = 2$ and $q = 2^8$.

One can also use Theorem 3.1 to compare several, say, M different independent strings. Let $U_i = (X_i^0, X_i^1, \ldots, X_i^R)^{\mathrm{T}}$ denote the $(R+1)$-dimensional vector of frequencies of m-letter

Table 1. The optimal values α^* and weights $\mathbf{w}$ for small R.

R	α^*	$\mathbf{w}$
0	3.59	1
1	4.77	[0.62, 0.38]
2	5.89	[0.47, 0.33, 0.20]
3	6.98	[0.37, 0.29, 0.20, 0.14]
4	8.06	[0.33, 0.25, 0.19, 0.14, 0.09]
5	9.13	[0.29, 0.23, 0.18, 0.14, 0.09, 0.07]
6	10.17	[0.25, 0.21, 0.18, 0.14, 0.09, 0.07, 0.06]
7	11.21	[0.23, 0.19, 0.17, 0.14, 0.09, 0.07, 0.06, 0.05]
8	12.24	[0.21, 0.18, 0.16, 0.14, 0.09, 0.07, 0.06, 0.05, 0.04]

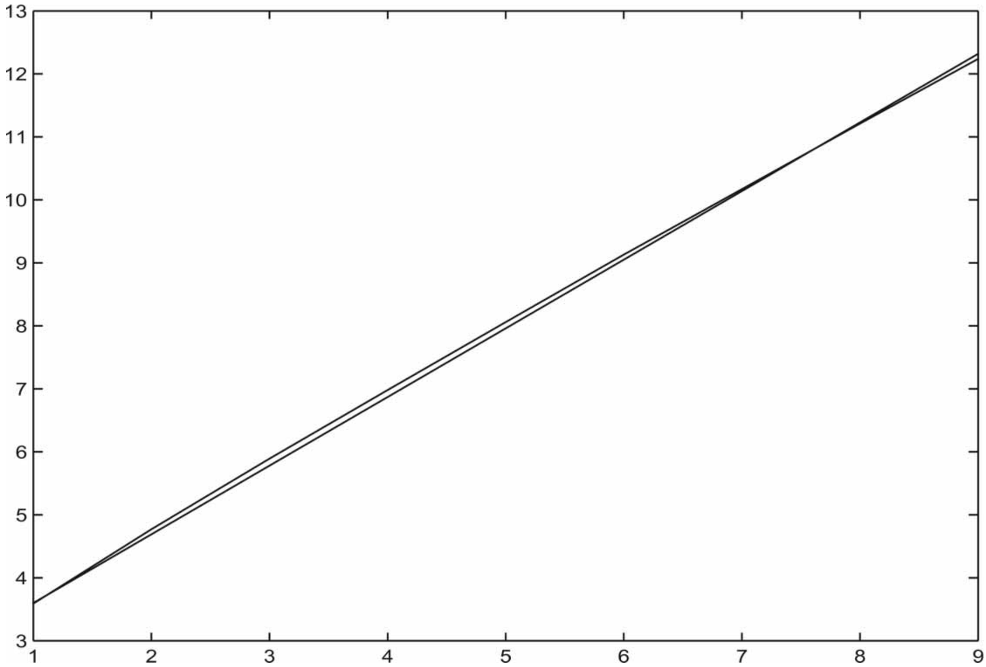

Figure 1. The plot of the optimal value α^* and its linear approximation.

aperiodic words appearing in the ith string, $i = 1, \ldots, M$. Assuming equal sample sizes, a test of the null hypothesis H_0: $\mathbf{E}U_1 = \mathbf{E}U_2 = \cdots = \mathbf{E}U_M$ can be based on the statistic $W = \Sigma_i (U_i - \bar{U})^{\mathrm{T}} \Sigma^{-1} (U_i - \bar{U})$, with Σ defined by equations (9) and (10). Under the null hypothesis, W has approximate χ^2-distribution with $(R + 1)(M - 1)$ degrees of freedom.

4. An example: testing block ciphers and other randomness sources

We start this section with testing of randomness applied to block ciphers. These ciphers are widely used and are important in cryptographic applications. Recently, the NIST carried out a competition for the development of the AES. Its goal was to find a new block cipher which could be used as a standard. Among the requirements was that its bit output sequence should look like a random string even when the input is not random.

Indeed, one of the basic hurdles for the 15 AES candidates was 'Randomness Testing of the AES Candidate Algorithms', whose aim was to evaluate these candidates by their performance as RNGs [13]. It is worth mentioning that some aperiodic words (namely, the templates 010111011, 110001010, $m = 9$, and 01011011, $m = 8$) have been used at earlier stages of the AES testing (in the so-called 128-bit key avalanche set), but in conjunction with the χ^2-statistic (as opposed to the Poisson approximation).

The winner of the competition, the Rijndael algorithm, and a runner-up, the Serpent algorithm, were used in our experiment involving randomness testing of their outputs by using the procedure described earlier with $m = -2$, $R = 5$. Both of these algorithms were implemented in C++MFC on two files of size 231 K each, 1000 times each. Each of the 1000 trials used a different 128-bit randomly chosen key. (In fact, the keys were chosen by self-encrypting the initial key as they passed randomness tests.)

Two modes of encryption were used. In the Electronic Code Book (ECB) mode, the input data were divided into equal size 128-bit blocks, and each block was encrypted one at a time. (Separate encryptions within different blocks are independent of other.) ECB is the weakest mode because no additional security measures are implemented besides the basic algorithm.

In the cipher block chaining (CBC) mode, the plaintext is also divided into equal size 128-bit blocks, but each encrypted block is xored with the next data block. This procedure makes each block dependent on previous blocks. Thus, to find the plaintext of a particular block, one needs to know the ciphertext, the key and the ciphertext of the previous block.

The first (non-random) text was a regular English text which happens to contain only 1116 different aperiodic pairs out of possible 65, 280. The expected numbers of frequencies of aperiodic words under the randomness hypothesis are as follows.

0	1	2	3	4	5
1791	6446	11604	13924	12532	9023

Encryption in the ECB mode by Rijndael after one round did not make it look much more random (table 2). Even after eight rounds, the numbers of aperiodic two-letter patterns were very far from those corresponding to the randomness hypothesis (table 3). (The P-values in the following tables are obtained from the normal approximation in Theorem 3.1 for a two-sided alternative.)

Just one round encryption in the CBC mode led to statistics confirming the randomness hypothesis. Indeed, the value of statistic $S/\sqrt{\mathbf{w}^{\mathrm{T}}\Sigma\mathbf{w}}$ is -0.09 with a large P-value.

The results for the Serpent algorithm turned out to be very similar, although randomness-like statistics were not attained after the first iteration in the CBC mode. Note that the Rijndael algorithm uses the key size and the block size to be 128, 192 or 256 bits and has a variable number of rounds. This number is 10 if both the block and the key are 128-bit long, it is 12 if the maximal length of the block or the key is 192 and it is 14 otherwise. There is an initial round key addition followed by these rounds. The Serpent algorithm encrypts a 128-bit plaintext into the 128-bit ciphertext in 32 rounds. Thus, Rijndael seems to achieve randomness faster, although the 'complexity' of the rounds plays a role too. Still the statistical characteristics of both algorithms did not change much after two rounds.

The second text was a file of zeros. As the ECB mode cannot be expected to lead to good results, we did not try it and give here the results only for the CBC encryption with two rounds.

One can see that the entries in tables 4–6 are very close to the theoretical values. Although all individual P-values (for a two-sided alternative) in these tables are fairly large, the values of

Table 2. Characteristics of the number of aperiodic words under encryption in the ECB mode by Rijndael after one round.

	0	1	2	3	4	5
Mean	51844	2381	1995	996	766	614
Standard deviation	632	161	126	54	42	32
P-value	0.00	0.00	0.00	0.00	0.00	0.00

Table 3. Characteristics of the number of aperiodic words under encryption in the ECB mode by Rijndael after eight rounds.

	0	1	2	3	4	5
Mean	5266	8265	11279	11058	9683	7319
Standard deviation	62	77	93	93	87	77
P-value	0.00	0.00	0.00	0.00	0.00	0.00

Table 4. Characteristics of the number of aperiodic words under encryption in the CBC mode by Rijndael after one round.

	0	1	2	3	4	5
Mean	1789	6443	11606	13929	12529	9023
Standard deviation	39	69	91	104	101	86
P-value	0.58	0.40	0.86	0.98	0.50	0.23

Table 5. Characteristics of the number of aperiodic words under encryption in the CBC mode by Rijndael after two rounds.

	0	1	2	3	4	5
Mean	1795	6457	11,609	13,928	12,523	9018
Standard deviation	39	69	87	101	99	84
P-value	0.79	0.75	0.47	0.89	0.87	0.60

Table 6. Characteristics of the number of aperiodic words under encryption in the CBC mode by Serpent after two rounds.

	0	1	2	3	4	5
Mean	1784	6419	11,556	13,873	12,481	8989
Standard deviation	41	67	86	101	97	81
P-value	0.46	0.93	0.29	0.74	0.86	0.90

statistic $\mathbf{S}/\sqrt{\mathbf{w}^{\mathrm{T}}\Sigma\mathbf{w}}$ in tables 5 and 6 are quite different: 0.62 and -3.68, respectively. It happens because the Serpent algorithm seems to produce fewer aperiodic words than randomness dictates, and this again gives an edge to Rijndael.

The results seem to confirm not only other methods that determined the AES competition winner, but also good qualities of our testing procedure, which is fairly easy to implement.

We also performed numerical experiments on several available generators. A random source of binary strings of length p can be obtained from a RNG which produces integer random numbers in the interval $[0, 2^p - 1]$.

As the first example built in the MATLAB system, RNG *unidrnd* function was tested. One hundred sequences having size of 231 kB were created by integer random numbers in [0, 255] generated by the function. The outcomes of 10 sequences are presented in table 7.

In the next example, we took 10 random files generated with an HG400 RNG–HG432 (at speed of 32 Mbit/s) whose files of size of 1024 kB are available at http://www.random.com.hr/products/hg400/data/.

The inner working of HG432 is described by Stipcevic [14]. The testing was performed with the same value of q. The cases for which the null hypothesis would be rejected at the significance level 0.05 are boldfaced (table 8).

Work is on the way at the ANSI X9F1 standards committee to develop and standardize a RNG that would use certain properties of the physical processes, such as the rates of the radioactive decay, to produce random numbers. The techniques described in this article could be useful in evaluating the properties of such generators. Indeed, one physical random bit generator is given by Jakobsson *et al.* [15], with the supporting data set in a form of binary file of the 92 million random bits (11,468,800 bytes) available at http://www.cs.nyu.edu/symbo1126. This file was divided into successive subfiles of the size 231 kB, which were analyzed with $q = 2^9$ and $q = 2^{10}$, with $3.6q^2$ leading bits of the file (table 9).

Table 7. Outcomes of the MATLAB RNG testing.

	0	1	2	3	4	5
Theoretical values	1790.7	6446.5	11604	13924	12532	9023
First sequence	1749	6495	11482	13994	12463	8961
P-values	0.1623	0.7272	0.1294	0.7224	0.2690	0.2570
Second sequence	1808	6406	11467	13918	12425	9042
P-values	0.6588	0.3071	0.1023	0.4784	0.1697	0.5793
Third sequence	1803	6455	11567	13922	12339	8952
P-values	0.6145	0.5423	0.3668	0.4919	0.0424	0.2274
Fourth sequence	1806	6498	11427	13871	12527	9032
P-values	0.6413	0.7395	0.05051	0.3255	0.4824	0.5377
Fifth sequence	1756	6486	11537	13953	12432	8976
P-values	0.2062	0.6888	0.2680	0.5958	0.1860	0.3104
Sixth sequence	1873	6271	11583	13973	12422	8966
P-values	0.9741	0.0144	0.4240	0.6598	0.1630	0.2742
Seventh sequence	1767	6421	11415	13891	12693	9045
P-values	0.2878	0.3755	0.03994	0.3886	0.9249	0.5916
Eighth sequence	1743	6357	11711	13793	12603	9009
P-values	0.1299	0.1326	0.8405	0.1328	0.7372	0.4414
Ninth sequence	1809	6521	11578	13646	12476	9080
P-values	0.6674	0.8234	0.4059	0.00919	0.3086	0.7258
Tenth sequence	1785	6330	11686	13881	12491	8948
P-values	0.4465	0.07344	0.7777	0.3566	0.3573	0.2149

Table 8. Results of the testing of the HG432 generator.

	0	1	2	3	4	5
Theoretical values	0.0107	0.1677	1.3099	6.8224	26.65	83.281
First file	0	0	0	7	32	75
P-values	0.4588	0.3411	0.1262	0.5271	0.8500	0.1821
Second file	0	0	0	7	25	91
P-values	0.4588	0.3411	0.1262	0.5271	0.3746	0.8012
Third file	0	1	0	6	21	84
P-values	0.4588	0.9790	0.1262	0.3764	0.1369	0.5314
Fourth file	1	0	1	4	22	87
P-values	1	0.3411	0.3933	0.1400	0.1839	0.6582
Fifth file	0	0	2	8	19	61
P-values	0.4588	0.3411	0.7267	0.6740	0.0692	**0.00731**
Sixth file	0	0	2	7	22	75
P-values	0.4588	0.3411	0.7267	0.5271	0.1839	0.1821
Seventh file	0	0	1	9	18	81
P-values	0.4588	0.3411	0.3933	0.7978	0.0469	0.4013
Eighth file	1	0	2	9	26	84
P-values	1	0.3411	0.7268	0.7978	0.4499	0.5314
Ninth file	0	0	4	6	26	72
P-values	0.4588	0.3411	0.9906	0.3764	0.4499	0.1082
Tenth file	0	0	1	7	29	69
P-values	0.4588	0.3411	0.3933	0.5271	0.6755	0.0588

Table 9. Results of a physical random bit generator testing ($q = 2^{10}$).

	0	1	2	3	4	5
Theoretical values	28651	103144	185658	222790	200511	144368
Data	28463	102739	185574	223542	199841	144246
P-values	0.1333	0.1038	0.4222	0.9444	0.0672	0.3741

The US government requires that all cryptographic modules used by the US Federal Agencies to protect sensitive data get validated to the FIPS 140-2 standard. This standard currently allows three RNGs; complying with at least one of them is mandatory. The results of this article can be used by standards developers to demonstrate the real strength of three currently adopted RNGs whose technical description can be found in Annex C of Federal Security standard FIPS 140-2 [16–18]. These three generators have passed the test based on equation (13) with similar P-values as in table 7.

To further study the aperiodic words test properties, the P-values of test statistics based on their frequencies for binary expansions of e, π, $\sqrt{2}$ and $\sqrt{3}$ were evaluated.

As consecutive P-values were sought, it was more convenient to employ a χ^2-statistic based on the pseudo-inverse of the limiting covariance matrix of the joint distribution of aperiodic word frequencies. In figure 2, the P-values are plotted against the first 50,000 digits of binary expansions of $\sqrt{2}$, $\sqrt{3}$, π and e. According to this data, P-values corresponding to $\sqrt{3}$ and e are somewhat smaller than those of $\sqrt{2}$ and π. The smallest P-values for $\sqrt{3}$ binary expansion occur in the block from 3447th to 3453th digits, (of order 0.03). Because of the multiple nature of the testing problem, they lack statistical significance to reject the random nature of these digits. Our results do not support the conjecture about the non-random appearance of digits in the expansion of $\sqrt{3}$ [19]. Notice that Good and Gover [20] applied the serial test to the study of binary digits in the expansion of $\sqrt{2}$, and Rukhin [3] employed the approximate entropy test. Similar to $\sqrt{3}$, these tests occasionally led to small P-values (about 0.0025), which, however, do not provide enough statistical significance against the randomness hypothesis.

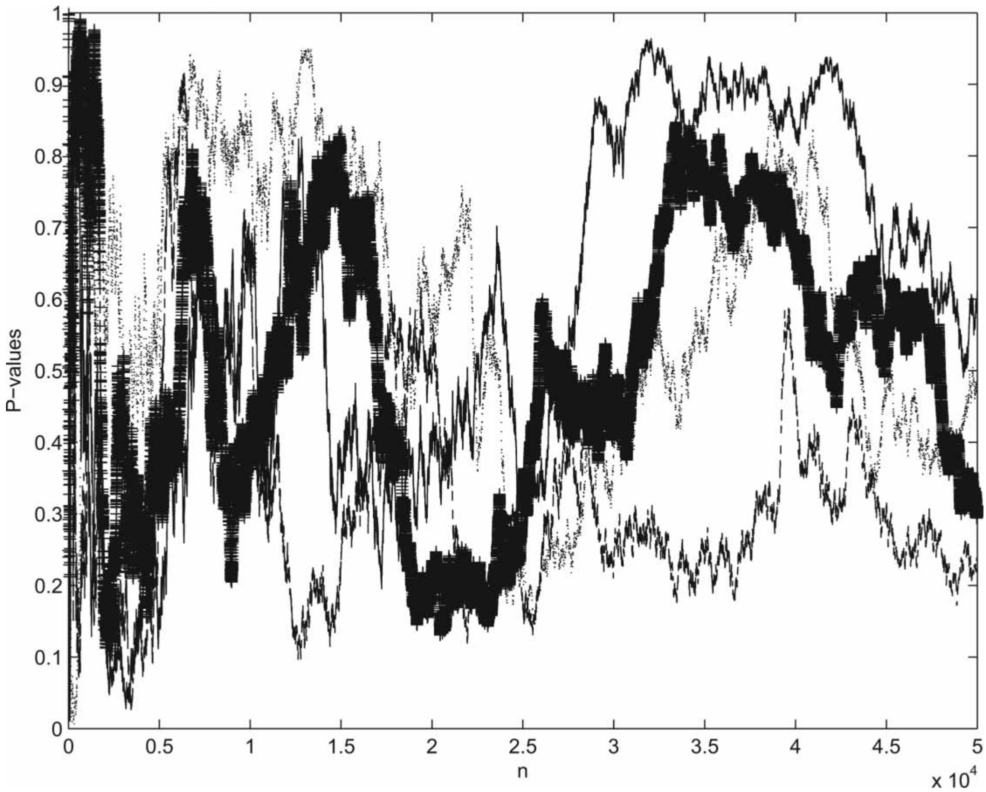

Figure 2. Consecutive P-values for binary expansions of $\sqrt{2}$ (the line marked by $+$), $\sqrt{3}$ (dashed line), π (dotted line) and e (solid line) when $m = 3$.

To sum up, the aperiodic words test could be a useful addition to the existing suite of tests for randomness [21].

Acknowledgements

Z. Volkovich is also affiliated with the Department of Mathematics and Statistics, University of Maryland at Baltimore County. A. L. Rukhin's research was supported by a grant no. MSPF-02G-068 from the National Security Agency. The authors are grateful to the referee for his helpful comments and to J. Soto and A. Roginsky for their interesting discussion.

References

[1] Knuth, D.E., 1997, *The Art of Computer Programming*, Vol. 2 (3rd edn) (Reading, MA: Addison-Wesley Inc.).
[2] Marsaglia, G., 1985, A current view of random number generation. *Computer Science and Statistics: Proceedings of the Sixteenth Symposium on the Interface* (New York: Elsevier Science Publishers), pp. 3–10.
[3] Rukhin, A.L., 2000, Approximate entropy for testing randomness. *Journal of Applied Probability*, **37**, 88–100.
[4] Barbour, A.D., Holst, L. and Janson, S., 1992, *Poisson Approximation* (Oxford: Oxford University Press).
[5] Reinert, G., Schbath, S. and Waterman, M.S., 2000, Probabilistic and statistical properties of words: an overview. *Journal of Computational Biology*, **7**, 1–46.
[6] Szpankowski, W., 2001, *Average Case Analysis of Algorithms on Sequences* (New York: Wiley-Interscience).
[7] Marsaglia, G. and Zaman, A., 1993, Monkey tests for random number generators. *Computers and Mathematics with Applications*, **9**, 1–10.
[8] Marsaglia, G., 1996, Diehard: a battery of tests for randomness. Available online at: http://stat.fsu.edu/geo/diehard.html
[9] Rukhin, A.L., 2002, Distribution of the number of words with a prescribed frequency and tests of randomness. *Advanced Applied Probability*, **34**, 775–797.
[10] Kolchin, V.F., Sevast'yanov, B.A., and Chistyakov, V.P., 1978, *Random Allocations* (Washington, DC: Whinston Sons).
[11] Guibas, L.J. and Odlyzko, A.M., 1981, Periods in strings. *Journal of Combinatorial Theory*, **30A**, 19–42.
[12] Mikhailov, V.G., 1989, Asymptotic normality of decomposable statistics from the frequencies of m-chains. *Discrete Mathematics and Applications*, **1**, 335–347.
[13] Soto, J. and Bassham, L., 2001, Randomness testing of the advanced encryption standard finalist candidates. Proceedings of AES Conference. Available online at: http://csrc.nist.gov/publications/nistir/ir6483.pdf
[14] Stipcevic, M., 2004, Fast nondeterministic random bit generator based on weakly correlated physical events. *Review of Scientific Instruments*, **75**, 4442–4449.
[15] Jakobsson, M., Shriver, E., Hillyer, B.K., and Juels, A., 1998, A practical secure physical random bit generator. Proceedings of the Fifth ACM Conference on Computer and Communications Security, San Francisco.
[16] American Bankers Association, 1998, Public key cryptography for the financial services industry: the elliptic curve digital signature algorithm (ECDSA), Annex A.4. ANSI X9.62.
[17] National Institute of Standards and Technology, 2000, Digital signature standard (DSS), Appendices 3.1 and 3.2, Federal Information Processing Standards Publication 186-2. Available online at: http://csrc.nist.gov/publications/fips/fips186-2/fips186-2-changel.pdf
[18] National Institute of Standards and Technology, 2005, Using the 3-key triple DES and AES algorithms, NIST-Recommended Random Number Generator Based on ANSI X9.31, Appendix A.2.4. Available online at: http://csrc.nist.gov/cryptval/rng/931rngext.pdf
[19] Pincus, S. and Kalman, R.E., 1997, Not all (possibly) 'random' sequences are created equal. *Proceedings of the National Academy of Sciences of the United States of America*, **94**, 3513–3518.
[20] Good, I.J. and Gover, T.N., 1967, The generalized serial test and the binary expansion of $\sqrt{2}$. *Journal of Royal Statistical Society*, **130A**, 102–107.
[21] Rukhin, A.L., Soto, J., Nechvatal, J., Smid, M., Barker, E., Leigh, S., Levenson, M., Vangel, M., Banks, D., Heckert, A., Dray, J. and Vo, S., 2000, A statistical test suite for random and pseudorandom number generators for cryptographic applications. *NIST Special Publication 800-22*, Department of Commerce. Available online at: http://csrc.nist.gov/rng/