E. V. Korotkov M. A. Korotkova

012-93

ENLARGED SIMILARITY OF NUCLEIC ACIDS SEQUENCES

Moscow 1993

HIGH EDUCATION GOVERNMENT COMMITTEE OF RUSSIA

MOSCOW PHYSICAL ENGINEERING INSTITUTE

Korotkov E.V. Korotkova M.A.

ENLARGED SIMILARITY OF NUCLEIC ACIDS SEQUENCES

Preprint 012-93

Confirmed by editorial board of institute

Moscow 1993

UDC 577.23

Korotkov E.V. , Korotkova M.A. Enlarged similarity of nucleic acids sequences; M. Preprint / MPEI 012-93 - 1993 - 30c.

The definition of nucleic acids sequence structure is introduced with the help of graph theory methods. The number of structures for the different length of sequences is established. The definition of enlarged similarity of nucleic acids sequences is introduced on the base of sequence structure notion. The mutual information between sequences is used as quantitative measure of structures similarity for two compared sequences. The method of mutual information calculation taking into account the correlation of bases in compared sequence is developed . The definitions of correlated similarity and evolution similarity between compared sequences are given.

1. Introduction.

The developing of the fast methods for nucleic acids sequences determination leads to accumulation of known sequences of different genes and introns and intergenic regions. The execution of human genome program will lead to sequencing of full human genome and genomes of other species (Watson,1990; Cantor,1990). The obtaining results will be used in science and medicine. However the such using is possible after understanding of sense of obtaining information. This problem is similar to the problem of understanding of unknown language text. It is necessary to determine the boarders and the sense of words and than to do this for sentences.

It is obvious that experimental approach gives more synonymous understanding of genetic text. But it is limited by finding of copies of the known fragments only. This approach has the relativity -narrow field of application and may characterize a very small part of defined sequences. The experimental methods find the protein coding regions, introns and small quantity of sequences taking part in genetic activity regulation (Watson et.al., 1983).

The theoretical approach has the more wide field of application. It gives the possibility of the comparison between any DNA sequences.

The well-known statistical methods have found the pairs correlations of bases in DNA sequences, some peculiarities of alternation nucleotides in coding and noncoding regions (Nussinov 1987; Computer analysis of genetic texts,1990).

The search of homology regions has led to discovering the sites in DNA sequences which have homology to each other.

The repeated sequences of different sizes, regions with small number of copies in genome (pseudogenes and some duplication sites) belong to such regions (Okada ,1991; Korotkov,1990). The incomplete homology and homology with deletions between sequences have been found also (Nussinov,1987; Computer analysis...,1990; Nucleic acids and

protein..., 1987). The schemes of evolution conformities of some sequences have been constructed on the base of analysis of homology between sequences. The models of the evolution divergence of DNA sequences have been developed also (Kimura,1983).

However, it is difficult to decide that two sequences are similar to each other if their homology is enough low. If any selected sequences are compared with the set of sequences from big date bank then the probability to find enough low level of homology is very high. It means that the similarity between ancient sequences can't be found by homology search. The some coding regions and some protein-binding sites, genes and repeats from evolutionary distinct species may be such sequences. It shows the necessity of new methods for analysis of genetic information.

The search method of enlarged similarity of nucleic acids sequences is considered in present work. This method uses the definition of DNA sequence structure. It is possible to apply the mathematical methods being developed in this work for analysis of any other texts.

2. The structure approach to the DNA sequences analysis.

The main idea of the structure approach to the DNA sequences analysis is to define a new notion of the sequence structure and to use this notion for elaboration of the method of similar sequences search. We mean that the structure is the type of DNA bases alternation and it is independent of DNA bases. The exact notion is given in the next section. We use unlabeled directed graphs for the structure notion.

We can illustrate the notion of sequence structure as follows. Let there is some storage cells. Each cell contains one symbol and all symbols are different. Let there is a program building the sequence of this symbols by getting the symbols from the cell by address. For example, program may put the symbol from 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 builds different sequences for different cells contents but the symbols alternation are the same for all those sequences.

While we have defined the notion of sequence structure, we may recognize do two sequences have the same or different structures. Sequences have the structures with the common and different parts in common case and the structure correlation is not the coincidence only. Moreover, the same structures are met rare because the number of different structures is large (Table.1). We will find the sequences with similar structures more often than with the equal structures. We can illustrate the obtaining of similar structures by the work of the same program when some cells contain more than one symbol and the choice of the symbol from a cell is arbitrary.

Structure similarity of sequences may be smaller or bigger and we must define the measure of similarity. We choose the mutual information between sequences as the measure of the structure similarity.

The method of search of sequences with structure similarity is following. Most similar sequences having the largest value of mutual information are chosen from the set of all sequences being analyzed. Comparison of the structures of those sequences shows the type of sequences similarity and enables to represent it's similarity visually. This structure comparison helps to find the biological reasons of sequences similarity.

The new class of MB1 repeats was found by structure approach (Korotkov,1990; Korotkov.1991; Korotkov,1992). The subclass of MB1 repeats was found independently early (Donehower et. al., 1989). The new type of mirror symmetry in human genome was found by this method also (Korotkov,1990; Korotkov,1991).

3. The sequence structure and structure similarity of sequences.

The notion of the sequence structure reflects the alternation of the letters in sequence. Let there are two sequences of letters from 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 letters alternation.

Example 1.

5'- ATCGAGCGACGATGA - 3'

5'- GCTAGATAGTAGCAG - 3'

These sequences are equivalent because there is one to one mapping of the first sequence to the second by the law A->G, G->A, T->C, C->T. This type of coincidence between two sequences is named the synonymous mapping one sequence to other. For such coincidence it may be restored each sequence if it is known the other sequence and the law of the mapping. We define that if the sequences have synonymous mapping than they have identical structures.

The convenient representation of these structures may be obtained by the using of graph theory. The bases of sequence are represented by points (nodes of graph). The direction from the beginning of the sequence to the end of it is represented by directed lines (arcs). Those lines link the points: the first with the second, the second with the third and so on. All points corresponding the same bases are linked with the same additional points (nodes) representing the bases types.

The exact definition of sequence structure is following. The structure of the sequence $B = \{b_1, b_2, \ldots, b\}$ (where $b_i \in \{A,T,C,G\}$ for $i \in \{1,\ldots,n\}$) is unlabeled oriented graph $G(B) = \langle V, U \rangle$ (Harary,1969); where V is the set of nodes, U is the set of arcs. $|V| = n+k$, where $k \in \{1,2,3,4\}$ and k is equal to the number of different bases in the sequence. The number of arcs $|U| = 2n - 1$ (see Supplement 1) .

The arcs set for the structure is the next.

1. For all i, $1 \leq i \leq n-1$, there exists the arc (V_i, V_{i+1});

2. For any i , $1 \leq i \leq n$ there exists only one arc (V_i, V_{n+l}), $1 \leq l \leq k$, such that:

a) for all j, $1 \leq j \leq n$, such that $b_i = b_j$ there is the arc (V_j, V_{n+l});

b) for all j, $1 \leq j \leq n$ such that $b_i \neq b_j$ and any arcs (V_j, V_s) and (V_i, V_m) there is m≠s.

3. There are not other arcs in G(B) structure.

Some qualities of the sequence structure are given in Supplement 2. The sequence structure may be built by the next algorithm. The application of this algorithm to concrete sequence is illustrated in Fig. 1.

Sequence structure constructing algorithm.

1. Rases of the sequence are written by vertical and points representing graph nodes are placed near each base (Fig. 1a). Every node represents the nearest base of sequence.

2. The nodes are linked by arcs. Each arc links the node, representing the base, with the node, representing the next base of sequence (Fig. 1b).

3. Four points for nodes representing bases types are drawn and are marked by letters A,T,C,G.

4. The arcs from every node representing the sequence base to the node representing base type with the same name are drawn (Fig. 1c).

5. All labels and the nodes, representing base types, which are not linked with any other nodes are erased (Fig. 1d).

This built graph corresponds to the definition of structure because the positions of nodes and the distance between them are not important.

The nodes representing base types are called type nodes. The number of type nodes in structure is equal to the number of different bases in sequence.

Different numbers of structures exist for different lengths of sequences. The numbers of structures for different lengths of sequences are represented in Table. 1.

It is possible to note that for the structure with one type

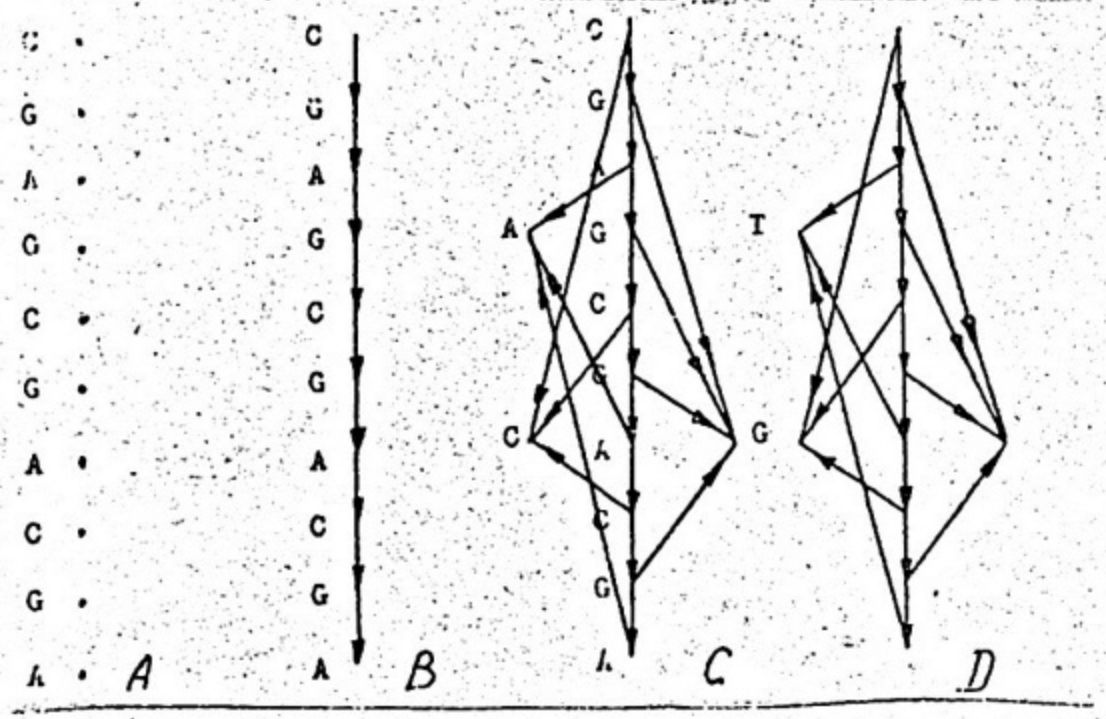

Fig.1. The illustration of the algorithm of constructing of the sequence structure.
A. Sequence and corresponding nodes.
B. Chain which shows the order of the bases.
C. "Prestructure" in which the links with the bases types are shown.
D. Structure of the sequence.

Table 1. The structures numbers for different length of sequence.

Length of sequence	Number of sequences	Total structures number	Structures number for different bases number in sequence		
			Number of different bases	Possible number of structures	Number of sequences for one structure
1	4	1	1	1	4
2	16	2	1	1	4
			2	1	12
3	64	5	1	1	4
			2	3	12
			3	1	24
4	256	15	1	1	4
			2	7	12
			3	6	24
			4	1	24
n	4^n	$\frac{4^{n-1}+2}{6}+2^{n-2}$	1	1	4
			2	$2^{n-1}-1$	12
			3	$\frac{3^{n-1}+1}{2}-2^{n-1}$	24
			4	$\frac{4^{n-1}-3^n-1}{6}+2^{n-2}$	24

node here is 4 different sequences having this structure. For every structure with 2 type nodes which corresponding to the sequences with 2 different bases types there are 12 sequences with this structure. For the structure with 3 or 4 type nodes here are 24 sequences with this structure. It means that for every sequence which is built of one type of bases here exist 3 other sequences with the same structure. For the sequence being built of 2 bases types here are 11 other sequences with the same structure. For the sequence with 3 or 4 different bases types there are 23 other sequences with the same structure. The number of sequences with the same structure depends on the number of bases types in sequence and does not depend on the sequence length.

Fig. 2 represents the sequence structures for small lengths and the building of new structures for increasing sequence length.

Different intuitively similar sequences have the same structures. So, the structure corresponding to two sequences in example 1 is presented in Fig. 3a. Insignificantly different sequences have insignificantly different structures. For example, if the second sequence is obtained from the first one by changing only one base, than the structures of this two sequences are differed by one arc only. The structures for such two sequences are presented in Fig. 3b. There the first sequence is the first sequence of Example 1, and the second sequence is obtained from the first by changing one base. Fig. 3c compares the structures of the first sequence of example 1 and of the sequence being obtained from it by deletion of one base.

Two bases sequences $A = \{a_1, a_2, \ldots, a_n\}$ with the structure $G(A)$ and $B = \{b_1, b_2, \ldots, b_n\}$ with the structure $G(B)$ are called the structure identical if the structures $G(A)$ and $G(B)$ are isomorphic.

If there is isomorphism between two structures then they may be drawn on place as identical if the same order of graph

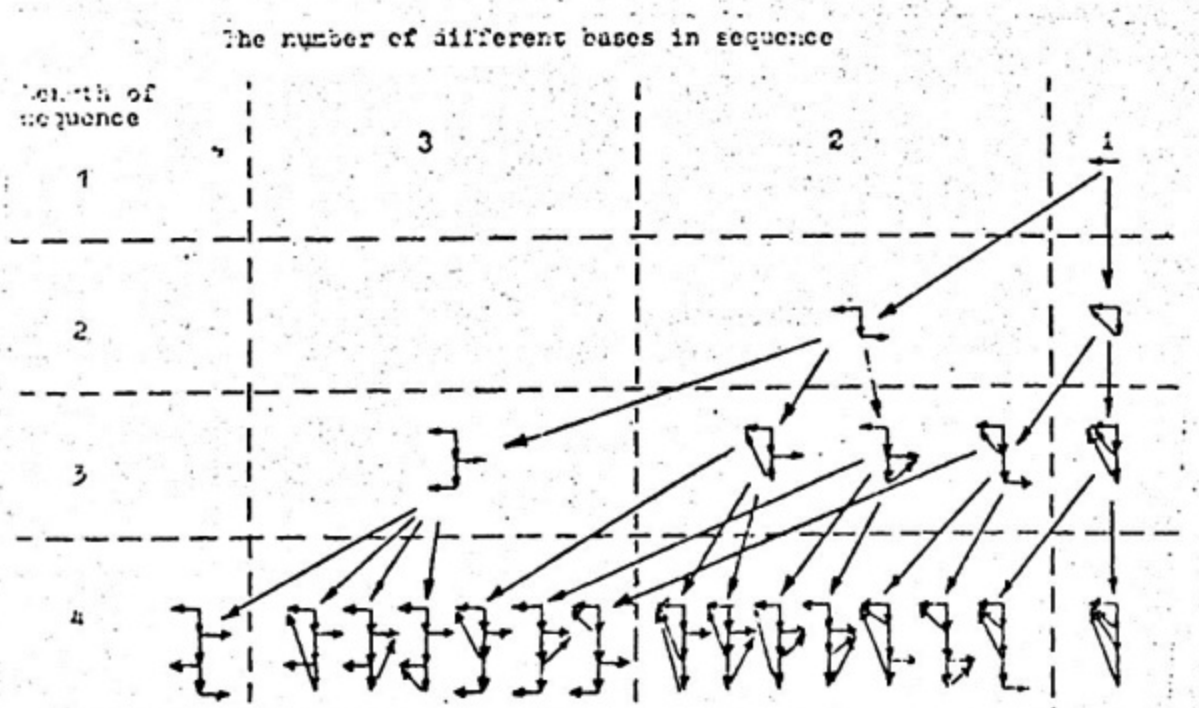

Fig.2. The structures of the sequences with length 1,2,3 and 4 bases and ways of their formation.

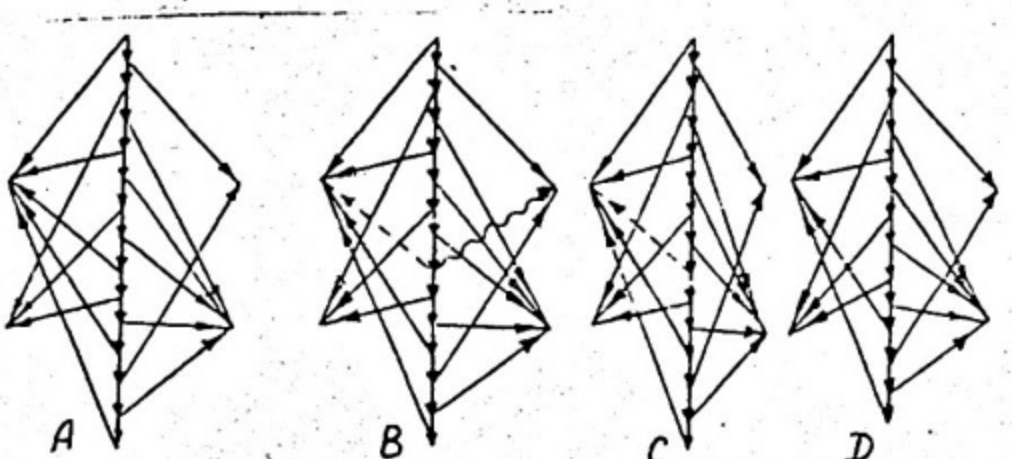

Fig.3. The structures of the sequences.
A. For sequences of the example 1:
...ATCGAGCGACGATGA...
...GCTAGATAGTAGCAG...
B. Comparison of the structures of sequences 1 and 2. The common arcs are unbroken. The arcs belonging to structure of sequence 1 only are shown as dotted lines and arcs belonging to structure of sequence 2 only are shown as wavy lines.
1- ATCGAGCGACGATGA
2- ATCGAGCGTCGATGA
C,D. The comparison of the structures of sequence 3 and sequence 4 (with deletion). The dotted lines show the arcs corresponding to deleted base.
3- ATCGAGCGACGATGA
4- ATCGAGCGCGATGA

drawing is applied. For example we propose the next manner of graph drawing. At the third algorithm step first type node corresponding the first base is placed to the left from the sequence. Second type of sequence bases equals to the second base of sequence if it is differed from the first or to n+1 base if first n bases of sequence are equal. The second type node corresponding to the second type of sequence bases is placed to the right from the sequence, symmetrically to the first type node. The third type node corresponding to the third type of sequence bases is placed under the first type node. The fourth type node is placed symmetrically to the third type node under the second one.

Structure identity is special case of structure similarity. While we have the definition of structure identity, we can determine are structures of two sequences identical or not. However, the comparison of two sequences which are not structure identical is more difficult task. Two sequences may be considered as structure similar in two cases. In the first case sequences structures have the most arcs in the same directions and a few arcs are differed in directions. The structure similarity in the second case is obtained by regular deflection from structure identity. For example for the last case, all arcs being directed to one type node in the first structure are directed to some two type nodes in the second structure. The structure similarity in the second case when it is obtained by regular deflection from structure identity we can explain by the next example.

Suppose we have four sequences $A_1 = \{a_1,\ldots,a_n\}$, $B_1 = \{b_1,\ldots,b_n\}$, $A_2 = \{a_{n+1},\ldots,a_{n+m}\}$, $B = \{b_{n+1},\ldots,b_{n+m}\}$. Sequences A_1 and B_1 are structure identical $G(A_1) = G(B_1) = G_1$, and sequences A_2 and B_2 are structure identical $G(A_2) = G(B_2) = G_2$ and the number of type nodes in G_i are no less than 3. We build the sequence $A = A_1 \cdot A_2 = \{a_1,\ldots,a_n,a_{n+1},\ldots a_{n+m}\}$ with the structure G(A) and the sequence $B = B_1 \cdot B_2 = \{b_1,\ldots,b_n,b_{n+1},\ldots,b_{n+m}\}$ with the structure G(B). In

the common case structures G(A) and G(B) are different as there is 24 manner of coincidences of the type nodes in structures G_1 and G_2, and the coincidences may be different in G(A) and G(B).

Suppose that the sequences A_1, A_2, B_1, B_2 are such as it is described above. Let the sequence A is built of the sequences A_1 and A_2, and sequence B is built by the same manner of the sequences B_1 and B_2. Than sequences A and B are not structure isomorphic in 23 cases of 24. But the sequences A and B are structure similar as A and B may be divided to structure identical sequences.

The exact definition of division of pair of sequences A and B being compared reflects the follows. First, each base of the sequences A and B enters exactly one sequence being built. Second, the order of bases in each subsequence including in sequences being built, is conservative. Third, both sequences A and B are divided by the same manner.

Let the sequence of indexes $\{1,2 \ldots,n\}$ is divided to two subsequences I and J, $I = \{i_1,\ldots. i_k\}$, $i_l < i_{l+1}$ for $1 \leqslant l \leqslant k-1$ and $J = \{j_1,\ldots,j_{n-k}\}$, $j_m < j_{m+1}$ for $1 \leqslant m \leqslant n-k-1$, such that $I \cap J = \emptyset$ and $\{i_1,\ldots,i_k\} \cup \{j_1,\ldots,j_{n-k}\} = \{1,\ldots,n\}$. Then for sequence pair $\langle A,B\rangle$, $A = \{a_1,\ldots,a_n\}$, $B = \{b_1,\ldots, b_n\}$ the set of two pairs $\{\langle A_1, B_1\rangle, \langle A_2, B_2\rangle\}$, where $A_1 = \{a_{i_1},\ldots,a_{i_k}\}$, $A_2 = \{a_{j_1},\ldots,a_{j_{n-k}}\}$, $B_1 = \{b_{i_1},\ldots,b_{i_k}\}$, $B_2 = \{b_{j_1},\ldots,b_{j_{n-k}}\}$ is the division of pair of sequences $\langle A, B\rangle$.

Pair of sequences may be divided to arbitrary number of sequence pairs. If the set of pairs $\{\langle A_1,B_1\rangle \ldots. \langle A_m,B_m\rangle\}$ $m > 1$ is the division of pair of sequences $\langle A,B\rangle$, and for some j, $1 \leqslant j \leqslant m$ the set $\{\langle A_{m+1},B_{m+1}\rangle, \langle A_{m+2},B_{m+2}\rangle\}$ is the division of pair $\langle A_j,B_j\rangle$, then the set of pairs $\{\langle A_1,B_1\rangle, \ldots, \langle A_{j-1},B_{j-1}\rangle, \langle A_{j+1},B_{j+1}\rangle,\ldots,\langle A_{m+1},B_{m+1}\rangle, \langle A_{m+2},B_{m+2}\rangle\}$ is the division of $\langle A,B\rangle$ pair.

Two sequences $A = \{a_1,\ldots a_n\}$ with the number of base types

$k_A \geq 2$ and $B = \{b_1,\ldots b_n\}$ with the number of base types $k_B \geq 2$ we call totally structure similar, if they are not structure identical and there exists the division of pair $\langle A,B\rangle$ to the set of pairs $\{\langle A_1,B_1\rangle,\ldots\langle A_m,B_m\rangle\}$, such that $m \leq \max\{k_A,k_B\}-1$ and A_i is structure identical to B_i for $1 \leq i \leq m$. Total structure similarity shows that two sequences may be divided to smaller number of structure identical pairs then it may be done for arbitrary pair of sequences.

We call the sequences A and B structure similar to the degree of similarity α, if there exists the division $\{\langle A_1,B_1\rangle, \langle A_2,B_2\rangle\}$ of $\langle A,B\rangle$ pair such that A_1 and B_1 are structure identical or structure similar and A_1 contains no less then α% bases of the sequences A. Let us note that there exists the division $\{\langle A_1,B_1\rangle, \langle A_2,B_2\rangle, \langle A_3,B_3\rangle, \langle A_4, B_4\rangle\}$ of an arbitrary sequence pair $\langle A,B\rangle$, $A=\{a_1,\ldots,a_n\}$, $B=\{b_1,\ldots,b_n\}$ such that A_i and B_i are structure identical for each i, $1 \leq i \leq 4$. But it is not mean the total structure similarity of this sequences. The examples of structure similar sequences and their structure similarity are shown in Fig. 5-8.

We can determine are the two sequences structure similar or not. But structure similarity of sequences may be smaller or greater. It is clear, that two structure identical sequences have more likeness in its structure than two total structure similar sequences. Or, for example, there is two pairs of sequences, $\langle A,B\rangle$ and $\langle C,D\rangle$ and A is structure similar to B, C is structure similar to D. If there exists a division $\{\langle A_1,B_1\rangle, \langle A_2,B_2\rangle\}$ such that A_i and B_i are structure identical for $i \in \{1,2\}$, but there is not division of $\langle C,D\rangle$ to two structure identical pairs, than A and B are more structure similar than C and D. So, we must have measure of structure similarity.

We choose the mutual information as the measure of structure similarity. If the mutual information of sequences A and B is greater than it for sequences C and D we

call A and B more similar than C and D. We can assert than two sequences are much similar or little similar.

Thus, in this section we had defined the notions of sequence structure and structure similarity. First, those notions allow for every two sequences to determine their similarity and to show similar subsequences and the kind of likeness. Second, the notions allow to draw this similarity visually. Third, the notions allow to find the structure identical subsequences with the same law of base alternation. Fourth, the notions allow to determine how two sequences may be divided to structure identical subsequences by the best manner (see Supplement 3).

4. The measure of nucleic acids sequences structure similarity.

The mutual information between two compared sequences is taken as the measure of their similarity (Korotkov ,1986; Korotkov ,1987, Korotkov,1991). The mutual information is calculated with the help of the matrix of bases coincidences between compared sequences. The number of each type of coincidence in matrix with dimension 4x4 is calculated when two sequences with L length are compared. The main diagonal elements show the number of AA, TT, CC and GG coincidences. The sum of all 16 elements of the matrix is equal to the length of compared sequences. The mutual information is calculated as:

$$I(1,2) = (H(1)+H(2)-H(1,2))L\ln 2 \quad . \qquad (1)$$

H(1) is the middle entropy of the first sequence for one base. H(2) is the middle entropy of the second sequence for one base. H(1,2) is the middle entropy of the "coincidences sequence" for one type of coincidences.

The "coincidences sequence" is the sequence which has the 16 letters alphabet. The bases pairs (first base from the first compared sequence and the second base from the second compared sequence) are the letters of coincidences sequence.

The middle entropy in any sequence for one letter is calculated as (Shannon,1948):

$$H = -\sum_{i=1}^{m} p_i \log p_i \quad . \qquad (2)$$

The size of the used alphabet is m. The middle probability of i letter in sequence is p_i. For DNA sequences m equals 4 and p equals the numbers of A, T, C and G bases in sequences divided by the length of compared sequences. It is supposed that the sequences of equal lengths are compared. For coincidences sequence m equals 16 and p_i are the probabilities of each of 16 letters.

The middle value of 2I(1,2) is equal 9. 2I(1,2) is distributed as χ^2 with 9 degrees of freedom when accidental sequences with 4 letters alphabet are compared (Kullback,1959).

All pairs of sequences possessing the similar structure have the same and maximal for given type of structure (with one, two, three or four bases) and sequences length mutual information. The mutual information as measure of sequences similarity allows to determine the degree of sequences similarity when the sequences structures similarity is imperfect. There is the full analogy with imperfect homology of sequences. The mutual information here is always less than it is for perfect structures similarity of sequences.

The mutual information allows to determine the degree of structure similarity for two sequences when compared sequences are not synonymous reflection of each other. If the one sequence and coincidences matrix are known than it is impossible to restore the other sequences here. The example of such complex relation between sequences is the coincidences of the sites of purines and pyrimidines between two compared sequences. Such relation between sequences is represented in their structures by the next way. The structures coincide after the joining in each structure the two nodes

corresponding the purines and two nodes corresponding the pyrimidines.

5. The bases correlation calculation in compared sequences.

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

$$H(1) = -\sum_{\alpha} (p(B_\alpha^k) \log_2 p(B_\alpha^k))/k \qquad (3)$$

$$H(2) = -\sum_{\beta} (p(B_\beta^k) \log_2 (B_\beta^k))/k \qquad (4)$$

$$H(1,2) = -\sum_{\gamma} (p(B_\gamma^k) \log_2 (B_\gamma^k))/k \qquad (5)$$

Where B_α^k, B_β^k and B_γ^k are the k-long chains in the first compared sequence, second compared sequence and in coincidence sequence with number α, β and γ. If the first compared sequence has m_1 letters alphabet, the second compared sequence has m_2 letters alphabet then the coincidence sequence has $m_1 \cdot m_2$ letters alphabet. The number of k-long chains in the first compared sequence equals to m_1^k, in the second compared sequence it equals to m_2^k and in the coincidence sequence the number of k-chains equals to $(m_1 \cdot m_2)^k$. For the DNA sequences $m_1 = m_2 = 4$. It is possible to calculate the number of being contained k-long chains $n(B_\alpha^k)$ for any sequence with length L (L >> k) where $\alpha = 1,2,\ldots m^k$. Then it possible to calculate the $p(B_\alpha^k)$ probability as:

$$p(B_\alpha^k) = n(B_\alpha^k)/(L-k+1) \qquad (6)$$

So, if two compared sequences have the enough length than base correlation for any length may be taken into consideration in I(1,2).

The conditional mutual informations F_2 , F_3 ,..... F_k correspond to absolute mutual informations I_2 (1,2), I_3 (1,2), ..., I_k (1,2).

$$F_k = kI_k(1,2) - (k-1)I_{k-1}(1,2) \qquad (7)$$

The mutual information F_{k+1} is 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} equals 0. If two accidental sequences are compared then $F_2 = F_3 = \ldots = F_k$.

The I (1,2) calculation is possible for compared sequences length more than 10^3 only because the coincidence sequence has 16 letters alphabet. It gives 256 pairs combinations of those letters. For sequences with lesser length it is possible the application of another two methods.

A. Analysis of bases coincidences which are distinguished from accidental bases coincidences.

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

$$N(i,j) = X(i)Y(j)/L^2 \qquad (8)$$

X(i) are A,T,C, and G number in the first sequence. Y(j) are A,T,C, and G number in the second sequence. L is the length of sequences.

Matrix N shows the numbers of coincidences of each type of pairs when two accidental sequences with given A,T,C and G numbers are compared. It is possible to determine the type of coincidence which is distinguished more from accidental coincidence when M and N matrixes are known:

$$f_A = \max \{ M(i,j) - N(i,j) \} \qquad (9)$$

It is possible to consider the two events when two sequences are compared. A is event which includes the bases coincidences being distinguished more from bases coincidences

for accidental sequences (with the same A,T,C and G composition). B is the event which includes the another types of bases coincidences. Event A may include the one from 16 possible bases coincidences. For example it may be AT. Then event B includes the another 15 types of bases coincidences, 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 middle probabilities of events A and B for comparison of two accidental sequences as:

$$e_A = X(i)Y(j)/L^2 \qquad (10)$$

$$e_B = 1 - e_A \qquad (11)$$

$$I' = f_A \ln f_A + f_B \ln f_B - f_A \ln e_A - f_B \ln e_B - L \ln L \qquad (12)$$

The number of event A appearance is f_A. The number of event B appearance is f_B. The 2I value is distributed with one degree of a freedom when two accidental sequences are compared. The I value may be calculated for sequences with correlation of the nearest bases. The definition of corresponding k-chains (k = 2,3,....) is required here. Let call $\{ D_\alpha^k \}$ the set of all possible k-chains for events A and B. Those k-chains are considered as sequences consist of zeros and ones. Site i is 1 if the bases pair corresponds to event A and i is 0 if the bases pair corresponds to event B.

Let call $\{X_\beta^k\}$ the set of all possible k-chains of the first compared sequence. Let call $\{Y_\gamma^k\}$ the set of all possible k-chains of the second compared sequence. The probability $p(X_\beta^k)$ may be calculated as:

$$p(X_\beta^k) = N(X_\beta^k)/(L-k+1) \qquad (13)$$

Analogous equation is used for $p(Y_\gamma^k)$. Two coinciding X_β^k and Y_γ^k k-chains form corresponding D_α^k k-chain. It is possible to number all X and Y chains and than to calculate the theoretical probability of the k-chain D_α^k as:

$$C(D_\alpha^k) = \sum_{\beta,\gamma\to\alpha} p(X_\beta^k)p(Y_\gamma^k) \qquad (14)$$

The observed number of k-chains $f(D_\alpha^k)$ is calculated using the coincidences sequence. Than the I_k' may be calculated as (Kullback,1959):

$$I_k' = \{\sum_\alpha f(D_\alpha^k)\ln f(D_\alpha^k) - \sum_\alpha^\nu f(D_\alpha^k)\ln C(D_\alpha^k) - (L-k)\ln(L-k) \qquad (15)$$

Such analysis of compared sequences may be done for considerable bases correlation length. It is required to determine for k=2 the probabilities of 4 chains only. For k=3 it is required to determine the probabilities of 8 chains only and so on. If the length of compared sequences is some hundred bases then such analysis may be done for bases correlation length which is equal to 6.

It is possible to introduce the conditional mutual informations F_k'. They are calculated using the absolute mutual informations I_k and the formula /7/. But the coincidences sequence is constructed using the definitions of A and B events.

In the event A there may be included one after another the coincidences which are most distinguished from coincidences for 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 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 other 13 types in event B.

Such analysis may find the certain structural law of sequences coincidences which may be linked with the organization of the two compared sequences. Let consider two sequences for example:

5'- A [G] C [C] A [T] C [T] C [G] A [G] T [C] C [C] T [T] G [C] T [T] C A - 3'
5'- G [A] C [T] G [C] G [C] A [A] T [A] G [T] G [T] A [C] G [T] G [C] G T - 3'

When GA,TC and CT coincidences are included in A event then there is periodicity of coincidences sequence. This periodicity increases the mutual information as the H(1,2) entropy is decreased for k-3.

Such analysis gives the possibility to determine the most characteristic coincidences types which are observed between two compared sequences. The comparison of the MB1 repeats to each other by this analysis has shown that MB1 repeats have similarity to each other when event A includes the some number of purine-pyrimidine coincidences.

The calculation of mutual information with the different bases correlation length allows to introduce the definition of correlation mutual information which is:

$$S_k = F_1 - F_k \qquad (16)$$

The correlation mutual information is the part of mutual information which is result of correlation of the neighboring bases in compared sequences. The F_k may be considered as "evolution" part of mutual information. The calculation of the correlation and evolution mutual informations allows to eliminate the cases of sequences similarity which arise from bases correlation in compared sequences and are not conditioned by the common evolution origin of compared sequences.

B. Introducing the new alphabet in compared sequences.

The calculation of the mutual information between compared sequences of small length may be done with using matrix of the less volume. This matrix is formed for the more simple alphabet. Such analysis allows to execute of the sequences comparison enough fast but it narrows of the sequences similarity class. The letters joining are the simple way of the new alphabet creation. For example, there are three ways of letters joining in pairs. According of this there are 9 matrixes types which may be constructed when two sequences are compared. Each matrix contains the information which is

partially contained in another matrixes.

There is the biological sense of the letters confluences in compared sequences. For example, A may primary be changed for A->G, G->A, T->C, C->T. It is class of substitutions which save the purine and pyrimidine sites. Comparison of such two very divergent sequences gives the low level of homology which is not statistically important. But those sequences have the 100% coincidences in purines and pyrimidines sites. The matrix with the dimension 2x2 is more suitable for search of such sequences similarity. The lines sings of the such matrix are A={A,G} and T={T,C} in the first compared sequence. The columns signs are A={A,G} and T={T,C} in the second compared sequence. The calculation of mutual information is executed by formula (1). The correlation of neighboring bases may be taken into consideration for mutual information by formula 5 (matrix dimension 2x2) for length of the compared sequences less than 100 bases . Such comparison of some human genome sequences have been done and the function of distribution of $2F_2$ value is well agreed with the distribution χ^2 with one degree of freedom (Korotkov ,1992).

6. Illustration of the structural approach application to DNA sequences analysis.

Let consider the similarity between two MB1 repeats from human genome , between t-RNA genes Asp and different repeats from human and gorilla genomes as examples of enlarged similarity between DNA sequences. The Fig 4a shows the similarity of the MB1 and MIB1 repeats from clones HSHLADC2 (Owerbach et al., 1986) and HSHP201 (Maeda et. al., 1984). The coordinates of MB1 and MIB1 repeats are shown near the sequences. The 2F' maximum is reached if AT,CG,TA,GC,GT and CA coincidences are included in the event A and the other 10 coincidences types are included in the event B. The $2F_1'$ equals to 50 for calculations without consideration of the bases correlation. $2F_2'$ equals to 49 for consideration of the bases

A
1112-CACTTtGCtG AtAaGGAaaC TGAGGCaCAG acAGGTTGAG TAtCTTGCCC AaAtTcAgGc AgCCtGTAAG
1256-GTAGAcTAcT TgTcTTTcgG ACTCTGgGTT acTCCAATTC ATcAAACGGA TcTcAcTgTc TaGAtAATTT
AGGCaGAGtC aGGATTtGAa CCCt-1205
TCCGcCTCgA cTCTAAcCTg AGGc-1163

B
1-GggATTGTAg TTCAATTGgT CAgAGCACCG CCcTGTCaAG gCGGAAgCTG CGGGTTcGAG CCCCGTC-67
180-AgcTGGAGGt TGCTGTGAgC TGtGATGCCA CTgCACTcTA cCAAGGgTGA CAAAGTgAGA CTCTATC-247

C
74-GCCCTGActG cCCCGAGcTT GGGcGTcGAA GGcGGAACtg TCCcGCCacG AGActGGttA-15
103-ATTTCAAgcG cTTTGAGgCC AAAgGCaGAA AAgGAAATat CTTcGTTtaA AAActAGoCA-163

Fig.4. Enlarged similarity of the sequences. The bases coincidences with are included in event A are shown by capital letters.

A. MB1 and MIB1 repeats from HSHLADC2 (25) and HSHP201 (26) clones.

B. T-RNA gene from CHNTRN3 (27) and Alu repeat from GCREG20 clone (28).

C. T-RNA gene Asp. from CHNTRN3 clone (27) and repeated human sequence from HSALPA clone (29).

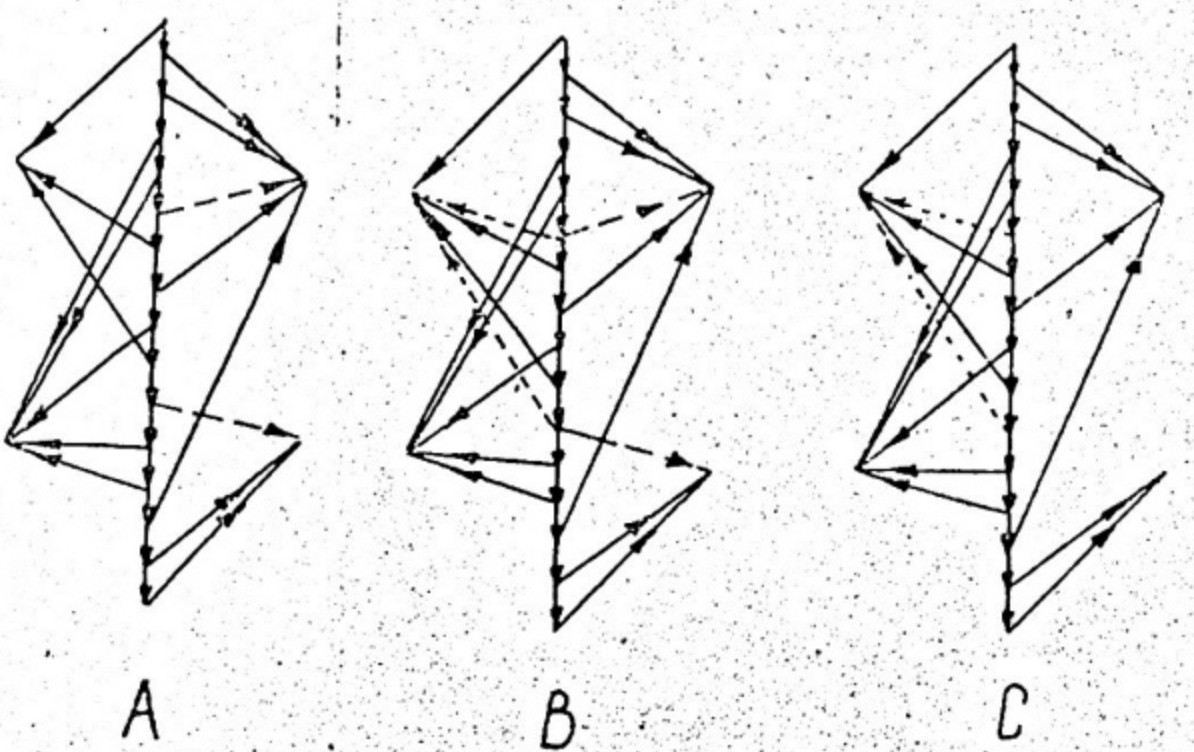

Fig.5. The comparison of the structures of sequences.
AGGTTGAGTACTTGCC
TCCAATTCATAAACGG
The first sequence is begun from 1144 base of HSHLPC2 clone and the second sequence is begun from 1224 base HSHP201 clone.

A. The structure of first sequence.

B. The coincidence of structures.

C. The structure of second sequence.

pairs correlation. $2F'_3$ equals 49 for consideration of the three bases correlation. The titles letters in compared sequences show the bases coincidences including in event A.

The Fig. 5 shows the structures comparison of repeats fragments which are represented in Fig. 4a. Because the graphical representation of structures on plane for long sequences is not clear the structures are compared for fragments only. The structures of being represented fragments with length of 16 bases are similar and are differed in two positions only.

The Fig. 4b shows the similarity of tobacco chloroplast t-RNA Asp from CHNTRN3 clone (Ohme et. al., 1985) and gorilla Alu repeats from GCREG20 clone (Daniels & Deininger, 1983). The 2F' maximum is reached when GA, CC, AG, TG, CT, TT, TC and AT coincidences are included in event A and other coincidences are included in event B. The $2F'_k$ equals 46, 46, 48 for k=1, 2, and 3. It is impossible to find this similarity between DNA sequences by methods of the homology search. Obtained result is agreed with early observed cases of homology t-RNA genes and different repeats sequences (Sacamoto & Okada, 1985; Okada, 1991). Fig. 6 represents the structures of those sequences fragments and their division to isomorphic structures. Fig. 7 shows the structures of other DNA fragments which are divided to two pairs of structures. The structures in first pair are isomorphic but in second pair they are similar only. Division of the second pair to two isomorphic pairs is clear. The number of structure identical pairs equals 3 and the total similarity condition is executed.

Fig. 4c shows the mirror similarity (Korotkov, 1991) between chloroplast t-RNA for Asp amino acid of CHNTRN3 clone (Ohme et. al., 1985) and repeated sequence from human genome of HSALPA clone (Weiss et, al, 1983). The maximum of mutual information is observed when coincidences CT, TC, AA, GG and GA are included in event A and other 11 coincidences are included

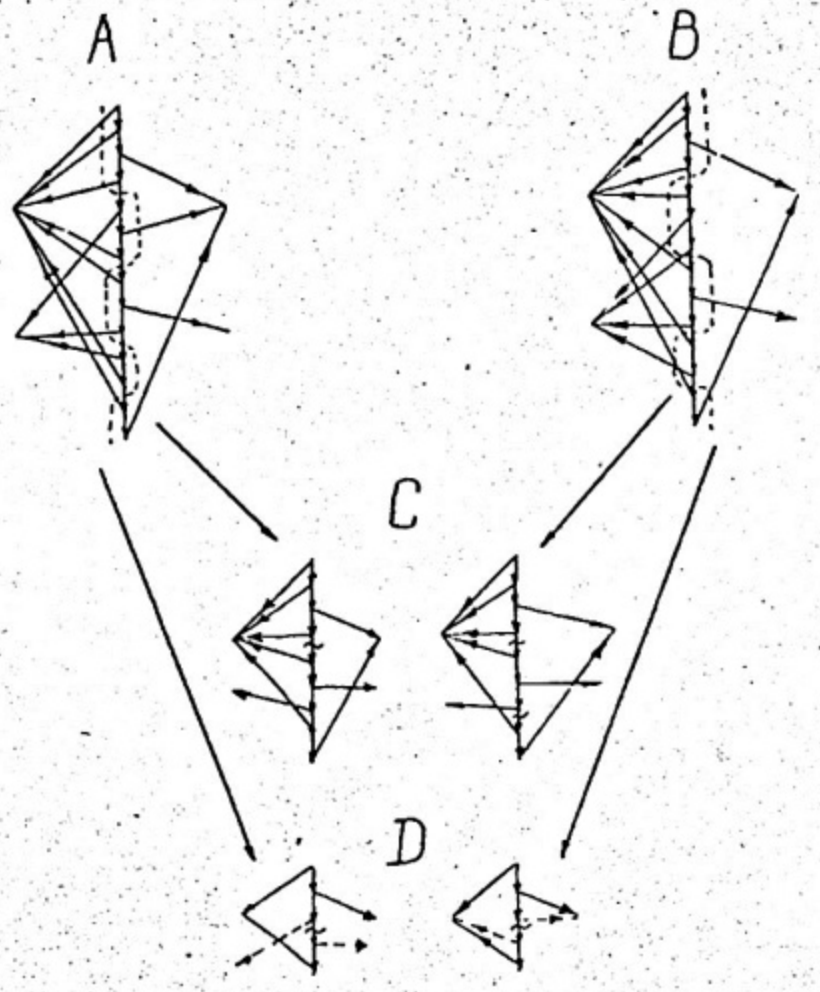

Fig.6. The comparison of the structures of sequences.
TTGTAGTTCAATTG
GGAGGTTGCTGTGA
The first sequence is the fragment of t-RNA gene Asp from CHNTRN3 clone from 5 base and the second sequence is the fragment Alu repeat from GCGER20 clone from 184 base.

A. The structure of the fragment from CHNTRN3 clone.
B. The structure of the fragment from GCGER20 clone.
C. The structures of the first pair of sequences in the division.
D. The structures of the second pair of sequences in the division.

The division of sequences is shown in figures A and B as dotted lines and in figures C and D as wavy lines.

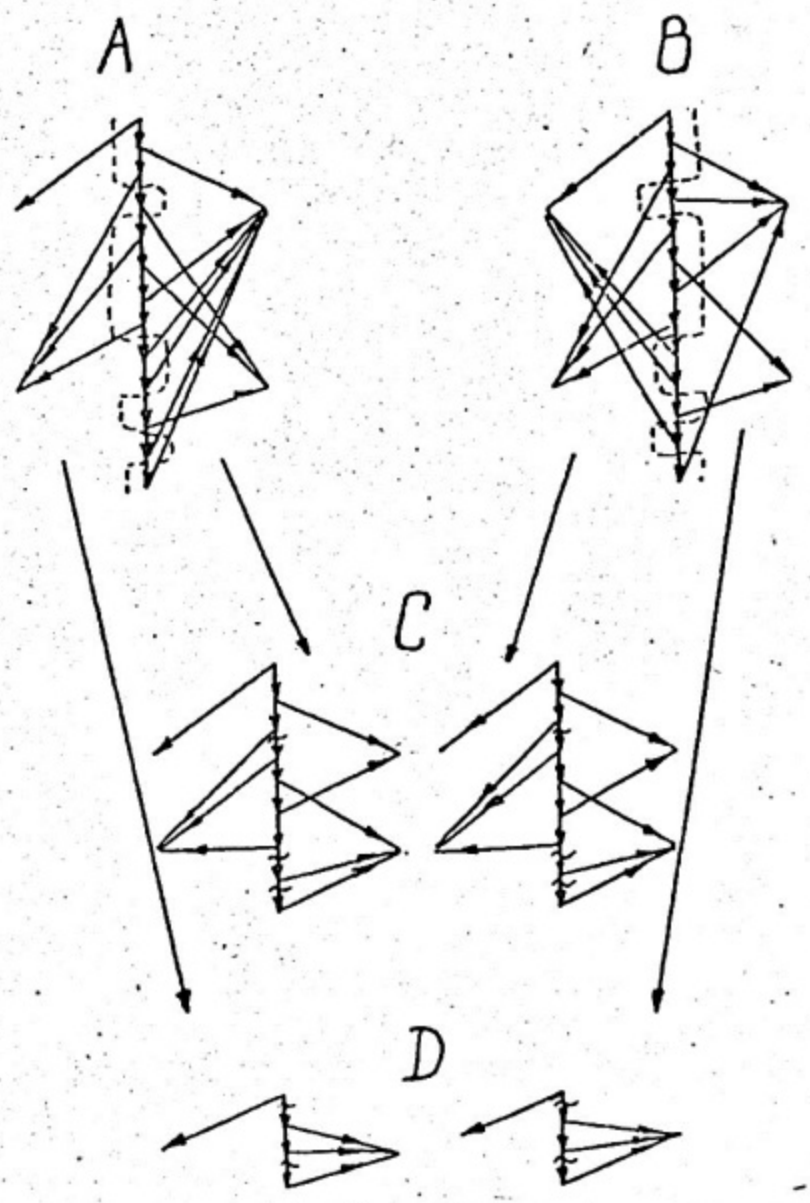

Fig.7. The comparison of the structures of sequences. TCAGAGCACCGAA - fragment Asp from CHNTRN3 clone which is begun from 19 base. CTGTGATGCCACT - fragment of Alu repeat from GCGER20 clone which is begun from 198 base.

A. The structure of the fragment from CHNTRN3 clone.
B. The structure of the fragment from GCGER20 clone.
C. The structure of the first pair of sequences in the division.
D. The structures of the second pair of sequences in the division. The division of the sequences is shown in figures A and B as dotted lines and in figures C and D as wavy lines.

in event B . The $2F_1'$, $2F_2'$ and $2F_3'$ values are 39, 38 and 45. The structures of two fragments of those sequences are compared and the isomorphic structures of their division are shown in Fig. 8.

7. The results of enlarged sequences similarity using and perspective for investigation.

The described in the present work methods of DNA sequences analysis are realized as Fortran programs complex. The analysis of EMBL date bank by method of enlarged similarity search of nucleic acids sequences has discovered the new MB1 family repeats in human and other genomes and the new type of mirror symmetry between DNA sequences. The MB1 family repeats was found in (Dohenover et. al., 1989) independently with less number copies in the family. The application of the enlarged similarity of nucleic acid sequences allowed to find the full number copies of MB1 repeats (about 3×10^5 copies in human genome).

The MB1 family repeats is common for very many mammals (Korotkov, 1992). The life time of MB1 repeats is more than 10^8 years (Korotkov , 1992). The MB1 family members are very diverted from each other and the most MB1 repeats are similar to each other by purine and pyrimidine sites only.

This approach has applied for classification of the t-RNA genes from different species also (Chaley & Korotkov, 1991). This classification of the t-RNA genes shows that consideration of enlarged similarity between sequences of the t-RNA genes allowed to find the more ancient similarity of tRNA genes than it is possible to do by the other theoretical methods now.

The mathematical methods for determination of the DNA sequences complexity are developing now (Gusev, 1991a; Gusev 1991b). Complexity of (0,1) sequences have been analyzed in works of A. N. Kolmogorov and disciples (Kolmogorov, 1987;

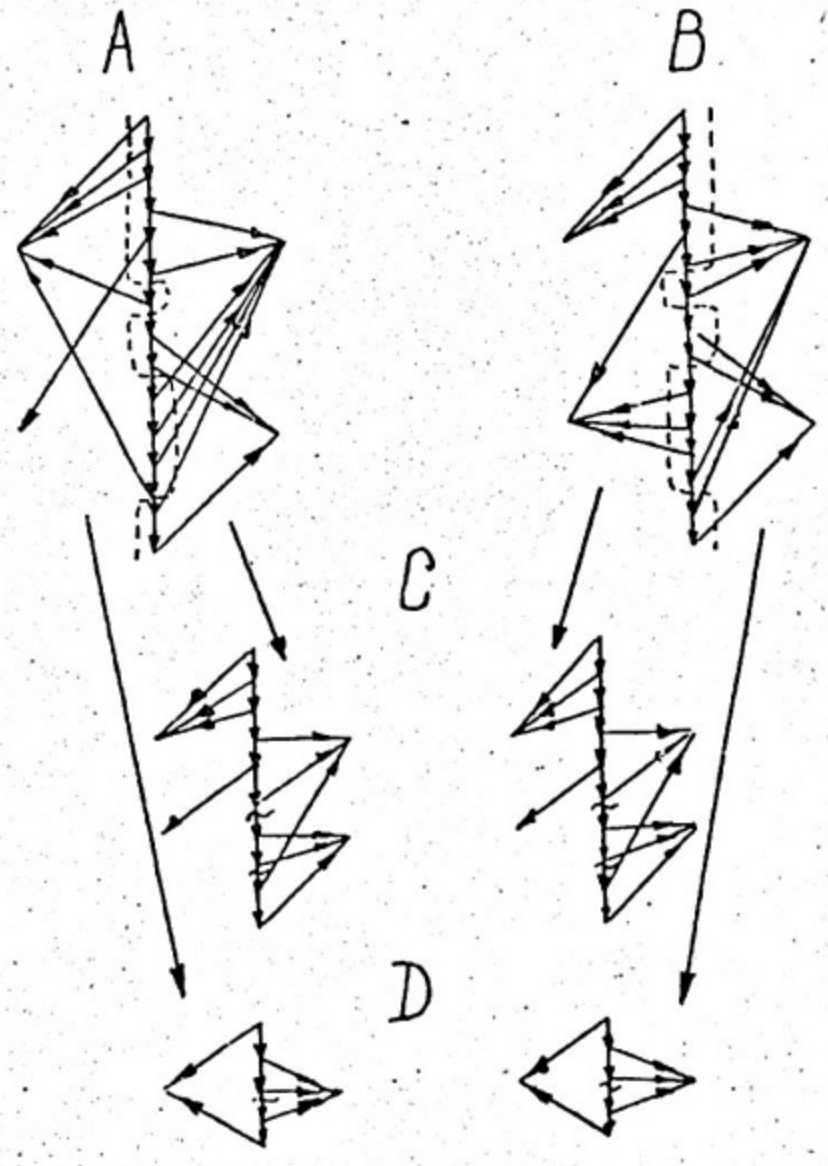

Fig 8. The comparison of the sequences structures. CCCGAGCTTGGGCGT - fragment of tobacco t-RNA Asp gene from CHNTRN3 clone from 73 base. TTGAGGCCAAAGGC - fragment of repeated sequence from HSALPA clone from 104 base.

A. The structure of the fragment from CHNTRN3 clone.
B. The structure of the fragment from HSALPA clone.
C. The structures of the first pair of sequences in the division.
D. The structures of the second pair of sequences in the division. The division of the sequences is shown in figures A and B as dotted lines and in figures C and D as wavy lines.

Zvonkin, 1973). However, there is not the universal method for determination of any law of the sequence construction and compressing of the sequence for excluding any law is impossible. The limited set of sequence transformation is used for sequences compression now. The enlarged similarity of DNA sequence and determination of mutual information is the way for appraisal of complexity of transformation one sequence to another sequence.

The further developing appraisal of transformation complexity of compared sequences is supposed by discovering the other constructed law in compared sequences. The developing of such investigation is important for analysis of genome sequences of human and other species.

References.

Cantor C.R. (1990). Science. V.248, 49-51

Chaley M.B. & Korotkov E.V.(1991). Izvestia Akad. Nauk SSSR. Seria of biology. N.6, 915-927.

Christofides N.N.(1975)."Graph theory. Algorithmic approach" Academic press, New-York, part 12, item 4.12 .

Computer analysis of genetic texts. (1990). M.D.Frank-Kamenetskii eds. Moscow. Nauka.

Daniels R.S. & Deininger P.I. (1983). Nucl. Acids Res. V.11, 7595-7610.

Donehower L.A., Slagle B.L., Wilde M., Daglington G. & Buted J.S.(19850. Nucl. Acids Res. v.17, 699-722.

Gysev V.D., Kulichkov V.A. & Chupachina O.M. (1991). Mol. Biol (Russian). V.25, 825-834.

Gysev V.D., Kulichkov V.A.& Chupachina O.M. (1991). Mol. Biol (Russian) V.25, 1080-1089.

Harary F. (1969)."Graph theory" Addison-Wesley publishing company. London.

Kimura M. (1983). The neutral theory of molecular evolution. Cambridge University press, Cambridge.

Korotkov E.V.& Korotkova M.A. (1985) Dokl. Akad. Nauk SSSR. V.280, 1472-1475.

Korotkov E.V.(1986). Dokl. Akad. Nauk SSSR. V.288, 1014-1017.

Korotkov E.V.(1987). Mol. Biol.(Russian). V.21, 478-483.

Korotkov E.V. (1990) Dokl. Akad. Nauk SSSR. V.311, 238-242.

Korotkov E.V.(1991) Mol. Biol.(Russian). V.25, 250-263.

Korotkov E.V.(1992). Izvestia Akad. Nauk SSSR. Seria of biology. N.4., 660-672.

Kolmogorov A.N. (1987). Information theory and theory of algorithms. Moscow. Nauka. P. 238-250.

Kullback S. (1959). Information theory and statistics. New-York, John Willey and Sons, Inc. .

Maeda N., Yahg f., Barnett D.R., Bowman B.H. & Smithies O. (1984) Nature. V. 309, 131-135.

Nucleic acids and protein sequences analysis. (1987) Eds. Bishop M.J. and Daw C.J. IRL Press.

Nussinov R.J. (1987) J. Theor. Biol. V. 125, 219-235.

Ohme M., Kamagashira T., Shinazaki K. & Sugiura M. (1985). Nucl. Acids Res. V. 13, 1045-1056.

Okada N. (1991). Trends in Ecology and Evolution. v. 6, 358-361.

Owerbach D., Rich C. & Taneja K. (1986). Immunogenetics. V. 24, 41-46.

Sacamoto K. & Okada N. (1985). J. Mol. Evol. 1985. V. 22. p. 134-140.

Shannon C.E. (1948). Bell System Tech. J. V. 22, 623-656.

Watson J.D., Tooze J. & Kurtz D.T. (1983) Recombinant DNA. Scientific American Books, New-York,.

Watson J.D. (1990). Science. V. 248, 44-49.

Weiss R.B., Mineura K., Henderson E.E., Duker N.J. & Deriel J.K. (1983). Biochemistry. V. 22, 4501-4507.

Zvonkin A.K. (1973). Uspechi mathematical nauk. V. 25, 83-124.

Supplement 1

The directed graph $D = \langle V, U \rangle$ consists of limited set V of nodes and set U of the well regulated nodes pairs. Any such pair (u,v) is called arc or oriented edge and is meant uv usually. The uv arc is going from node u to node v. The uv arc is called incident to u and v. It is said that v is adjacent from u and u is adjacent to v. The degree of outcome $d^-(v)$ is called the number of nodes which are adjacent from v and the degree of income $d^+(v)$ is called the number of adjacent to v nodes.

The route in directed graph is called the alternating sequence of nodes and arcs $V_1, x_1, v_2, \ldots, x_n, V_{n+1}$ in which each arc x_i is V_i, V_{i+1}. The way is route having different arcs (Harary, 1969).

The two graphs G and H are isomorfic when there exists synonymous reflection between their nodes saving the adjacent relation.

Supplement 2.

Let us note, that for every $G(B) = \langle V, U \rangle$ structure, $V = \{v_1, v_2, \ldots, v_{n+k}\}$ one node v_1 has the degrees $d^+(v_1) = 0$, $d^-(v_1) = 2$, $n-2$ nodes $v_2, \ldots, v_{n-1}$ have the degrees $d^+(v_i) = 1$, $d^-(v_i) = 2$, the one node v_n (equilibrium node) have the degrees $d^+(v_n) = d^-(v_n) = 1$ and k nodes (type nodes) have the degrees $d^-(v_{n+i}) = 0$ and the degrees sum $d^+(v_{n+1}) + \ldots + d^+(v_{n+k}) = n$.

The node v_1 represents the first base b_1 of sequence. The equilibrium node V_n represents the last base b_n of sequence. The nodes $v_2, v_3, \ldots, v_{n-1}$ represent the other n-2 bases of sequence $b_2, \ldots, b_{n-1}$. The each node representing the base of sequence has the one entering arc directed from preceded base representing node (with the exception for v_1 node which has not entering arc) and one going out arc being directed to corresponding type node, and one goinf out arc being directed to the node representing the next base of sequence (with the exception for v_n node). The type nodes represent the base types of sequence. Its number is equal to the number of different bases in sequence and each type node has the entering arcs from the nodes corresponding the same bases.

It is obvious that primary sequence may be restored from the marked structure if the type nodes are marked by concrete base types.

Supplement 3.

Let we choose two orbitraty divisions $D = \{\langle A_1, B_1 \rangle, \ldots, \langle A_m B_m \rangle\}$ and $D' = \{\langle A'_1, B'_1 \rangle, \ldots, \langle A'_m, B_n \rangle\}$ of one sequence pair $\langle A, B \rangle$, such that A_i and B_i are structure identical for $1 \leq i \leq m$ and A'_i and B_i are structure identical for $1 \leq j \leq n$. We say that the division D' is better than the division D if $m > n$, or $m = n$ and $|A'_1| > |A_1|$, or $m=n$ and there exists j such that $|A'_i| = |A_i|$ for $1 \leq i \leq j-1$ and $|A'_j| > |A_j|$. We say that the division D is the best division of sequence pair if there does not exists the better division of the same pair.

We can obtain the best division by algorithm in (Christofides,1975) if it is used for detect maximum. The coincidence matrix is used for building the division. The first pair $\langle A_1, B_1 \rangle$ with maximum $|A_1|$ we obtain from coincidence matrix. Then the chosen matrix element we make equal to 0 and use the new matrix for search the second pair and so on. Obtained division is the best always and we can define are two sequences similar or not.

Eugene V.Korotkov
Maria A.Korotkova

Enlarged similarity of nucleic acids sequences

Manager of edition Korotkova M.A.

ISBN 5-7262-0073-X

	Confirmed for print 1.07.93		Format 60x80 1/16
P.L. 2.0	Reg.publ.l. 2.0	C-012-93	Edition 75 copy Z. 1200

Moscow Physical Engineering Insitute,
Printing hourse of MPEI, 115409, Moscow, Kashirskoe shosse,31