Hindawi
Advances in Fuzzy Systems
Volume 2018, Article ID 4028417, 10 pages
https://doi.org/10.1155/2018/4028417

Research Article

Search of Fuzzy Periods in the Works of Poetry of Different Authors

Artur Nor [1] **and Eugene Korotkov** [1,2]

[1] *National Research Nuclear University "MEPhI", Kashirskoe Highway, 31, 115409, Moscow, Russia*
[2] *Institute of Bioengineering, Research Center of Biotechnology of the Russian Academy of Sciences, Leninsky Ave. 33, bld. 2, 119071, Moscow, Russia*

Correspondence should be addressed to Eugene Korotkov; genekorotkov@gmail.com

Received 4 April 2018; Revised 2 July 2018; Accepted 16 July 2018; Published 16 August 2018

Academic Editor: Ferdinando DiMartino

We applied a new method for the identification of fuzzy periods and the insertion and deletion of characters were taken into consideration while studying the works of poetry. The technique employs genetic algorithm, dynamic programming, and the Monte Carlo method. In the present work, the technique was applied to poems written by the famous Russian and foreign classics. A total of 95 poems were studied; and fuzzy periods possessing high statistical significance were identified with more than half of the poems under study. The existence of correlation between the stressed vowel letters in a poem with the position of the fuzzy periods was shown. The present study shows that a work of poetry contains both semantic component and fuzzy periods of letters; hence a poem could have psychological impact on the audience.

1. Introduction

Works of poetry could be considered as a superposition of the semantic content and of the acoustic wave determined by a certain sound alternation periodicity. In relation to this, a poet is capable of combining the semantic content with a certain acoustic wave in a work of poetry. A certain periodicity of sounds alternation in a work of poetry is understood as an acoustic wave. If the meaning of a poetic text is easily understood by each person, the acoustic wave embedded in a work of poetry will be perceived rather intuitively, as some musicality, often fascinating the listeners and exposing them to a certain psychological impact [1]. In order to understand the mechanism of the acoustic wave impact on listeners, it would be very interesting to attempt quantitatively identifying and studying the acoustic wave embedded in a work of poetry, in the form of a certain periodicity of the poetic text [2, 3]. To solve this problem, it seems important to develop and apply new mathematical methods that could quantitatively demonstrate the existence of an acoustic wave in a work of poetry in the form of fuzzy periods and provide the quantitative characteristics of the periodicity found. This task seems to be important, since the quantitative determination of acoustic waves would ensure the classification of existing acoustic waves in the works of poetry. Thus, we could correlate a certain type of acoustic wave and its impact on a listener. After introducing such an important concept as fuzzy periods [4, 5], we could illustrate it with an example. Under the fuzzy periods, we shall obtain the mean of such periods, where the similarity between individual periods is insignificant or is missing at all; and the periodicity becomes statistically significant only on a certain set of periods (more than 2) [6]. Fuzzy periods could be demonstrated with an example. Let us consider a sequence in the following form:

(qzwrt)(qzwrt)(qzwrt)(qzwrt)(qzwrt)(qzwrt)(qzwrt)...

The given sequence is characterized by a perfect periodicity consisting of 5 letters. In this study, each period is highlighted in parentheses, for clarity. There is absolute similarity between the separate periods and it is easily identified using the techniques described previously. Considering a case in the position of each period, a definite and limited set of alphabet letters could be found; for example, such set of letters for each

period position is shown as follows: {q,i,u,s,t}; {u,c,i,a,s,r}; {o,p,f,g,l,k,w}; {a,b,n,m,v}; {p,f,g,h,t,j,r}.

Now, let us create a character sequence taking from each set a letter with the use of random technique and corresponding to the period position; then, the sequence can be obtained in the following form:

(iroap)(tufng)(sslmt)(uawaj)(qcgbf)(siknh)(sipvr) ...

The resulting character sequence lacks absolute periodicity. However, it should be noted that given the sufficient length of this sequence, it could be seen that, in the position of each period, only certain alphabet letters are located. Such a sequence is characterized by fuzzy periods, which could not be identified by pairwise comparison of any two periods but could be detected using a certain set of periods (more than 2).

Nowadays, several mathematical techniques are employed for the detection of fuzzy periods in character and numerical sequences. These include the wavelet transform [7] and the Fourier transform [8]. Previously, the information decomposition (ID) technique was developed [4]. The difference between the ID technique and the Fourier transform lies in the fact that the ID technique could be used for character sequence analysis without recoding it into a numerical series. Such a method of analysis makes it possible to obtain results that are unattainable with the Fourier transform. This allowed the fuzzy periods in DNA sequences [5], amino acid sequences [6], and of several works of poetry to be revealed [4]. However, the ID technique, like other methods previously discussed, does not allow the finding of a statistically significant fuzzy period with insertions and deletions of characters, which in case of literary works could be registered in connection with pronunciation peculiarities. For example, certain sounds may not be pronounced at all or may be pronounced with a certain accent. Consequently, most of the fuzzy periods contained in the sequence could not be determined using the previously developed methods.

As of today, there are mathematical approaches based on dynamic programming that allow the accurate identification of fuzzy periods of time series or character sequence in the presence of characters insertion or deletion [9, 10]. All these techniques are used to construct the multiple alignment of periods; and they are based either on performing the pairwise alignment of periods, followed by the subsequent creation of a guide tree, or on the search for embryos or common words in periods. Thereafter, the initial multiple alignment of periods is provided; and the optimization thereof is carried out in one way or another, including the use of hidden Markov models, iterative procedures, and some other techniques [10–12]. However, all the developed approaches do not ensure construction of the multiple alignment, if the statistically significant pair alignment is missing in the analyzed sequences. It does not allow the creation of a statistically significant guide tree for the progressive alignment; or the sequences are that different that they do not provide searching for the statistically significant embryos or common words. It turns out that nowadays, it is impossible to construct a multiple alignment for significantly different sequences (periods). In this case, it could be argued that all the developed approaches are "blind" and will not identify a statistically significant multiple alignment in the significantly different sequences (periods). Such an alignment could be found, if it would be possible to construct a multiple alignment through the direct application of dynamic programming for all the analyzed sequences. But this is the so-called NP-complete problem [13, 14]; and such an approach requires gigantic computer resources that are not available at present; and it is difficult to think about its creation in the nearest future.

Previously, a new technique was developed for identifying the fuzzy periods in character sequences, which took into consideration the insertions and deletions of characters [15, 16]. This technique is based on the new solution of the NP-complete problem regarding the sequences (periods) multiple alignment. This method employs genetic algorithm, techniques aimed at optimizing weight matrices, dynamic programming, and the Monte Carlo method. It enables identification of the fuzzy periods of a character sequence with insertions and deletions in previously unknown positions. It is important to note that this analysis requires only the symbolic sequence itself (the text of the poetic work) and other information about the poetic work, including the placement of stresses and features of pronunciation, are not required. In the given work, this approach was applied while searching for fuzzy periods in the poems of famous Russian and English-speaking poets. We showed that, in more than half of the works of poetry, it is possible to find fuzzy periods. This study shows that a work of poetry contains both semantic component and fuzzy periods, which could be responsible for the psychological impact of a poem on the audience. Fuzzy periods can be a reflection of the sound "wave" which exists in a poetic work.

2. Fuzzy Periods Search Technique Algorithm Used with Consideration of Characters' Insertions and Deletions

At the beginning of the work, the poetics is transformed in such a way that all the spaces are deleted, uppercase letters are changed to lowercase, and punctuation marks are changed to spaces (Figure 1, Paragraph 1). Thus, the character sequence is created on the basis of the transformed work of poetry for further evaluation. In Figure 1, Paragraph 2, the $Q(n)$ set of random matrices having the $k \times n$ dimension is generated, where n is the period length and k is the size of the original alphabet sequence. In Figure 1, Paragraph 3, modification and optimization of random matrices are performed, which is required for constructing the S sequence alignment.

Then, in Figure 1, Paragraph 4, a search was conducted for a matrix that possesses the greatest value of the similarity function, when the S sequence is aligned. For this purpose, genetic algorithm and dynamic programming are applied. At each phase of the genetic algorithm and for each matrix and the S sequence, we calculate the maximum value of the E_{max} similarity function using dynamic programming. In this case, E_{max} appears to be a fitness function; and each matrix becomes a genotype. Then, to the $Q(n)$ set of matrices we apply the genetic algorithm, which causes the mutation, multiplication, and destruction of matrices. As a result, we

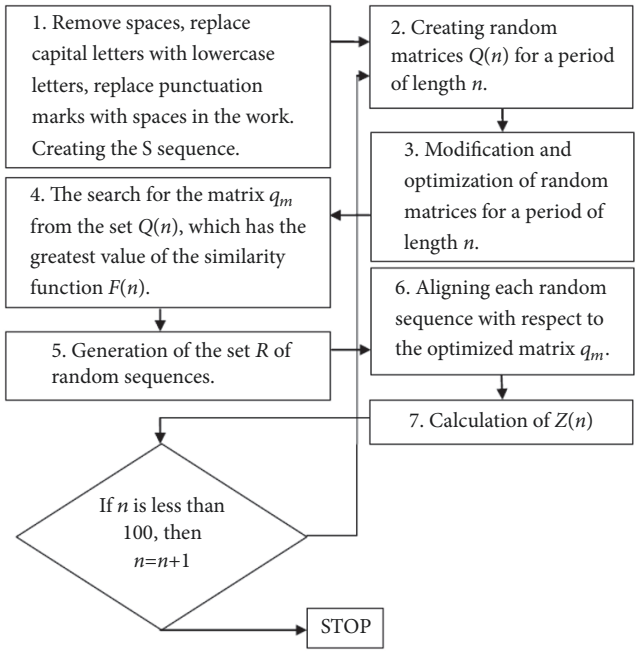

Figure 1: The main stages of the algorithm mused for calculation Z(n) of the analyzed sequence S.

find such M_{max} matrix that possesses the greatest E_{max} value; and we denote it as mE_{max}. In order to estimate the statistical significance of mE_{max}, we generate the R random sequences set (Figure 1, Paragraph 5). This set is generated using S sequence random mixing. Then, for each sequence out of the R set, the maximum value of the E_{max} similarity function for the M_{max} matrix is determined (Figure 1, Paragraph 6). This makes it possible to calculate the average value and variance for the E_{max} with the R set and then calculate $Z(n)$ (Figure 1, Paragraph 7). These calculations were performed for n values from 2 to 100. As a result of the algorithm, the dependence of Z on n was obtained and was denoted as $Z(n)$. Let us consider Paragraphs 1-7 in more details.

3. Main Phases of the Technique and the Algorithms Used

3.1. Spaces Removal, Replacing Uppercase Letters with Lowercase Letters and Replacing Punctuation Marks with Spaces. We submitted a poem, which we would like to study for the presence of fuzzy periods, to the program input. The program replaces all uppercase letters with lowercase letters and deletes all spaces (Figure 1, Paragraph 1). Next, the program replaces punctuation marks such as a dot, comma, dash, colon, interrogative and exclamation marks, and also the end of each line to the space character. The space plays the role of a pause [17]. A space is included in the alphabet as an additional character and, thus, an alphabet of the k characters size is used ($k = 34$ for Russian works and $k = 27$ for works in English). At the program output, a transformed work consisting of only the characters of the given alphabet is obtained. This work is regarded as the S sequence with the N length.

The method of preparing the text can introduce certain distortions into the periodicity, which is in the poetic works. However, an incorrect method can only worsen the statistical importance of periodicity, since the shortcomings in the preparation of the text are compensated by the creation of additional insertions or deletions. This means that, for this text preparation, we may find that not all periods are significant. However, those that are discovered exist in the analyzed poems.

3.2. Creating Random Matrices for the n Length Period. The $k \times n$ dimension random matrices, where k is the size of the alphabet and n is the period length, are generated as follows. Each element of $m(i, j)$,i=1,...,k, j=1,... n matrix was randomly filled with equal probability of either 0 or 1. A total of 10^5 of such matrices are created for each n length period, where n varies consecutively in steps of 1 from 2 to 100. Out of the created random matrices and for each n length period, we selected only 10^3 of such matrices, which in the $k \times n$ dimension space were located at a certain value from each other. For this purpose, the matrix in the $k \times n$ space is considered as a point; and we should take only those points that are located at a distance not less than D_0 from each other. The distance between the points is calculated as

$$D = \sqrt{\sum_{i}^{k} \sum_{j}^{n} \left(m_1(i, j) - m_2(i, j)\right)^2} \quad (1)$$

Here $m_1(i, j)$ is the M_1 random matrix element, and $m_2(i, j)$ is the M_2 random matrix element. The matrix (point in the $k \times n$ space) was added to the $Q(n)$ set, if the D distance between it and every already included matrix (point) in this set was greater than the D_0 value. The first generated random matrix was immediately included in the $Q(n)$ set (Figure 1, Paragraph 2).

3.3. Modification and Optimization of Random Matrices for the n Length Period. Then, a modification of the generated random matrices of the $Q(n)$ set (Figure 1, Paragraph 3) was performed. This was done with the goal of ensuring that the mE_{max} distribution functions from different matrices of the $Q(n)$ set and at the R random sequences set were identical. For this purpose, the algorithm described in [15] was used. To do this, each matrix was modified, so that the R^2 and K_d values would be identical for all the matrices.

$$R^2 = \sum_{i=1}^{k} \sum_{j=1}^{n} m(i, j)^2 \quad (2)$$

$$K_d = \sum_{i=1}^{k} \sum_{j=1}^{n} m(i, j)\, p(i, j) \quad (3)$$

where $m(i, j)$ is the matrix element and $p(i, j) = f(i)t(j)$, while $\sum_{i,j} p(i, j) = 1$, $f(i) = b(i)/N$, where $b(i)$ is the number of the i type characters in the S sequence with $\sum_i b(i) = N$ and $t(j) = 1/n$ for any j. Equation (3) is the equation of a sphere in the $k \times n$ space with R radius. Equation (4)

is the equation of a plane in the $k \times n$ space. Then, the modified matrix was optimized using genetic algorithm and dynamic programming [15] (Figure 1, Paragraph 3). The genetic algorithm was applied immediately to the entire $Q(n)$ set of matrices, in order to create such an M_{max} matrix and such a S subsequence that would have the greatest E_{max}. E_{max} is the maximum similarity function when searching for local alignment [18] of the S sequence, in respect to a certain matrix of the $Q(n)$ set [19]. Each matrix in the genetic algorithm appears to be the genotype, and E_{max} here acts as the fitness function. This procedure has already been described in [15]. As a result of this algorithm operation, we obtained the M_{max} matrix (let us call it the mM_{max}), as well as a fragment of the S sequence (let us call it S'), which possessed the maximum value of the similarity function, when it was aligned with the mM_{max} matrix.

3.4. Generation of the R Random Sequences Set. Random sequences were generated using the S' subsequence (Figure 1, Paragraph 5). The R random sequences set was created using the random mixing of characters in the original S' subsequence. The size of this set contains 200 sequences. The random sequence was created on the basis of the original S' subsequence, by randomly mixing the sequences. For this purpose and using the random numbers sensor, the r sequence was generated with a length of N', where N' is the length of the initial S' subsequence. Then, the r sequence was regularized in ascending order and the permutations made were memorized. Thereafter, the permutations made in the r sequence were applied to the S' subsequence. In total, 200 random sequences were created and were included in the R set.

3.5. Random Sequences Alignment in respect to the Optimized Matrix. For the obtained optimized M_{max} matrix (Section 3.3), the E_{max} average value and value variance were calculated. To do this, we constructed the local alignment of each random sequence out of the R set in respect to the optimized M_{max} matrix (Figure 1, Paragraph 6) [20]. Using this algorithm, we searched for the best local alignment between each sequence out of the R set and the sequence of column numbers of the optimized M_{max} matrix. For this purpose, the matrix for the E similarity function was filled using the optimized $m_{max}(i, j)$ matrix:

$$E(i,j) = \max \begin{cases} 0 \\ E(i-1, j-1) + m_{max}(s(i), l) \\ E(i, j-1) - d \\ E(i-1, j) - d \end{cases} \quad (4)$$

where $s(i)$ is the character sequence element and d is the price for the character insertion or deletion from the alphabet in the S character sequence. Here, i and j vary from 1 to N, $l = j - n \cdot \mathrm{int}((j-1)/n)$. This means that the matrix column with the l number always corresponds to the j index. The E matrix has the $N \times N$ dimension, where N is the length of the character sequence. After filling the F matrix, the following value was used: $E_{max} = E(N, N)$.

3.6. Z(n) Calculation. As a result of calculations using formula (5), the value for each random sequence out of the R set was found. Then, the $\overline{E_{max}}$ average value and the $D(E_{max})$ variance were calculated using the E_{max} set obtained for the $Z(n)$ random sequences as follows:

$$Z(n) = \frac{E_{max} - \overline{E_{max}}}{\sqrt{D(E_{max})}} \quad (5)$$

All calculations were performed for the n period length from 2 to 100.

4. Constructing Multiple Alignment

After completing the algorithm operation, a specific n period length was selected that possessed the greatest Z value. For this period length, a local alignment was constructed, which consisted of two sequences located one below the other. The first sequence was a sequence of indices that periodically varied from 1 to n (denoted as *index S*). The second sequence is the character sequence, and, namely, the transformed poem (denoted as *symbol S*). Local alignment was employed to construct multiple alignment in the following way. The local alignment was divided into short fragments as follows: if the index in the *index S* sequence reached the n value, then the alignment fragment was cut from the entire local alignment and so on until the very end of the local alignment. Thus, a set of the alignment fragments was obtained. It should be noted that both in the *index S* and in the *symbol S* sequences the insertion character or the "$*$" deletion character could be contained. Then, the multiple alignment was constructed using the obtained alignment fragments. In details, the process of constructing multiple alignment is described in [21].

5. Calculating the Chi-Square Distribution Using Multiple Alignment and Its Transfer into Normal Arguments

For the multiple alignment columns, the chi-square distribution was calculated. The column numbers were the period positions and the "$*$" character was not involved in the calculation. The number of letters in the entire local alignment was counted and denoted by L. The number of I type letters was also counted in the entire local alignment and was denoted as $u(i)$, where I is the letter from the alphabet. Then the i type letter probability was calculated within the entire local alignment, as $p(i) = u(i)/L$. The number of letters was counted without taking in to consideration the "$*$" in the j column, and it was denoted as $V(j)$, where j varied from 1 to n. The $f(i, j)$ value denoted the number of i letters in the j column. As a result, the chi-square value was calculated for each column with the (n-1) degree of freedom:

$$x^2(j) = \sum_i \frac{(f(i,j) - p(i) \cdot V(j))^2}{p(i) \cdot V(j)} \quad (6)$$

Thereafter, the obtained chi-square distribution was transformed into the arguments of normal distribution. For this purpose, the Wilson-Hilferty approximation was used, based on the fact that the $(x_v^2/v)^{1/3}$ distribution and the increasing v were approaching normal distribution with the $\mu = 1-2/(9\cdot v)$ mathematical expectation and the $\sigma = \sqrt{2/(9\cdot v)}$ variance [22]. As a result, we obtained a formula for converting to a normal distribution with the number of degrees of freedom equal to (*n-1*):

$$w\left(x^2(j)\right) = \frac{\left(x^2(j)/(n-1)\right)^{1/3} - (1-2/(9\cdot(n-1)))}{(2/(9\cdot(n-1)))^{1/2}} \quad (7)$$

where n is the period length.

6. Calculating Mutual Information

In order to check the relationship between stresses in a poem and period positions, we calculated the mutual information between the stressed vowels and the period positions. In the multiple alignment, the stressed vowel letters were replaced by 1 and other letters were replaced by 0. Then, the number of zeros (0s) and ones (1s) in each multiple alignment column was counted. As a result, the $2\times n$ dimension table was constructed; and the first line indicated the number of zeros (0s) in each column, whereas the second line indicated the number of ones (1s) in each column. In the end, mutual information was calculated according to the following formula [23]:

$$I = \sum_{i=1}^{r}\sum_{j=1}^{n} a_{ij}\ln\left(a_{ij}\right) - \sum_{i=1}^{r} a_{i\cdot}\ln\left(a_{i\cdot}\right) - \sum_{j=1}^{n} a_{\cdot j}\ln\left(a_{\cdot j}\right) + N\ln(N) \quad (8)$$

where $a_{i\cdot} = \sum_j^n a_{ij}$, $a_{\cdot j} = \sum_i^r a_{ij}$, $N = \sum_i^r \sum_j^n a_{ij}$, and a_{ij} are the table elements. The *2I* value was distributed as the chi-square with (*r-1*)(*n-1*) degree of freedom. In order to convert the normal distribution, formula (7) was used. Thus, the w value reflects the correlation of the stressed vowel letters and all the other letters with the period positions. If a correlation is present, the $w(2I)$ values should be greater than 4.0.

7. Study of Artificial Sequences

The developed algorithm was first applied to the study of artificial sequences, one of which was a sequence in the following form: $[abcdefg]_{45}$ (the set of *abcdefg* letters was repeated 45 times); the sequence had an alphabet of 7 letters. Then random substitutions were introduced to this sequence (the number of random substitutions was indicated in % of the initial sequence length). Figure 2 shows that the application of the developed algorithm to an artificial sequence randomly changed by 50%. It could be seen that the Z value takes the maximum value at the 7 letters period length, while the maxima are significant for length periods that are multiples of 7, but these maxima gradually decrease.

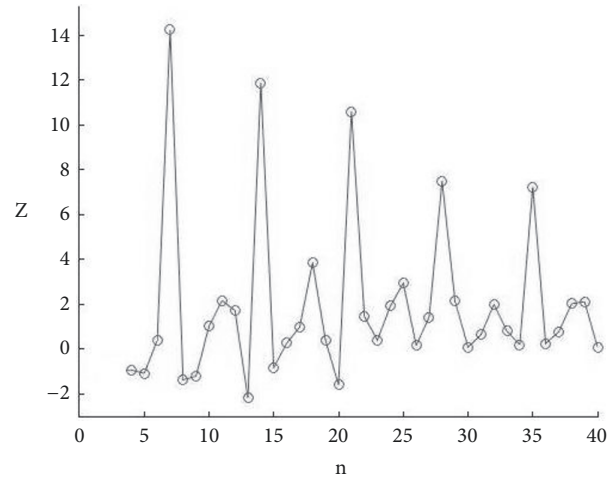

Figure 2: Graph of $Z(n)$ for an artificial sequence randomly changed to 50%.

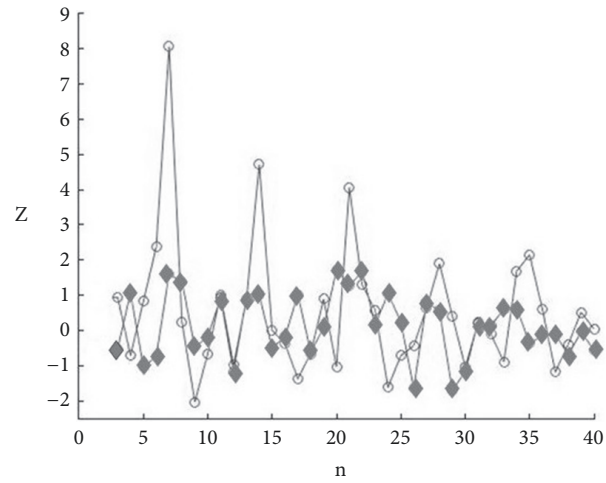

Figure 3: Graph $Z(n)$ for an artificial sequence containing 60% random base substitutions with the addition of 8 inserts and 12 deletions (circles) and $Z(n)$ for a random sequence (rhombus).

Figure 3 shows that the result for the artificial sequence is randomly changed by 60% and involves the addition of 8 inserts and 12 deletions of letters, as well as for the random sequence (obtained by random mixing of the initial sequence). It could be seen that there are no fuzzy periods in the random sequence. The test results show that the technique confidently identifies fuzzy periods in the presence of insertions and deletions of characters, as well as of the substitution of random characters.

8. Searching for Periodicity in Works of Poetry

Afterwards, the developed algorithm was applied to identify fuzzy periods in the works of poetry written in Russian and English languages. Certain results were presented which contained the discovered fuzzy periods in the poems of famous classics. The poem by A. Pushkin, the Russian classic,

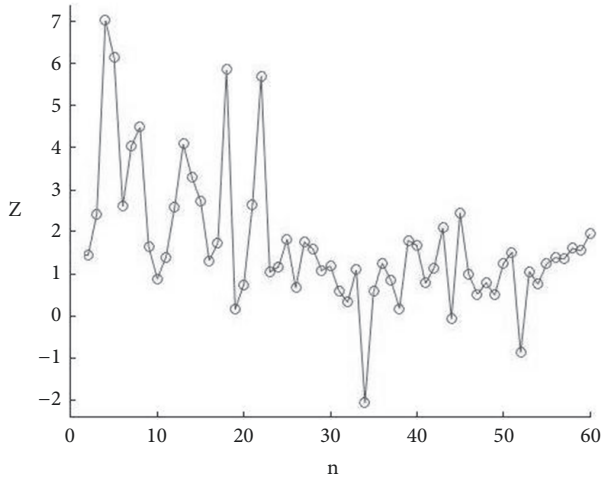

Figure 4: Graph Z (n) for the poem A.S. Pushkin "I remember a wonderful moment..."

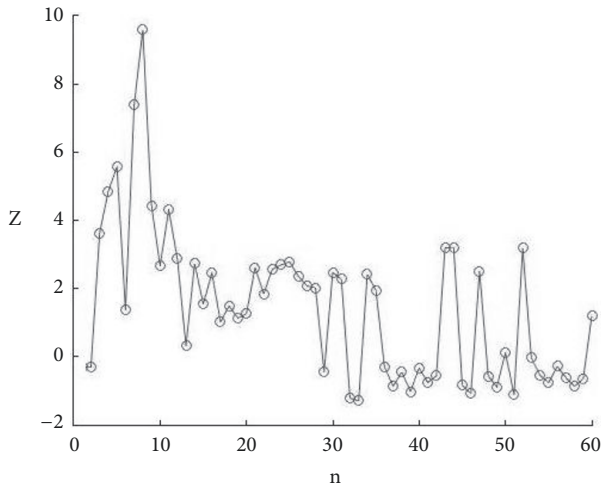

Figure 5: Graph $Z(n)$ for William Blake's poem "Spring."

entitled "I remember a wonderful moment..." was written with the iambic tetrameter (iamb is a two-syllable verse meter with stress on the second syllable in the foot; the foot in this case consists of two syllables). After applying a set of programs to the poem, a graph of the Z dependence upon the n period length was obtained (Figure 4). Z=7.04 and the maximum value exceeds the Z_0 threshold value and is reached at n=4, while the fuzzy period length is equal to 4. The Z_0 threshold value was determined experimentally by calculation based upon the random sequences obtained from the converted poem (initial sequence) by adding a large number of random substitutions to it. It was calculated that after taking into consideration the probability of the Z>6.0, accidental occurrence was less than 5% for all the analyzed poems.

William Blake's poem entitled "Spring" was written in the two-legged trochee (trochee is the two-syllable verse meter with stress on the first syllable in the foot). The Z=9.58 maximum value (Figure 5) is reached at n=8, which means

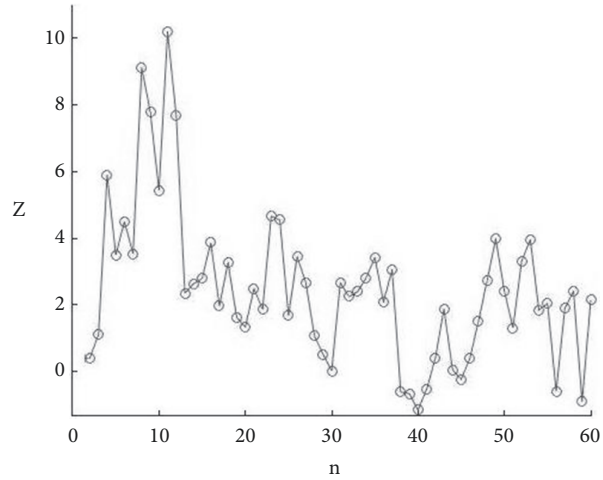

Figure 6: Graph $Z(n)$ for William Blake's poem "The Lamb."

that the length of the fuzzy period is equal to 8. It should be noted that, in some cases and for lengths close to a period with the maximum Z, for example, at n=7, a sufficiently great value of Z is registered. This is due to the addition of superfluous inserts and deletions and, thus, periodicity close to that found is being simulated. The $w(2I)$=4.68 mutual information value was calculated, which indicated the presence of a correlation between the stressed vowels and the period positions. An evaluation of the multiple alignment positions (Section 5) shows that the 6th period position is the most significant.

The poetic meter of the famous poem by William Blake called "The Lamb" is the base trochee. The Z=10.17 maximum value (Figure 6) is reached at n=11; i.e., the length of the fuzzy period is equal to 11. It is interesting to note that the Z values of the length period are greater and are close to n=11. However, this happens with the additional inserts and deletions, thus simulating the main period. The mutual information value for this poem is equal to $w(2I)$ = 7.96, which indicates a strong relationship between stresses in the poem and the period positions. A study of the multiple alignment positions (Section 5) shows that the 10th period position is the most significant.

The poem "Fire and Ice" by Robert Frost was studied; the poem was written with the iambic tetrameter with a variable number of feet stops in the lines, either 4 or 8. The Z=7.55 maximum value (Figure 7) was reached at n=13; i.e., the length of the fuzzy period is equal to 13 letters.

The mutual information value is $w(2I)$=6.00. After calculating the chi-square distribution, it turned out that the most significant is the 4th period position in the multiple alignment. Table 2 presents the multiple alignment. By selecting the most common letter in each period position in the multiple alignment, the following set of letters will be received: *iresoleshith*. After substituting all the stressed letters in the poem with 1s and all the remaining letters with 0s, it becomes absolutely evident that there is a relation between the stresses in the poem and the period positions in the multiple alignment, because the first period position practically consists of 1s (Table 3).

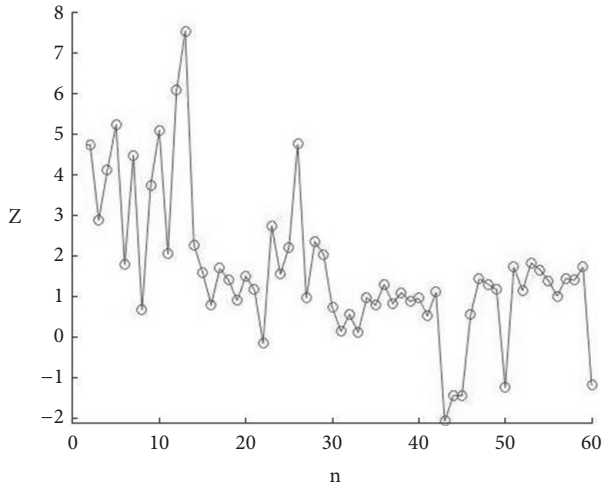

Figure 7: Graph Z (n) for the poem by Robert Frost "Fire and ice."

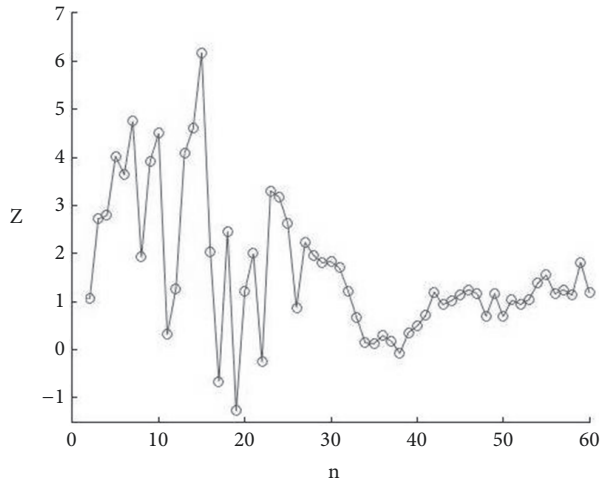

Figure 8: Graph $Z(n)$ for the poem by George Gordon Byron "Remember thee."

In an additional study, we considered the possibility of the existence of an interrelation between the positions of the fuzzy period and the stressed letters in the poem. For example, the result is given for the poem "Fire and Ice" by Robert Frost. In this example, the placement of stresses was done manually. To do this, all the percussive letters in the poem were replaced by 1 and all the other letters by 0. After this, it became evident that there is a relationship between the stresses in the poem and the positions of the period in the multiple alignment, so the first position of the period consists of almost only 1 (Table 3).

The poem by George Gordon Byron "Remember Thee" was written with the iambic tetrameter. The Z=6.16 maximum value (Figure 8) was obtained at n=15; i.e., the length of the fuzzy period is equal to 15. The mutual information value is $w(2I)$=4.90; and the most significant is the 11th position of the period in the multiple alignment, which practically consists of the letter "h." If the most popular letter is selected in the position of each period of the multiple alignment and in case of the same number of certain letters in the column, the one used most rarely in this poem is selected, then the following set will be obtained: *lremombertheea*.

Table 1: Lengths of fuzzy periods found in 95 works of poetry of different authors.

Author	Number of analyzed works of poetry	Lengths of fuzzyperiods found in the works of poetry
Pushkin A.	15	2,4,7,10
YeseninS.	10	2,6,8,11
Blok A.	5	3,5,10,11
Tutchev F.	5	2,7,8,15
Fet A.	5	2,4,6,9
Mayakovsky V.	7	2,3,5,8
Shakespeare W.	5	10,15,18,37
Byron D.	5	8,16,28
Frost P.	20	5,8,13,15,20
Blake W.	18	4,6,8,10,11,23

It should be noted that, in English, it is not the letter that strikes but the sound. Since the sound can consist of several vowel letters, then for the sake of certainty, the first letter in the sound was considered (marked) by the stressed letter. Concerning the very arrangement of accents in this poem, we arranged them according to the poetic size (iambic, trochee, etc.). Therefore, in the case of chorea, which is characterized by an alternating sequence, a shock and then an unstressed sound, the poem was placed stress. However, to search for the periods themselves, as earlier noted, only the text itself is used and no other information is required.

In total, 95 poems by Russian and foreign poets were studied. In more than half of the poems studied, fuzzy periods with the Z>6.0 value were found. The other half also had fuzzy periods, but the Z level is lower than 6.0. These results could be explained by the fact that not all poems have a "clear structure" and rhyme. They also combine poetic dimensions that make it difficult to detect the periodicity that is often used. There is another explanation, which is connected with the fact that in many cases a large number of insertions or deletions of symbols in the text are required to notice fuzzy periods. Such causes can lead to a relatively low level of statistical significance of fuzzy periods.

Table 1 shows that the short lengths of the fuzzy period are mostly often encountered in Russian poems, whereas in English poems the lengths of the fuzzy period are longer. This can be explained by the structure of the language. For example, in the Russian language there is a frequent alternation of vowel and consonant letters, and in the English language a case is more widely spread, where several consonant or vowel letters are consecutive, which in turn prolongs the period.

9. Conclusions

The present study aimed at evaluating the efficiency of a new technique in searching for fuzzy periods which are accompanied by insertions and deletions [15] in the texts of

Table 2: Multiple alignment of Robert Frost's poem "Fire and Ice." The zero line shows the positions of the period.

0)	*	*	1	2	*	3	*	4	*	5	6	7	8	9	*	10	*	*	11	12	13
1)	*	*	*	*	*	*	*	*	*	s	o	m	e	s	*	a	*	*	y	t	h
2)	e	w	o	r	l	d	*	w	*	i	l	l	e	n	*	d	*	*	i	n	f
3)	*	*	i	r	*	e	*		*	s	o	m	e	s	*	a	*	*	y	i	n
4)	*	*	i	c	*	e	*		*	f	r	o	m	w	*	h	*	*	a	t	*
5)	*	*	i	v	*	e	t	a	*	s	t	e	d	o	*	f	*	*	d	e	s
6)	*	*	i	r	*	e	*		i	h	o	l	d	w	i	t	*	*	h	t	h
7)	*	*	o	s	*	e	*	w	*	h	o	f	a	*	*	v	*	*	o	r	f
8)	*	*	i	r	*	e	*		*	*	b	u	t	i	*	f	*	*	i	t	h
9)	*	*	a	*	*	d	*	*	*	t	o	p	e	r	*	i	s	*	h	t	w
10)	*	*	i	c	*	e	*		*	i	t	h	i	n	*	k	*	*	i	k	n
11)	*	*	o	w	*	e	*	*	*	n	o	*	u	g	*	h	*	*	o	f	h
12)	*	*	a	t	*	e	*		*	t	o	s	a	y	t	h	*	*	a	t	f
13)	*	*	o	r	*	d	*	e	*	s	t	r	u	c	*	t	*	*	i	o	n
14)	*	*	i	c	*	e	*		i	s	a	l	s	o	*	g	r	e	a	t	
15)	*	*	a	n	*	d	*	w	*	o	u	l	d	s	*	*	*	*	u	f	f
16)	*	*	i	c	*	e	*		*	*	*	*	*	*	*	*	*	*	*	*	*

Table 3: Multiple alignment with the replacement of letters by 0 and 1 for the work of Robert Frost "Fire and Ice." The null line shows the positions of the period.

0)	*	*	1	2	*	3	*	4	*	5	6	7	8	9	*	10	*	*	11	12	13
1)	*	*	*	*	*	*	*	*	*	0	0	0	0	0	*	1	*	*	0	0	0
2)	0	0	1	0	0	0	*	0	*	0	0	0	1	0	*	0	*	*	0	0	0
3)	*	*	1	0	*	0	*	0	*	0	0	0	0	0	*	1	*	*	0	0	0
4)	*	*	1	0	*	0	*	0	*	0	0	0	0	0	*	0	*	*	1	0	*
5)	*	*	0	0	*	0	0	1	*	0	0	0	0	1	*	0	*	*	0	0	0
6)	*	*	1	0	*	0	*	0	0	0	1	0	0	0	0	0	*	*	0	0	0
7)	*	*	1	0	*	0	*	0	*	0	0	0	1	*	*	0	*	*	0	0	0
8)	*	*	1	0	*	0	*	0	*	*	0	0	0	1	*	0	*	*	0	0	0
9)	*	*	1	*	*	0	*	*	*	0	0	0	1	0	*	0	0	*	0	0	0
10)	*	*	1	0	*	0	*	0	*	0	0	0	1	0	*	0	*	*	0	0	0
11)	*	*	1	0	*	0	*	*	*	0	1	*	0	0	*	0	*	*	0	0	0
12)	*	*	1	0	*	0	*	0	*	0	0	0	1	0	0	0	*	*	0	0	0
13)	*	*	1	0	*	0	*	0	*	0	0	0	1	0	*	0	*	*	0	0	0
14)	*	*	1	0	*	0	*	0	0	0	1	0	0	0	*	0	0	1	0	0	0
15)	*	*	0	0	*	0	*	0	*	1	0	0	0	0	*	*	*	*	0	0	0
16)	*	*	1	0	*	0	*	0	*	*	*	*	*	*	*	*	*	*	*	*	*

poems. The applied mathematical method analyzes the text of a poetic work and does not require any other information. It should also be noted that we were unable to find a statistically significant periodicity in ordinary novels both in the Russian and in the English languages. This shows that the fuzzy periods within the linguistic texts are observed exclusively in works of poetry. In general, the results of the present study suggest that a poet uses certain acoustic waves, when writing a poem. It could be noted that poets use a fairly diverse set of acoustic wave lengths, when creating a poem. Probably the fuzzy periodicity of the text is reflection of such acoustic waves. It could be assumed that the acoustic wave is rather important to ensure the psychological impact on the audience, when reciting a poem.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

In this work, we used the resources of the high-performance computing center of the National Nuclear Research University "MEPhI."

Supplementary Materials

Guide for launching programs: (1) Generation and selection of random matrices for each length of the period n (n = 2..100) by the program ENG_ComparisonMatrix. The program code in the file "ENG_ComparisonMatrix.c". This program must be run alternately for each length of the period n, which changes in the program code in the variable "n." For example, in order to generate a plurality of matrices for a period length of 2, it is necessary to (a) open the file "ENG_ComparisonMatrix.c" and (b) assign a value of 2 to the variable "n"; it should be "n = 2;". (c) In the program code at the end of the filename to store the matrices to make 2, the following should turn out: FILE ∗ fp1 = fopen ("matrix_n_2.txt", "w"); (d) save changes to the program file and compile. As a result of the program, we get the file "matrix_n_2.txt", which will contain the matrix. And the file "report_number_matrix.txt", in which the number of generated matrices will be stored (this amount is useful in the program ENG_POEM). After starting the program for each length of the period, we get a set of files of the form: "matrix_n_2.txt", "matrix_n_3.txt", "matrix_n_4.txt", and so on to "matrix_n_100.txt". In the program file "ENG_POEM.c" in the array int Number [101] add the value of the number of matrices for each length of the period. For example, for the length of the period n = 2, 1062 matrices were obtained, and then in the array it is necessary to fill the third position and then get: int Number [101] = {0, 0, 1062, ...}. (2) Choosing a poem for research: we go in the folder with poems "95 poem" and open the folder with the desired author and then copy the text from the file with the poem to the file "poem.txt" in the folder with the program Changetext. (3) Changetext program converts the original poem into a sequence for research (removes spaces, punctuation marks replaces with spaces). The program code in the file "Changetext.c". The input file of the program is a file with a poem "poem.txt". The output file of the program is the file "poemSequence.txt". (4) The mixingSeq program generates 200 random sequences from the original sequence in the file "poemSequence.txt". The program code in the file "mixingSeq.c". The input file of the program is the file "poemSequence.txt". The output file of the program is the file "Sequences.txt". (5) The ENG_POEM program checks poems in English for periodicity. The program code in the file "ENG_POEM.c". This program must be run alternately for each length of the period n, which changes in the program code in the variable "n" with the parameters d and Kd given for each poem. Parameters d and Kd are indicated for each poem in the file "d_Kd_poem.txt". Taking the values of these parameters, it is necessary to fill them in the code of the program "ENG_POEM.c". For example, in order to run the program for the length of the period n = 2, with the parameters d = 5, Kd = -0.2, it is necessary to (a) open the file "ENG_POEM.c" and (b) assign a value of 2 to the variable "n", and it should be "n = 2;" and (c) in the code of the program at the end of the file name to store the matrices to make 2, the following should turn out: ∗ Matrix = fopen ("matrix_n_2.txt", "r"); (d) in the program code at the end of the file name for the report to make 2, the following should turn out: ∗ Report = fopen ("report_n_2.txt", "a"); (i) assign a value of 5 to the program variable "d" and it should be "d = 5;". (f) Assign the value -0.2 to the variable "Kd" and it should be "Kd = -0.2;" and (g) save changes to the program file and compile. Input files of the program are the following files: "Sequences.txt" and "matrix_n_2.txt" ("matrix_n_2.txt", in this case n = 2, and if another length of periodicity is investigated, then it is necessary to replace 2 with another value). The output file of the program is the file "report_n_2.txt", and this file stores the local and multiple alignment of the investigated sequence, as well as local alignments for random sequences. At the end of this file, a value of statistical significance for the period n = 2 is indicated in a line of type Z (2) = 4.800792. Carrying out the calculations for each length of the period n we get a set of files of the form: "report_n_2.txt", "report_n_3.txt", "report_n_4.txt",..... "report_n_100.txt". Then a graph is constructed from the values of Z (2), Z (3), Z (4), ... Z (100). *(Supplementary Materials)*

References

[1] H. Wang, P. Mok, and H. Meng, "Capitalizing on musical rhythm for prosodic training in computer-aided language learning," *Computer Speech and Language*, vol. 37, pp. 67–81, 2016.

[2] F. Orsucci, A. Giuliani, C. Webber Jr., J. Zbilut, P. Fonagy, and M. Mazza, "Combinatorics and synchronization in natural semiotics," *Physica A: Statistical Mechanics and its Applications*, vol. 361, no. 2, pp. 665–676, 2006.

[3] O. A. Rosso, H. Craig, and P. Moscato, "Shakespeare and other English Renaissance authors as characterized by Information Theory complexity quantifiers," *Physica A: Statistical Mechanics and its Applications*, vol. 388, no. 6, pp. 916–926, 2009.

[4] E. V. Korotkova, M. A. Korotkovaa, and N. A. Kudryashova, "Information decomposition method to analyze symbolical sequences," *Physics Letters Section A: General, Atomic and Solid State Physics*, vol. 312, no. 3-4, pp. 198–210, 2003.

[5] E. V. Korotkov, M. A. Korotkova, and J. S. Tulko, "Latent sequence periodicity of some oncogenes and DNA-binding protein genes," *Computer Applications in the Biosciences*, vol. 13, no. 1, pp. 37–44, 1997.

[6] V. P. Turutina, A. A. Laskin, N. A. Kudryashov, K. G. Skryabin, and E. V. Korotkov, "Identification of amino acid latent periodicity within 94 protein families," *Journal of Computational Biology*, vol. 13, no. 4, pp. 946–964, 2006.

[7] Z. R. Struzik, "Wavelet methods in (financial) time-series processing," *Physica A: Statistical Mechanics and its Applications*, vol. 296, no. 1-2, pp. 307–319, 2001.

[8] V. Afreixo, P. J. S. G. Ferreira, and D. Santos, "Fourier analysis of symbolic data: a brief review," *Digital Signal Processing*, vol. 14, no. 6, pp. 523–530, 2004.

[9] G. Benson, "Tandem repeats finder: a program to analyze DNA sequences," *Nucleic Acids Research*, vol. 27, no. 2, pp. 573–580, 1999.

[10] M. Pellegrini, "Tandem repeats in proteins: Prediction algorithms and biological role," *Frontiers in Bioengineering and Biotechnology*, 2015.

[11] J. Jorda and A. V. Kajava, "T-REKS: identification of Tandem REpeats in sequences with a K-meanS based algorithm," *Bioinformatics*, vol. 25, no. 20, pp. 2632–2638, 2009.

[12] S. Kurtz, J. V. Choudhuri, E. Ohlebusch, C. Schleiermacher, J. Stoye, and R. Giegerich, "REPuter: The manifold applications of repeat analysis on a genomic scale," *Nucleic Acids Research*, vol. 29, no. 22, pp. 4633–4642, 2001.

[13] I. Elias, "Settling the intractability of multiple alignment," *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology*, vol. 13, no. 7, pp. 1323–1339, 2006.

[14] H. T. Wareham, "A Simplified Proof of the NP- and MAX SNP-Hardness of Multiple Sequence Tree Alignment," *Journal of Computational Biology*, vol. 2, no. 4, pp. 509–514, 1995.

[15] V. M. Pugacheva, A. E. Korotkov, and E. V. Korotkov, "Search of latent periodicity in amino acid sequences by means of genetic algorithm and dynamic programming," *Statistical Applications in Genetics and Molecular Biology*, vol. 15, no. 5, pp. 381–400, 2016.

[16] F. E. Frenkel, M. A. Korotkova, and E. V. Korotkov, "Database of Periodic DNA Regions in Major Genomes," *BioMed Research International*, vol. 2017, Article ID 7949287, 9 pages, 2017.

[17] A. Oras, *Pause Patterns in Elizabethan and Jacobean Drama: An Experiment in Prosody*, 1960.

[18] T. F. Smith and M. S. Waterman, "Identification of common molecular subsequences," *Journal of Molecular Biology*, vol. 147, no. 1, pp. 195–197, 1981.

[19] A. A. Laskin, E. V. Korotkov, M. B. Chaley, and N. A. Kudryashov, "The locally optimal method of cyclic alignment to reveal latent periodicities in genetic texts. The NAD-binding protein sites," *Molekularna Biologija*, vol. 37, no. 4, pp. 663–673, 2003.

[20] S. B. Needleman and C. D. Wunsch, "A general method applicable to the search for similarities in the amino acid sequence of two proteins," *Journal of Molecular Biology*, vol. 48, no. 3, pp. 443–453, 1970.

[21] F. Sievers and D. G. Higgins, *Multiple Sequence Alignment Methods*, 2014.

[22] E. B. Wilson and M. M. Hilferty, "The Distribution of Chi-Square," *Proceedings of the National Acadamy of Sciences of the United States of America*, vol. 17, no. 12, pp. 684–688, 1931.

[23] S. Kullback, *Information Theory and Statistics*, Dover, New York, NY, USA, 1997.