Enlarged Similarity of Nucleic Acid Sequences

Eugene V. KOROTKOV1 and Maria A. KOROTKOVA2

117312, Bioengineering Center, Russian Academy of Sciences, 60 Oktyebra prospect, 7/1, Moscow, Russia1 and Department of Cybernetics, Moscow Physical Engineering Institute, 115409, Kashirskoe chosse, 31, Moscow, Russia2

(Received 20 February 1996; revised 13 May 1996)

Abstract

The concept of nucleic acid sequence base alternations is presented. The number of base alterations for the sequences of different length is established. The definition of "enlarged similarity" of nucleic acids sequences on the basis of sequence base alterations is introduced. Mutual information between sequences is used as a quantitative measure of enlarged similarity for two compared sequences. The method of mutual information calculation is developed considering the correlation of bases in compared sequences. The definitions of correlated similarity and evolution similarity between compared sequences are given. Results of the use of enlarged similarity approach for DNA sequences analysis are discussed.

Key words: DNA sequence; computer analysis; sequence base alteration; mutual information; enlarged similarity

1. Introduction

The development of fast methods of nucleic acid sequence determination led to the accumulation of sequences of different genes, intrones and intergenic regions. The execution of the “Human Genome” program should result in sequencing of the whole human genome and genomes of other species.1,2 The results obtained will be used in science and medicine. However, such use is possible only after understanding the meaning of the information obtained. This problem is analogous to trying to decipher a text in an unknown language. First of all, it is necessary to find the borders, and the meanings of words and then to do the same for sentences. It is obvious that experimental approach gives a clearer understanding of a genetic text. But it is limited by the fragmentary nature of the text. This approach has a relatively narrow field of application and may characterize a very small part of determined sequences. The experimental methods make it possible to find coding regions, intrones and a small quantity of sequences, taking part in the process of genetic activity regulation.3 The theoretical approach has a wider field of application in that several DNA sequences can be compared. Well-known statistical methods help to find correlations of base pairs in DNA sequences, as well as some peculiarities of the alteration of nucleotides in coding and noncoding regions.4,5

The search for homologous regions led to the discovery of the sites in DNA sequences that are homologous to each other. Repeated sequences of different size, regions with small number of copies in genome (pseudogenes and some duplication sites) are classified as homologous regions. Incomplete homology and homology with deletions were also found.4,5

However, it is difficult to decide that two sequences are similar to each other if their homology is rather low. If any selected sequence is compared with the set of sequences from a big data bank, then the possibility of finding an accidental homology of low level is very high. This means that the similarity between ancient sequences can not be found by homology searches on a statistically important level. Certain coding regions and certain protein-binding sites, genes and repeats from evolutionarily distinct species could serve the examples of sequences similarity of which it is difficult to be found on statistically important level by homology search only. It shows the necessity for development of new methods of genetic information analysis.

A new method of the search of nucleic acid sequences enlarged similarity is presented. The method enables the determination of DNA sequence base alteration. it is possible to apply the mathematical methods developed for any other text analysis.
2. Results and Discussion

2.1. The enlarged similarity approach to DNA sequence analysis

The general idea of the enlarged similarity approach to DNA sequence analysis is to define a new concept of the sequence base alternation and to use this concept for elaboration of the method for the search of similar sequences. We mean that the base alteration is a type of DNA base order and it is independent of DNA base properties. The exact definition is given in the next section.

We can illustrate the concept of the sequence base alternation as follows. Let there be some storage cells. Each cell contains one symbol and all symbols are different. Let there be a computer program for building of the sequence from symbols by address transport from cells. For example, the program may put the symbol from the second cell to the first place, the symbol from the third cell to the second place, the symbol from the first cell to the third place and so on. The program can build different sequences from different cell content, but the alternation of symbols is the same for all those sequences. After we define the concept of the sequence base alternation, we can determine whether two sequences have the same or different base alternations. The same base alternation for two sequences is a special case of the base alternation similarity. We can illustrate the obtaining of similar base alternations by the computer program model where some cells contain more than one symbol and the choice of the symbol from a cell is arbitrary. We say that two sequences have enlarged similarity if they have the same or similar base alternation.

The enlarged similarity of sequences can be more or less close and we must define the measure of similarity. We choose the mutual information between sequences as a measure of the enlarged similarity.

The method for a search of sequences with enlarged similarity is as follows. The majority of similar sequences having the largest value of mutual information are chosen from the set of all sequences to be analyzed. The comparison of the base alternations in these sequences shows the type of sequence similarity. So, the comparison of base alternation helps to find the biological sense of the sequences similarity.

It is important to consider the biological processes connected with the enlarged similarity between DNA sequences. It is possible to assume that the enlarged similarity may be created by evolutionary changes in DNA sequences. Firstly, enlarged similarity may be found between DNA sequences having a common origin from one ancient sequence that diverged during evolution. These sequences may have a low and statistically unimportant level of homology as a result of many base changes, but transitions occur more frequently than transversions and such sequences may have about 80–100% coincidences in sites of purines and pyrimidines. The enlarged similarity studies of DNA sequences may reveal this similarity of DNA sequences on a statistically important level. Secondly, if base changes in DNA sequences occur in the presence of several limitations, then there is a possibility for the creation of more complex relations between DNA sequences. Those relations differ from homology, complementarity or purine-pyrimidine relations and may be found between coding regions of DNA sequences or between protein-binding sites. Coding regions and protein-binding sites have limitations of base changes because the biological function of a sequence should be preserved.

So, the search for enlarged similarity between DNA sequences could become an important method for revealing ancient genome duplications and divergences of some repeats, functional signals and genes. The BLAST26 and FASTA27 programs, as well as programs for homology search4,5 are not able to find statistically important similarities between DNA sequences if the sequences have a low level of homology or if there are relations of another type between bases.

2.2. Sequence base alternation and enlarged similarity of sequences

The concept of sequence base alternation reflects the sequential order of the letters. For the sequences having one-to-one mapping to another sequence, we define that those sequences have the same base alternation. Let there be two sequences of letters from the alphabet \{A,T,C,G\} as in example 1. It is clear intuitively that these two sequences are similar and their similarity is conditioned by the order of the letter alternation.

Example 1.

\[ 5'\text{ATCGAGCGACGTAGA}-3' \]
\[ 5'\text{GCTAGATGACGTACG}-3' \]

The base alternation of the sequence \( B = \{b_1, b_2, \ldots, b_n\} \) of letters from the alphabet \( \{a_1, a_2, \ldots, a_k\} \) is the set \( S = \{S_1, S_2, \ldots, S_k\} \) where \( S_i \) is the sequence of the numbers of letter \( a_j \) occurrences in sequence \( B \). For the alphabet containing 4 letters A, T, G, and C, the base alternation consists of 4 sets \( S_A, S_T, S_C \) and \( S_C \). If a sequence does not contain a letter \( a_i \), then the sequence \( S_i \) of its numbers of occurrences is empty. Let us consider two sequences of example 1.

For the first sequence \( S_A = \{1,5,9,12,15\} \), \( S_T = \{2,13\} \), \( S_C = \{3,7,10\} \), \( S_G = \{4,6,8,11,14\} \) and the base alternation \( S_1 = \{1,5,9,12,15\}, \{2,13\}, \{3,7,10\}, \{4,6,8,11,14\} \) for the second sequence from the example 1 \( S_A = \{4,6,8,11,14\} \), \( S_T = \{3,7,10\} \), \( S_C = \{2,13\} \), \( S_G = \{1,5,9,12,15\} \) and the base alternation \( S_2 = \{4,6,8,11,14\}, \{3,7,10\}, \{1,5,9,12,15\}, \{2,13\} \) is easy to see that \( S_1 \) and \( S_2 \) sets are the same set, because reordering the elements does not change the set.

Different numbers of base alternations exist for se-
sequences of different lengths. The numbers of base alternations for sequences of different length are represented in Table 1. It is possible to note that the base alternation containing one non-empty set of numbers has 4 different sequences with this base alternation. For every base alternation with two non-empty sets of numbers corresponding to the sequences with 2 different base types, there are 12 sequences with this base alternation. For the base alternation containing 3 or 4 non-empty sets of numbers there are 24 sequences with this kind of base alternation. It means that for every sequence that is built of one base type, 3 other sequences exist with the same base alternation pattern. For each sequence built of 2 base types, there are 11 other sequences with the same base alternation. For sequences with 3 or 4 different base types, there are 23 other sequences with the same base alternation. The number of sequences with the same base alternation pattern depends on the number of base types in the sequence and does not depend on the sequence length. Sequences that differ insignificantly in one-to-one mapping have insignificantly different base alternations. For example, if the second sequence differs from the first one at only one base, then the base alternations of these two sequences differ by the position of one number only.

Let us consider two sequences, one of which is the first sequence of example 1 and the second one differs at one base.

Example 2.
5'-ATCGAGCGACGATGA-3'
5'-GTTAGATAGTAGCAG-3'

For the first sequence from example 2 the base alternation is $S_1 = \{ \{1,5,9,12,15\}, \{2,13\}, \{3,7,10\}, \{4,6,8,11,14\} \}$. For the second sequence the base alternation $S_2 = \{ \{1,5,9,12,15\}, \{13\}, \{2,3,7,10\}, \{4,6,8,11,14\} \}$. It is easy to see that base alternations differ by the position of one number only. Two sequences $A = \{a_1, a_2, \ldots, a_n\}$ with the base alternations $S_A$ and $B = \{b_1, b_2, \ldots, b_n\}$ with the base alternation $S_B$ are called the identical base alternations if the base alternations $S_A$ and $S_B$ are identical. Base alternation identity is a special case of alternation similarity. The enlarged similarity may be obtained by the regular declination from base alternation identity. The regular declination from the base alternation identity is shown in example 3. Letter A in the first sequence corresponds to two letters G and C in the second sequence.

Example 3.
5'-ATCGAGCGACGATGA-3'
5'-GCTAGATACTAGCAC-3'

The base alternation of the first sequence is $\{ \{1,5,9,12,15\}, \{2,13\}, \{3,7,10\}, \{4,6,8,11,14\} \}$ and the base alternation of the second sequence is $\{ \{1,5,12\}, \{2,9,13,15\}, \{3,7,10\}, \{4,6,8,11,14\} \}$. Two sets in base alternations are identical. Two other sets are characterized by such property: sets $\{1,5,9,12,15\}$ and $\{1,5,12\}$ have common numbers and all other numbers, that are absent in the set of the second base alternation, are included in the other set of the second base alternation.

We define the base alternation similarity of the first type for the sequences that have not got identical base alternation and have 3 or 4 different base types each. The sequence A with the base alternation $S^A = \{ S_1^A, S_2^A, S_3^A, S_4^A \}$ and the sequence B with the base alternation $S_B = \{ S_1^B, S_2^B, S_3^B, S_4^B \}$ have base alternations similar to the first type, if for every $S_j^A$ there exist $S_j^B, S_k^B$ such that $S_j^A \subseteq S_j^B \cup S_k^B$ and for every $S_j^B$ there exist $S_j^A, S_k^A$ such that $S_j^B \subseteq S_j^A \cup S_k^A$ (or each base type of one sequence corresponds to one or two base types of other sequence). We define the alternation similarity

| Table 1. The numbers of base alternations for sequences of different lengths. |
|---------------------------------------------------------------|
| Length of sequence | Number of sequences | Total number of base alternations | Number of base alternations for different number of bases in the sequence |
|-------------------|-------------------|-------------------------------|------------------------------------------------------------------|
|                   |                   |                               | Number of different bases | Possible number of base alternations | Number of sequences for one base alternation |
| 1                 | 4                 | 1                             | 1                       | 1                                | 4                                                  |
| 2                 | 16                | 2                             | 1                       | 1                                | 4                                                  |
| 3                 | 64                | 5                             | 1                       | 1                                | 4                                                  |
| n                 | $4^n$             | $(4^{n-1} + 2)/6 + 2^{n-2}$   | 1                       | 1                                | 4                                                  |

Downloaded from https://academic.oup.com/dnaresearch/article-abstract/3/3/157/342343 on 29 July 2018
of the second type for sequences that have not got a base alternation identical or similar to the first type and have 4 different base types each. Sequence A with the base alternation $S_A = \{S_1^A, S_2^A, S_3^A, S_4^A\}$ and sequence B with the base alternation $S_B = \{S_1^B, S_2^B, S_3^B, S_4^B\}$ have base alternations similar to the second type if for every $S_i^A$ there exist $S_j^B, S_k^B, S_l^B$ such that $S_i^A \subseteq S_j^B \cup S_k^B \cup S_l^B$ and for every $S_i^B$ there exist $S_j^A, S_k^A, S_l^A$ such that $S_i^B \subseteq S_j^A \cup S_k^A \cup S_l^A$ (or each base type of one sequence corresponds to one, two or three base types of the other sequence). Sequences from example 3 correspond to the first type of the alternation similarity. If the majority of bases in two sequences corresponds to the alternation similarity and the rest bases decline from it, we say that it is imperfect alternation similarity. We define that sequences have enlarged similarity if they have identical base alternation or similar base alternation of any type. There may be various cases of the alternation similarity. The degree of enlarged similarity is characterized by the mutual information that has maximum value when the sequences have identical base alternations.

2.3. Measure of the nucleic acids sequences enlarged similarity

The mutual information between two compared sequences is taken as the measure of their similarity.\(^6\)\(^7\)\(^8\)\(^9\) The mutual information is calculated with the help of the matrix of base coincidences between compared sequences. The number of all base coincidences is calculated in a 4 x 4 matrix when two sequences with length L are compared. The main diagonal elements show the number of AA, TT, CC and GG coincidences. Then sum of all 16 elements of the matrix is equal to the length of compared sequences. The mutual information is calculated as:

$$H(1, 2) = (H(1) + H(2) - H(1, 2))L \ln 2$$

(1)

$H(1)$ is the average entropy of the first sequence for one base. $H(2)$ is the average entropy of the second sequence for one base. $H(1, 2)$ is the average entropy of the "coincidence sequence" for one type of coincidences. The "coincidence sequence" is the sequence that has a 16-letter alphabet. The base pairs (first base from the first compared sequence and the second base from the second compared sequence) are the letters of the coincidence sequence. The average entropy in any sequence for one letter is calculated as (10):

$$H = \sum_{i=1}^{m} p_i \log_2 p_i$$

(2)

where the size of the alphabet used is $m$. The average probability of letter $i$ in a sequence is $p_i$. For DNA sequences, $m$ is equal to 4 and $p_i$ is equal to the numbers of A, T, C, and G bases in sequences divided by the length of compared sequences. It is supposed that sequences of equal lengths are compared. For the coincidence sequence, $m$ is equal to 16 and $p_i$ is the probability for each of 16 letters. The average value of $2H(1, 2)$ is equal to 9. $2H(1, 2)$ is distributed with 9 degrees of freedom when accidental sequences a with 4-letter alphabet are compared.\(^11\) The mutual information as a measure of sequence similarity makes it possible to find the degree of sequence similarity when the alternation similarity of sequences is imperfect. There is a complete analogy with an imperfect homology of sequences. The mutual information here is always less than it is for the perfect base alternation similarity of sequences. The mutual information makes it possible to determine the degree of the alternation similarity for two sequences when compared sequences are not a synonymous reflection of each other. If one of the sequences and the matrix of coincidences is known, than it is impossible to restore another sequence here. An example of such a complex relation between sequences is the coincidences of the sites of purines and pyrimidines in two compared sequences. Such relations between sequences are represented in their base alternations by the following: base alternations coincide after joining the two sets corresponding to purines and the other two sets corresponding to pyrimidines in each base alternation.

The mutual information is a suitable measure of DNA sequence similarity because DNA sequence relations can be revealed at once. It is very important because the type of base coincidences is not known beforehand. The mutual information makes it possible to estimate the probability of accidental sequence similarity as well. We may consider DNA sequences as related to each other if the mutual information is greater than some cut-off value.\(^9\) It is possible to determine the type of relations between compared sequences if we can analyze the matrix of coincidences. Other methods of searching for enlarged similarity are very difficult because they require testing of all possible cases of DNA sequence relations.

The number of those relations is large, and that makes calculations complicated. Besides, modifications of BLAST\(^26\) and FASTA\(^27\) programs are not possible because they may reveal only one type of sequence relation (homology, for example). So, new methods of DNA sequence comparison are required which use mutual information as a measure of DNA sequence similarity, as well as of new software.

2.4. The account of base correlation in compared sequences

The frequencies of the k-long chains are calculated for evaluation of the mutual information for sequences with correlated nearest bases. If the alphabet of sequence has m letters then $m^k$ k-long chains are possible. The mutual information $I_k(1, 2)$ is also determined using formula 1.
but values $H(1)$, $H(2)$, and $H(1,2)$ are calculated as:

$$H(1) = \sum_{\alpha} p(B^k(\alpha)) \log_2 p(B^k(\alpha))$$

$$H(2) = \sum_{\beta} p(B^k(\beta)) \log_2 p(B^k(\beta))$$

$$H(1,2) = \sum_{\gamma} p(B^k(\gamma)) \log_2 p(B^k(\gamma))$$

Here $B^k(\alpha)$, $B^k(\beta)$, and $B^k(\gamma)$, are the $k$-long chains in the first and second of compared sequences, as well as in the coincidence sequence with numbers $\alpha$, $\beta$, and $\gamma$. If the first compared sequence has an $m_1$-letter alphabet and the second compared sequence has an $m_2$-letter alphabet, then the coincidence sequence has an $m_1m_2$-letter alphabet. The number of $k$-long chains in the first compared sequence is equal to $(m_1)^k$, in the second compared sequence it is equal to $(m_2)^k$ and in the coincidence sequence the number of $k$-chains is equal to $(m_1m_2)^k$. For DNA sequences $m_1 = m_2 = 4$. It is possible to calculate the number of $k$-long chains $n(B^k(\alpha))$ for any sequence with length $L$ ($L \gg k$), where $\alpha = 1, 2, \ldots, m^k$. Then it is possible to estimate the $p(B^k(\alpha))$ probability as:

$$p(B^k(\alpha)) = n(B^k(\alpha))/(L-k+1)$$

So, if two compared sequences have enough length, then the base correlation for any length may be considered in $I(1,2)$. The conditional mutual information $F_2, F_3, \ldots, F_k$ correspond to the absolute mutual information $I_2(1,2), I_3(1,2), \ldots, I_k(1,2)$.

$$F_k = kl_k(1,2) - (k-1)I_{k-1}(1,2)$$

The mutual information $F_{k+1}$ is a conditional mutual information of the next element if $k$ preceding elements are known. If $k$ preceding elements determine the $(k+1)$ element of the coincidence sequence, then $F_{k+1}$ is equal to 0. If two accidental sequences are compared, then $F_2 = F_3 = \ldots = F_k$. The $I_2(1,2)$ calculation is possible for compared sequences with length more than $10^3$ only, because the coincidence sequence has a 16-letter alphabet. It gives 256 pairs of combinations of those letters. For sequences with smaller length, the application of another method is possible.

### 2.5. Analysis of base coincidences distinguished from accidental base coincidences

It is possible to consider the matrix of accidental coincidences $N(4,4)$ which is calculated as:

$$N(i,j) = X(i)Y(j)/L^2$$

$X(i)$ are the number of A, T, C, and G in the first sequence, and $Y(j)$ are A, T, C, and G in the second sequence. $L$ is the length of a sequence. Matrix $N$ shows the number of coincidences for each type of base pair when two accidental sequences with given A, T, C, and G numbers are compared. It is possible to find the type of coincidence that is distinguished from the accidental coincidence more, when $M$ and $N$ matrices are known:

$$f_A = \max \{M(i,j) - N(i,j)\}$$

It is possible to consider two events when two sequences are compared. $A$ is the event that includes the base coincidences, that could be distinguished more from base coincidences for accidental sequences (with the same A, T, C, and G composition). $B$ is the event that includes other types of base coincidences. The event $A$ may include one of 16 possible coincidences of bases. For example it may be AT. Then the event $B$ includes other 15 types of coincidences of bases, for example: AA, TT, CC, GG, AG, GA, TC, CT, AC, CA, GT, TG, CG, GC, TA. When the matrix $M$ is known, it is possible to calculate the average probability for events $A$ and $B$ and to compare two accidental sequences as:

$$p_A = X(i)Y(j)/L^2$$

$$p_B = 1 - p_A$$

$$I = f_A \ln f_A + f_B \ln f_B - f_A \ln p_A - f_B \ln p_B - L \ln L$$

The number of the event $A$ appearance is $f_A$. The number of the event $B$ appearances is $f_B$. $I$ value is distributed with one degree of freedom when two accidental sequences are compared. The $I$ value may be calculated for sequences with correlation for the nearest bases. The definition of corresponding $k$-chains ($k = 2, 3, \ldots$) is presented below. Let us name $\{D^k(\alpha)\}$ the set of all possible $k$-chains for the events $A$ and $B$. Those $k$-chains are considered as sequences consisting of zeros and ones. Site $i$ is 1 if the base pair corresponds to event $A$, and $i$ is 0 if the base pair corresponds to event $B$. Let us call $\{X^k(\beta)\}$ the set of all possible $k$-chains of the first of compared sequences. Let us name $\{Y^k(\gamma)\}$ the set of all possible $k$-chains of the second of compared sequences. The probability $p(X^k(\beta))$ may be estimated as:

$$p(X^k(\beta)) = N(X^k(\beta))/(L-k+1)$$

The analogous equation is used for $p(Y^k(\gamma))$. Two coinciding $X^k(\beta)$ and $Y^k(\gamma)$ $k$-chains form a corresponding $D^k(\alpha)$ $k$-chain. It is possible to number all $X^k(\beta)$ and $Y^k(\gamma)$ chains and then to calculate the theoretical probability of the $k$-chain $D^k(\alpha)$ as:

$$C(D^k(\alpha)) = \sum_{\beta,\gamma} \{p(X^k(\beta))p(Y^k(\gamma))\}$$

The observed number of $k$-chains $f(D^k(\alpha))$ is calculated using the coincidence sequence. Then the $I_k$ may be cal-
The correlation of the mutual information is the part of the mutual information that is a result of correlation of the neighboring bases in compared sequences. The $F_k$ may be considered as an “evolutionary” part of the mutual information. Calculation of the correlation and evolution parts of the mutual information makes it possible to eliminate the cases of sequence similarity that arise from base correlation in compared sequences and are not conditioned by the common evolution origin of compared sequences.

2.6. Results of application of the enlarged similarity method to DNA sequence analysis

Let us consider the similarity between two MB1 repeats from the human genome. Figure 1 shows the similarity of the MB1 and MIB1 repeats from clones HSHLADC2 (13) and HSHP201 (4). The positions of MB1 and MIB1 repeats are shown. The $2F'$ maximum is reached if AT, CG, TA, GC, and CA coincidences are included in event A, and the other 10 types of coincidences are included in event B. The $2F'$ is equal to 50 for calculations without consideration of base correlation. $2F'$ is equal to 49 for consideration of a three-base correlation. Capital letters in compared sequences show the base coincidences included into the event A and B.

In event A we may include one after another the coincidences that are mostly distinguished from coincidences from accidental sequences with the same A, T, C, and G composition. For example, the second step is to include AT and CA coincidences in event A and to include the other 14 types of coincidences in event B. The third step is to include AT, CA, and CC coincidences in event A and to include the other 13 types of event B. Such analysis could find the certain order of sequence coincidences, that may be linked with the organization of two compared sequences. Let us consider two sequences in example 4.

Example 4.

5'-aGcCaTcTcGaGtCcCtTgCtTcC-3'
5'-gAcTgCgCaAtAgTgTaCgTgC-3'

When GA, TC and CT coincidences are included in A event, then the periodicity of the coincidence sequence could be observed. This periodicity increases the mutual information when the $H(1,2)$ entropy is decreased for $k = 3$.

The calculation of mutual information with different base correlation lengths makes it possible to introduce the definition of the mutual information correlation, that is:

$$S_k = F_1 - F_k$$

Downloaded from https://academic.oup.com/dnaresearch/article-abstract/3/3/157/342343 by guest on 29 July 2018
**Figure 2.** Comparison of base alternations MB1 and MIB1 repeats from HSHLADC2 and HSHP201 clones. A. $S_1$-base alternation MB1 repeat from HSHLADC2 clone, $S_2$-base alternation of MIB1 repeat of HSHP201 clone. B. Representation of base alternations $S_1$ and $S_2$ after excluding 8 numbers. C. Base alternations of alternation identical fragments.

This approach was applied for classification of the tRNA genes from different species. The classification of the tRNA genes shows that evaluation of enlarged similarity between sequences of the tRNA genes made it possible to find the more ancient similarity of tRNA genes than is possible by other theoretical methods. Latent periodicity of many human genes was found by this method also. Most parts of human genes have latent periodicity with lengths ranging from 5 to 120 base pairs. It may be related to the evolutionary origin of genes. The mathematical methods for determination of the DNA sequence complexity are being developed at present. The complexity of (0,1) sequences have been analyzed in try between MB1 and B1 repeats. The members of MB1 family repeats also were found by Donehower et al. These repeats are now known as MIR repeats. The application of the similarity of nucleic acid sequence base alternation analysis made it possible to find very divergent copies of MB1 repeats. A limited version of the enlarged similarity method made it possible to find about two times more MB1 repeats than is possible with the BLAST or FASTA programs. The majority of MB1 repeats are similar to each other in terms of purine and pyrimidine sites only. All members of MB1 family repeats in EMBL data bank can be found only by using enlarged similarity of nucleic acid sequences method.
the works of A. N. Kolmogorov and colleagues.\textsuperscript{24,25} However, it is not a universal method for the determination of the law of DNA sequence construction and information compressing of the sequence for excluding any law is impossible. The limited set of sequence transformation is now used for sequences compression now. The enlarged similarity of DNA sequence analysis and determination of mutual information is the best way to evaluate the complexity of the transformation of one sequence to another.

The method of finding of the enlarged similarity between DNA sequences is still being developed, taking into account the possibility of gaps and insertions in compared DNA sequences. This method is a generalization of the Needleman-Wunsch method\textsuperscript{28} The publication is in preparation now. This method makes it possible to reveal more cases of similarity between DNA sequences because the mutation process includes such changes in base sequences as insertions and gaps.

Further development of the evaluation of transformation complexity of compared sequences is given impetus by the discovery of other laws for the construction of DNA sequences. The development of this investigation is important for further analysis of the human genome and genomes of other species.

References

1. Watson, J. D. 1990, The human genome program at the National Institute of Health, Science, 248, 44-49.
2. Cantor, C. R. 1990, Orchestrating the human genome project, Science, 248, 49-51.
3. Alberts, B., Bray, D., Lewis, J., Raff, M., Roberts, K., and Watson, J. D. 1989, Molecular biology of the cell, Garland Publishing, Inc., New York, London.
4. Bishop, M. J. 1994, Guide to human genome computing, Harcourt Brace and Company, London.
5. Sandor, S. 1994, Computational methods in genome research, Plenum Publishing Corporation, New York.
6. Korotkov, E. V. and Korotkova, M. A. 1986, Analysis of human DNA sequences containing the Alu repeats, Dokl. Akad. Nauk SSSR, 288, 1014-1017.
7. Korotkov, E. V. 1987, Images of human Alu repeats in clones of human genomes, Mol. Biol. (Russian), 21, 478-483.
8. Korotkov, E. V. 1991, New family widespread mirror-reflected MBl repeats in human genome, Mol. Biol. (Russian), 25, 250-263.
9. Korotkov, E. V. 1994, Fast method of purine-pyrimidine relation search between DNA sequences, DNA Sequences, 4, 411-413.
10. Shannon, C. E. 1948, A mathematical theory of communication, Bell System Tech. J., 22, 623-656.
11. Kullback, S. 1959, Information theory and statistics, John Wiley and Sons, Inc., New York.
12. Korotkov, E. V. 1990, Family of mirror B1-like sequences in human genome, Dokl. Akad. Nauk SSSR, 311, 238-242.
13. Owerbach, D., Rich, C., and Taneja, K. 1986, Characterization of 3HLA-DR beta genes isolated from an HLA-DR 3/4 insulin dependent diabetic patient, Immunogenetics, 24, 41-46.
14. Maeda, N., Yahg, F., Barnett, D. R., Bowman, B. H., and Smithies, O. 1984, Duplication within the haptoglobin HP2 gene, Nature, 309, 131-135.
15. Korotkov, E. V. 1992, MBl repeats in some mammal genomes, Izv. Akad. Nauk SSSR, Biological seria, 4, 660-672.
16. Donehower, L. A., Slagle, B. L., Wilde, M., Daglington, G., and Buted, J. S. 1989, Identification of a conserved sequence in the non-coding regions of many human genes, Nucl. Acids Res., 17, 699-722.
17. Smit, A. F. A. and Riggis, A. D. 1995, MIRs are classic tRNA-derived SINES that amplified before the mammalian radiation, Nucl. Acids Res., 23, 98-102.
18. Jurka, J., Zietkiewicz, E., and Labuda, D. 1995, Ubiquitous mammalian-wide interspersed repeats (MIHs) are molecular fossils from the mesozoic era, Nucl. Acids Res., 23, 170-175.
19. Chaley, M. B. and Korotkov, E. V. 1991, Global classification of genes tRNA from many species, Izv. Akad. Nauk SSSR, Biological seria, 6, 915-927.
20. Korotkov, E. V. and Korotkova, M. A. 1995, Latent periodicity of DNA sequences of some human genes. DNA Sequence, 5, 353-358.
21. Korotkov, E. V., Korotkova, M. A., and Tuko, J. S. 1996, Latent sequence periodicity of some oncogenes and DNA binding protein genes, CABIOS, in press.
22. Gusev, V. D., Kulichkov, V. A., and Chupachina, O. M. 1991, The complex analysis of genomes. I. The measure of complexity and the classification of searched structural laws, Mol. Biol. (Russian), 25, 825-834.
23. Gusev, V. D., Kulichkov, V. A., and Chupachina, O. M. 1991, The complex analysis of genomes. II. The regions of gomology in bacteriophage lambda, Mol. Biol. (Russian), 25, 1080-1089.
24. Kolmogorov, A. N. 1987, Information theory and theory of algorithms, Nauka Press, Moscow, pp. 238-250.
25. Zvonkin, A. K. 1973, The complexity of ending objects and the development of information approach with using the algorithm theory, Uspechi mathematical nauk, 25, 83-124.
26. Altshul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. 1990, Basic local alignment search tool. J. Mol. Biol., 215, 403-410.
27. Pearson, W. R. and Lipman, D. J. 1988, Improved tools for biological sequence comparison, Proc. Natl. Acad. Sci. USA, 85, 2444-2448.
28. Needleman, S. B. and Wunch, C. D. 1970, A general method applicable to search for similarities in the amino acids sequences of two proteins, J. Mol. Biol., 48, 443-453.