# Application of Paired Correlation Algorithms for the Distance Matrices Between DNA Chains

BORIS MELNIKOV[1], ELENA MELNIKOVA[2]
[1]Faculty of Computational Mathematics and Cybernetics, Shenzhen MSU– BIT University, China, No. 1, International University Park Road, Longgang District, Shenzhen, 518172, CHINA
[2]Department of Information Technologies and Artificial Intelligence, Russian State Social University, No. 4, Wilhelm Pieck str., Moscow, 129226, RUSSIA

*Abstract:* - This paper is a continuation of some previous works by the authors. We consider various algorithms for calculating distances between genomes of similar species (we use primarily mitochondrial DNA, mtDNA) and various distance matrices between the same genomes obtained on the basis of these algorithms. We can say, just to simplify the situation a little, that all our publications on the subject of DNA analysis are associated with various applications of metrics set on such matrices. The paper also has a second subject, i.e., the study of the obtained distance matrices using special statistical characteristics. We consider two matrices obtained for the mtDNA of 32 species of monkeys; the species were selected so that they all belong to different genera. For them, we have obtained 2 matrices of distances between genomes corresponding to the Jaro –Winkler's and Needleman –Wunsch's algorithms. Next, we considered all the triangles obtained in these matrices, and for each of them we used a specially calculated badness. It is actually a measure of the deviation of the resulting triangle from some acuteangled isosceles one. For two sequences of such badness, we have considered variants of paired correlation. At the same time, in addition to the two standard pair correlation algorithms (Spearman's and Kendall's ones), we also considered a new algorithm proposed by us. The reason for considering this new algorithm is as follows. In the usual way of calculating the correlation, we consider only the set of pairwise values of two random variables, without taking into account the pairs themselves. Vice versa, in both of the mentioned pair correlation algorithms, despite their slight difference, we consider only the order of the elements in these pairs, not paying attention to the values themselves; we specifically note that this also applies to Spearman's criterion, which is usually written about as being more suitable for measurements made on an ordinal scale. In our proposed algorithm, we tried to take into account both the value of both random variables and their order in pairs. The results obtained are of interest. Thus, the "pole" variants (i.e., the usual correlation formula and standard pair correlation algorithms) show some (though very small) correlation between two sequences of 4960 pairs of triangles: from 0.1 to 0.4, depending on the specific algorithm, on whether preliminary normalization was carried out, etc. And the "intermediate" variant (taking into account both the order of pairs and the values of random variables) showed a complete lack of connection: the absolute value of the correlation coefficient did not exceed 0.006. Even more interesting is another result obtained in the work, which can be called a small connection between two well-known algorithms for determining the distances between genomes, namely, algorithms of Jaro –Winkler and Needleman –Wunsch.

*Key-words:* - paired correlation, distance matrix, DNA chain.

Received: March 13, 2024. Revised: Agust 7, 2024. Accepted: September 11, 2024. Published: October 9, 2024.

## 1. Introduction and Motivation

This paper is a continuation of some previous works by the authors; among these works, we note, first of all, [1], [2], [3], [4], see also the links from the mentioned papers. We should immediately note that a more complete title of the paper could be the following: "Application of paired correlation algorithms for comparative evaluation of algorithms for distances between DNA chains".

First, let us define the main term about which there is some ambiguity; at the same time, we should note that such ambiguity was discovered by the author primarily on various Internet resources (and to a much lesser extent in scientific publications), and it was noticed in two languages: English and Russian (publications in other languages were "not investigated"). This ambiguity is due to the following. By definition, we always consider correlation for any two objects, hence the incorrect use of the adjective "paired" in relation to the noun "correlation" may arise: any variant of correlation from a similar point of view on Internet resources can be called "paired". Using the correct terms agreed with [5] and other sources, we call pair correlation (or paired correlation) the correlation of two random variables given *by some pairs of values of these quantities* corresponding to each other.

Certainly, a lot of this has been done a long time ago, however, according to the authors of this paper, it has not been done to the end. It is already clear from the previous paragraph that we consider the correlation of two quantities:

- either "in the most unusual way", i.e., actually taking into account the pairs of values of these quantities, but not taking into account the comparative order of the elements of the pairs "within" each of these two quantities;
- or vice versa, i.e., taking into account this order only, but not taking into account the values themselves.

Some details are described in Preliminaries. We tried to take into account both of these characteristics at the same time.

In contrast to the standard approaches briefly described above, we tried to take into account this simultaneous consideration of two different characteristics as follows. Unlike the formula that claims to be universal for any variant of

paired correlation [1], we consider the formula for a pair of pairs of values (like the usual algorithms of paired correlation, Spearman's and Kendall's ones), and the final value of the paired correlation is obtained based on all possible pairs of pairs. However, unlike Spearman's and Kendall's criteria, we take into account the values of these elements of pairs themselves. For more information, see Sections II and V.

Now let us move on to a brief description of the subject area. From the title of our paper [2], we can conclude that it really has two subjects - exactly, algorithms and the data in question; we have already started talking above about the data itself, i.e. about DNA chains. Thus, to these data under consideration, we refer the application of the correlation algorithms we are considering to DNA analysis. Namely, we consider the DNA of monkeys of 32 species; for more information, see Section III.

Now let us look at all this a little more specifically. Earlier, in previous papers, in particular in the ones cited above, we identified various variants of *the badness* of different algorithms that calculate the distances between DNA chains; that is, we work with *algorithms to analyze algorithms*. We shall have to quote our usual thought "about three species"; this text has to be included in almost all our papers on the topic of DNA analysis.

Thus, the quotation is as follows. Let us consider three following species: human (H), chimpanzee (C) and bonobo (B). According to biologists,

- the ancestors of chimpanzees and bonobos diverged about 2 500 000 years ago,
- and the ancestors of humans with both of them diverged about 7 000 000 years ago,

see Fig. 1. As we shall see below, *the exact values are not particularly important*.

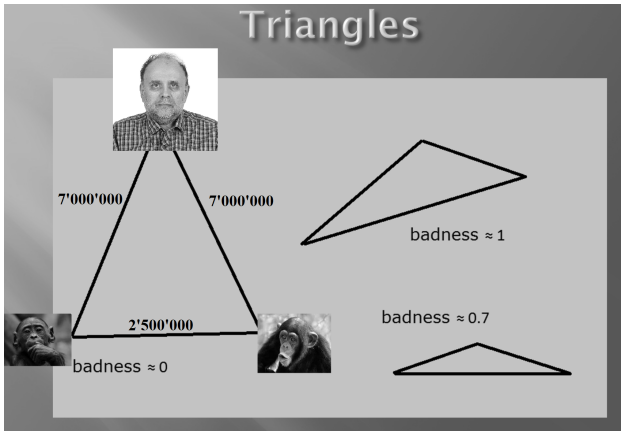



Fig. 1. Some triangles and their approximate badness

Then, the following question arises:

- *why* H should be closer to B comparing S?
- or vice versa: *why* it should be closer to C comparing B?

Obviously, the answer to both these questions is negative, i.e., in other words, the explanation of the greater intimacy cannot exist.

It is very important that all of this can be attributed not only to the mentioned species (H, B, and C), but also to any three species; only the specific values of proximity (or distances, which are usually measured by subtracting proximity from 100%) will be different. Therefore, in the matrix *of distances* between the genomes, all the received triangles should ideally be *acute isosceles* ones. Let us note that the N-dimensional matrix contains $\frac{N\cdot(N-1)}{2}$ values of the distances between its elements [3] that make up $\frac{N\cdot(N-1)\cdot(N-2)}{6}$ triangles. For instance, in considered matrices of size 32, there are 496 distances between pairs forming 4960 triangles.

Thus, it would not be an exaggeration to say that all our publications on the subject of DNA analysis are associated with various applications of metrics set on such matrices. In particular, some of our works are devoted to the restoration of such matrices, [3] etc. In particular, this work is related to the application of special statistical characteristics to them and to their analysis, and to obtain conclusions interesting for biology on this basis.

Much of the above text can be considered *motivation for carrying out work on our topic*. Moreover, the following can also be added to this motivation. Quite a long time ago, the authors suggested that the Jaro – Winkler's and Needleman – Wunsch's algorithms give little similar results to each other [4]. It was to verify this assumption that the calculations were carried out, and we believe that based on this paper, we have demonstrated a way of such verification.

This paper has the following structure.

Section II is the first part of preliminaries. In it, we consider some usual statistical characteristics used in the paper.

Section III is the second part of preliminaries. In it, we consider some previous results of our work.

In Section IV, we describe the object of the research of this paper. Namely, we list the species of monkeys we are considering, all belonging to different genera. After that, we present the distances calculated for the mt DNA of these

[1] However, in our opinion, it is *not* such a universal formula; some details are below.

[2] See also the possible more detailed title given at the beginning of the paper.

[3] In some specific algorithms, including those available on the Internet, the distance from some kind of A to some other kind of B may not coincide with the distance from B to A. We try not to consider such algorithms, or, at least, in such situations, we take a half-sum of distances as the answer.

[4] Let us copy the text from some our previous papers with some changes.

> The *difference* between genomes is very *different* in *different* studies, although the vast majority of both scientific and popular scientific papers give the distance between the genomes of humans and chimpanzees ranges from 0.5% to 2% (i.e., the similarity is from 98% to 99.5%). For example, according to [6], the genomes of humans and chimpanzees are "identical by more than 98.5%", and this statement is very often quoted "as the ultimate truth".

However, in our situation, everything is even much worse, than in the given example, and in the rest of this paper, this fact will be demonstrated using a small value paired correlation of the badness of distance triangles obtained by applying the two mentioned algorithms to the same species.

species in the form of two tables; everything is considered for two different distance calculation algorithms.

Section V could be considered as the main one. In it, we consider the approach to calculation of the pair correlation proposed by us.

In Section VI, we consider some results of computational experiments and give some discussion of them.

Section VII is the conclusion. In it, we formulate some direction of the future work.

## 2. Preliminaries A. Some Used Statistical Characteristics

This section is the first part of preliminaries. In it, we consider some usual statistical characteristics used in the paper, are agreed with [5]; sometimes, however, we use "some more mathematical" notation, for example, we do not use $MXY$ etc. The two random variables under consideration are denoted by $X$ and $Y$; their observed implementations are denoted in the same way with the corresponding subscripts, i.e.,

$$X_i \quad \text{and} \quad Y_i \quad \text{for} \quad i = 1, 2, \ldots, N.$$

Firstly, let us formulate *the usual definition of correlation*: recall that the pair correlation coefficient can be calculated using the usual formulas:

$$R(X, Y) = \frac{\text{cov}(X, Y)}{\sigma_X \cdot \sigma_Y},$$

where

$$\text{cov}(X, Y) = M_{X \cdot Y} - M_X \cdot M_Y.$$

In our further tables and program fragments, this variant of the coefficient *will have the number* $0$.

Secondly, let us formulate *some modificated Kendall's correlation coefficient* [5]. For it, we define *the number of discrepancies* ("entropy coefficient"): a discrepancy holds if for some pair $(i, j)$ where $i \neq j$, we have

$$X_i > X_j \quad \text{but} \quad Y_i < Y_j. \tag{1}$$

Let us denote the number of such discrepancies by $\text{entr}(X, Y)$, or simple $E$ in the next formula.

Since the maximum possible number of such discrepancies is $\frac{N \cdot (N-1)}{2}$, we shall consider the modificated Kendall's correlation coefficient by

$$1 - \frac{4 \cdot E}{N \cdot (N-1)};$$

this value is equal to $1$ in case of $0$ discrepancies, and is equal to $-1$ in case of maximum possible number of discrepancies. In our further tables and program fragments, this variant of the coefficient *will have the number* $2$.

Note that we could calculate this coefficient as follows. We define the "entropy coefficient" considered before for each pair of pairs by (1), then we calculate the sum of these coefficients and divide the result by the value $\frac{N \cdot (N-1)}{2}$ already used earlier.

However, different publications provide different versions of criticism of the Kendall criterion, but the authors of the current paper consider such a flaw to be the most important: it does not give very adequate results with a large number of coincidences in the values of the considered random variables. Therefore we shall also consider the following *"very modificated" Kendall's correlation coefficient*.

It is most convenient to consider it as a search for pairs of pairs, like in the last remark. However, unlike (1), we also use values $0$ (not only $1$ and $-1$): the value $0$ is selected if and only if the values of at least one of the random variables in the considered pairs match.

In our further tables and program fragments, this variant of the coefficient *will have the number* $3$. A fragment of the program for options 2 and 3 (both the modificated Kendall's correlation coefficients) is shown on Fig. 2.

```
if (nReg==2) return
    (pairOne.GetA()-pairOne.GetB())*(pairTwo.GetA()-pairTwo.GetB()) < 0 ?
        -1 : 1;
    // -1 if the pair is "incorrect" and +1 if it is "correct"
else if (nReg==3) { // a more complicated version of the previous one:
    // we take into account the equality of 0 in one of the pairs
    double rOne = pairOne.GetA()-pairTwo.GetA();
    if (::IsNulla(rOne)) return 0;
    double rTwo = pairOne.GetB()-pairTwo.GetB();
    if (::IsNulla(rTwo)) return 0;
    return (rOne*rTwo) < 0 ? -1 : 1;
}
```

Fig. 2. The part of the text of the function for the modificated Kendall's correlation coefficient

Thirdly, the *Spearman's correlation coefficient* is calculated in the usual way, i.e.

$$\frac{\sum_{i=1}^{n} (x_i - M_X) \cdot (y_i - M_Y)}{\sqrt{n \cdot \sigma_X \cdot \sigma_Y}}$$

This is an equivalently modified formula from [5]. In our further tables and program fragments, this variant of the coefficient *will have the number* $1$.

We note in advance that in Section V, our version of the calculation of the pair correlation will also be given. In our further tables and program fragments, our variant of the coefficient *will have the number* $4$.

## 3. Preliminaries – B. On Some Previous Results of the Authors' Work

This section is the second part of preliminaries. In it, we consider some previous results of our work.

Firstly, consider DNA again and using the triangular norm for the threes of distances. Unlike Fig. 1, the new figure shows the concrete values of badness of the concrete triangles, the details are below [6].

The real calculations allowed to compare the quality of the algorithms themselves for estimating distances between DNA

[5] We should immediately note that the correlation calculated in any way between the usual Kendall's correlation coefficient and our variant is always equal to $1$ ("correlation between correlations"), this is easily obtained by trivially considering the formulas.

[6] In exceptional cases, the three may not form a triangle. Then we consider the badness to be very large (significantly greater than 1, within a certain set number $M$). However, there are very few such situations in real computing.

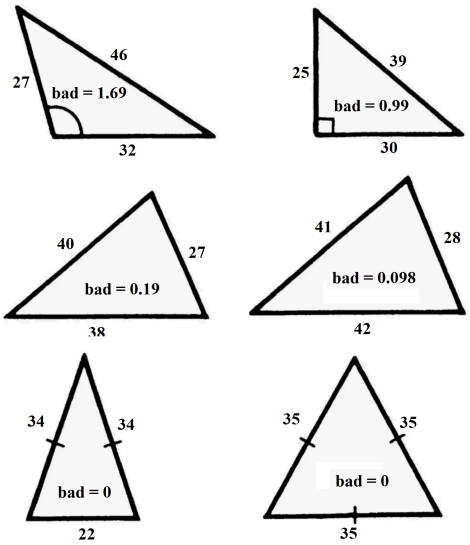



Fig. 3. The triangles and their badness

chains (*"heuristics for comparing heuristics"*). It is important that:

- the distance estimation algorithm
- and the "heuristics for comparing heuristics"

are *in no way related to each other*.

Thus, there are various algorithms to determine the distances between genomes. This raises not only the usual questions about the adequacy of the corresponding mathematical models, but also on the comparative evaluation of these models. For some different algorithms of this type:

- Needleman – Wunsch, [7],
- Smith – Waterman, [8] etc., which could be considered as a modification of previous one,
- Damerau – Levenstein, [9] etc.,
- Melnikov – Panin, [10], which could be considered as a modification of previous one,
- Jaro – Winkler (2 versions), [11],
- van der Loo, [12],

we consider the matrices of distances between the genomes; in our computational experiments (see below some of its descriptions), we used five different algorithms and made corresponding distance matrices, in which the number of genomes reached 100.

The total value of the badness is usually considered equal to the sum of all the badness of the triangles; as we already said, there are, to say, 4960 for dimension $n = 32$.

Note that we consider this matrix much more often, than the matrix *of closeness* that is usually considered in other publications. For instance, the main diagonal of the matrix of closeness contains all 1's, while the main diagonal of the matrix of distances contains all 0's.

The following Table I shows some versions of badness counted for some triangles (and "triangles", if the triangle inequality is violated). The mini-algorithms for calculating

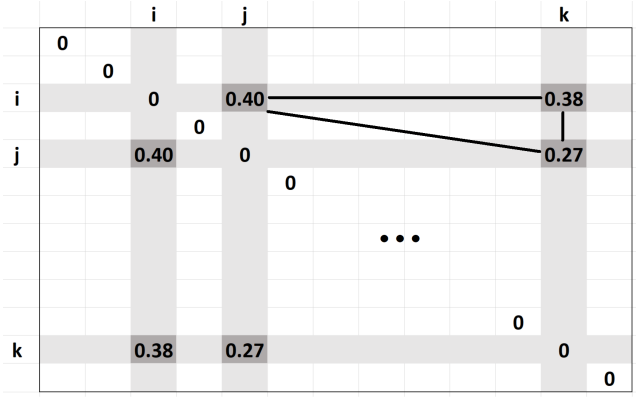
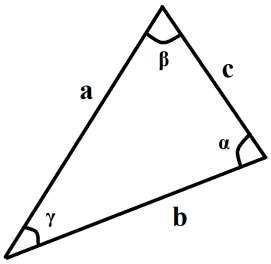



Fig. 4. The distance matrix with the triangles it forms

the values of different badness are shown in the header row of this table. Note that after some computational experiments described, among other, in the papers cited before, we came to the conclusion that version (0) is the best [7]; we shall use it in the rest of this paper.

Here are some additional comments for it. We round it up to an integer of degrees, therefore, the sum may not be the same as 180. For the sides and the angles, we suppose that $a \geqslant b \geqslant c$, and $\alpha \geqslant \beta \geqslant \gamma$. The badness of the kind (4) used in some our previous papers is not shown here. We also consider triples of lengths that do not form triangles; as already noted, this is extremely rare in real calculations (usually less than 0.1%); such triples are called triangles in quotation marks.




Some of the results of previous work are shown in Table II. In it, the titles of the algorithms are given in columns, and the variants of the badness are given in columns; both have already been described above in this section. Let us add only the following.

- The numbers of heuristic algorithms are marked with the first letters of the authors' surnames, see before.
- Column "Time" includes the time for filling in the table of dimension about $32 \times 32$ with the algorithm in question (to get all the values of distance matrix, the processor clock speed is $\approx 2.4$ GHz).
- The object of the study (the species under consideration) will be shortly discussed in the next section.
- Column "Vio." includes the average number of violations of the triangle inequality for all generation problem instances. We recalculated this number "per 1 000 elements" and rounded the result to integers; therefore, to say, the value 12 corresponds approximately to 1.2 %.

[7] We propose to show that this mini-algorithm is better than the others, in approximately the same way that in this paper we are trying to compare two different algorithms for obtaining a distance matrix. However, the comparison indicated in this footnote is not included in the subject of this paper, it will be the subject of one of the following publications.

TABLE I
SOME THREES (THE TRIANGLES AND "TRIANGLES") AND SOME VERSIONS OF THEIR BADNESS

| **Sides** $a, b, c$ | **Angles** $\alpha, \beta, \gamma$ | **Bad. (0)** $(\alpha-\beta)/\gamma$ | **Bad. (1)** $(\alpha-\beta)/\pi$ | **Bad. (2)** $(\alpha-\beta)/\alpha$ | **Bad. (3)** $(a-b)/a$ | **Bad. (5)** $(a-b)/c$ |
|---|---|---|---|---|---|---|
| 1 1 1 | 60 60 60 | 0 | 0 | 0 | 0 | 0 |
| 5 5 4 | 66 66 47 | 0 | 0 | 0 | 0 | 0 |
| 42 41 28 | 72 68 39 | 0.10 | 0.04 | 0.05 | 0.02 | 0.04 |
| 19 18 17 | 66 60 55 | 0.11 | 0.07 | 0.09 | 0.05 | 0.06 |
| 10 9 8 | 72 59 50 | 0.26 | 0.14 | 0.18 | 0.10 | 0.13 |
| 6 5 5 | 74 53 53 | 0.39 | 0.23 | 0.28 | 0.17 | 0.20 |
| 13 12 5 | 90 67 23 | 1.00 | 0.25 | 0.25 | 0.08 | 0.20 |
| 5 4 3 | 90 53 37 | 1.00 | 0.41 | 0.41 | 0.20 | 0.33 |
| 12 6 5 | – | | | 1.09 | | |
| 20 6 5 | – | | | 1.81 | | |

TABLE II
SOME RESULTS OF OUR PREVIOUS PAPERS

| **Algorithm** | **Time (h.)** | **Vio.** | **Bad. (0)** | **Bad. (1)** | **Bad. (2)** | **Bad. (3)** | **Bad. (4)** |
|---|---|---|---|---|---|---|---|
| D–L | 27 | 0 | **0.155** | **0.0522** | **0.121** | 0.0527 | 0.351 |
| **N–W** | 2.1 | 0 | **0.101** | **0.0314** | **0.0692** | **0.0290** | **0.205** |
| J–W | 2.3 | 12 | 1.331 | 0.501 | 0.600 | 0.154 | 0.580 |
| M–P | 28 | 0 | **0.155** | **0.0527** | **0.122** | **0.0482** | **0.323** |
| S–W | 28 | 14 | 0.200 | 0.0732 | 0.150 | 0.0608 | **0.320** |

TABLE III
THE CONSIDERED MONKEY SPECIES IN THE ALPHABETICAL ORDER

| **No.** | **Species of monkeys** |
|---|---|
| 1 | Allenopithecus nigroviridis |
| 2 | Ateles belzebuth |
| 3 | Brachyteles arachnoides |
| 4 | Cacajao calvus |
| 5 | Callimico goeldii |
| 6 | Callithrix jacchus |
| 7 | Carlito syrichta |
| 8 | Cebuella pygmaea |
| 9 | Cephalopachus bancanus |
| 10 | Cercocebus atys |
| 11 | Cercopithecus albogularis |
| 12 | Chlorocebus sabaeus |
| 13 | Colobus angolensis |
| 14 | Erythrocebus patas |
| 15 | Galago moholi |
| 16 | Gorilla gorilla |

| **No.** | **Species of monkeys** |
|---|---|
| 17 | Lagothrix lagotricha |
| 18 | Leontopithecus rosalia |
| 19 | Macaca fascicularis |
| 20 | Macaca fuscata |
| 21 | Mandrillus leucophaeus |
| 22 | Nasalis larvatus |
| 23 | Nycticebus coucang |
| 24 | Papio anubis |
| 25 | Presbytis melalophos |
| 26 | Pygathrix nemaeus |
| 27 | Rhinopithecus roxellana |
| 28 | Saguinus oedipus |
| 29 | Saimiri boliviensis |
| 30 | Semnopithecus entellus |
| 31 | Tarsius dentatus |
| 32 | Theropithecus gelada |

- The and higher accuracy of calculations, apparently, is not interesting here. Note that these violations of the triangle inequality *exist for standard algorithms* of calculating distances between genomes, then such violations are not our problems.
- The remaining columns (badness) have the same meaning as before. The variant of badness (4) is described in more detail in [1] [8].

Based on the calculation results, we can see some advantage of Needleman – Wunsch algorithm over other algorithms.

Now, we are ready to formulate *the main motivation* to perform all our work related to DNA analysis algorithms. Thus, the most important matter in this case is the following one:

> *can we talk about the effectiveness of such algorithms and the adequacy of these models based on the analysis matrices of the distance between the genomes only, without the involvement of biologists?*

The authors of this paper believe that this question should be answered in the affirmative: *yes, we can!*

# 4. The Object of the Research

In this section, we describe the object of the research of this paper.

Firstly, let us list the species of monkeys we are considering, see Table III. It is important to remark, that all the species belonging to different genera: apparently, this fact leads to a more or less successful distribution of the elements of the distance matrix.

After that, we present the distances calculated for the mt DNA of these species in the form of two tables; everything is considered for two different distance calculation algorithms. Namely, for our article we have reviewed the algorithms of Jaro – Winkler and Needleman – Wunsch [9].

Table IV is the calculated distance matrix for the Jaro – Winkler's algorithm. The species numbers correspond to those shown in Table III. The peculiarity of this algorithm is that it gives *very close* answers for these types; therefore, the 3-digit numbers shown in the table correspond to 3 decimal places after 0.0, for instance, 541 means 0.0541.

Table V is the calculated distance matrix for the Needleman – Wunsch's algorithm. The species numbers also correspond to those shown in Table III. This algorithm gives *not very close* answers for these types; therefore, the 3-digit numbers shown in the table correspond to 3 decimal places after 0. (not 0.0), for instance, 375 means 0.375. It is important to note that such an 10 times increase in values does not change any of the values of the badness of the triangles we are considering: indeed, considering the first triangle of the Table IV, the sides 0.0541, 0.0677, and 0.0635, we can say that its badness is exactly equal to the badness of the triangle with the sides 0.541, 0.677, and 0.635.

(In general, as follows from the previous material, we can work with the Table IV and V, as well as with any other tables built on the same principle, simply as with *tables of integers*: the values of badness that we are interested in will be the same.)

The values of the average badness (notation $\delta$) are shown at the bottom of both tables. It is important that these values are very small (in both cases, we also indicated triangles with sides differing by 1, the badness of which is approximately equal to the average badness of 4960 triangles of the corresponding table). From our point of view, the resulting "averaged" triangles (with the sides 10.5, 9.5, and 8.5 for the first example and with the sides 11, 10, and 9 for the second example) are visually almost indistinguishable from equilateral triangles [10].

# 5. The Proposed Approach to Calculation of the Pair Correlation

This section could be considered as the main one. We consider the approach to calculation of the pair correlation proposed by us.

First, it is necessary to say how exactly the sequences of triangles are obtained, the sequences of the badness of which are the subject of analysis using various pair correlation algorithms. The answer to this is very simple: for fixed vertices having numbers 1 and 2, we consider as the third all other possible options in ascending order, then fix vertices 1 and 3 (instead of 1 and 2) and do the same, etc.

Thus, we obtain two different sequences of badness for the same sequence of triangle numbers. For these sequences, we calculate the pair correlation in all the methods described above (recall that they were designated from (0) to (3)), and, in addition, we also use method (4), which we shall briefly describe further. We also remind you that in this method, we tried to take into account both the relative values of the elements in pairs (like methods (1), (2) and (3)) and their exact values (like method (0), i.e. in the case of the usual calculation of the correlation coefficient).

Thus, like methods (2) and (3), we consider the set of pairs of pairs: the first pair is $X_i$ and $X_j$ (for random variable X implementations), and the second one is $Y_i$ and $Y_j$ (for Y). Similarly like methods (2) and (3), each value can be in the range from $-1$ to 1 (with the usual meaning of these values), and the final correlation value is obtained by averaging all obtained values (in our case, 4960 values).

For these pairs, we obtain the value shown on Fig. 5. In it, values $X_i$ and $X_j$ are on the left side, and values $Y_i$ and $Y_j$ are on the right side.

[8] Note that below, we shall equally designate the numbers related to the previously discussed methods of calculating the pair correlation, as well as the numbers for the badness: for instance (2) is the second method for correlation and also the second badness. However, there will be no misunderstandings (ambiguities), it will always be clear from the context what exactly is meant.

[9] The authors express their gratitude to the post-graduate students Li Jiamian and Mu Jingyuan (Shenzhen MSU – BIT University, China), who have calculated the tables given below.

Note in advance that the tables can be copied from the pdf-file and easily processed using any computer programs.

[10] In some our previous works, another variant of badness was also considered, i.e. $\sigma$, not $\delta$. The strict definition of $\sigma$ is of little interest for this work, but when considering the previously cited articles, this should be taken into account.

TABLE IV
THE MATRIX OBTAINED BY APPLYING THE JARO – WINKLER'S ALGORITHM
TO 32 SPECIES OF MONKEYS (NO MORE THAN ONE SPECIES FROM EACH GENUS)

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 | 31 | 32 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 000 | 541 | 677 | 583 | 592 | 541 | 589 | 536 | 562 | 633 | 465 | 610 | 530 | 370 | 512 | 565 | 545 | 800 | 624 | 640 | 520 | 556 | 548 | 562 | 515 | 570 | 726 | 524 | 511 | 589 | 589 | 540 |
| 2 | 541 | 000 | 635 | 387 | 342 | 369 | 396 | 381 | 386 | 733 | 600 | 686 | 463 | 542 | 409 | 549 | 349 | 722 | 698 | 708 | 515 | 440 | 401 | 543 | 462 | 455 | 681 | 388 | 452 | 464 | 383 | 532 |
| 3 | 677 | 635 | 000 | 665 | 676 | 627 | 668 | 626 | 670 | 714 | 728 | 739 | 666 | 678 | 655 | 777 | 617 | 731 | 744 | 760 | 737 | 661 | 663 | 767 | 692 | 680 | 690 | 646 | 648 | 710 | 661 | 753 |
| 4 | 583 | 387 | 665 | 000 | 334 | 396 | 385 | 384 | 396 | 767 | 630 | 727 | 457 | 579 | 422 | 577 | 383 | 677 | 723 | 733 | 546 | 447 | 411 | 571 | 442 | 434 | 637 | 403 | 474 | 447 | 378 | 568 |
| 5 | 592 | 342 | 676 | 334 | 000 | 384 | 395 | 321 | 397 | 777 | 644 | 736 | 481 | 584 | 433 | 570 | 375 | 672 | 742 | 751 | 554 | 451 | 421 | 579 | 429 | 444 | 650 | 418 | 498 | 453 | 393 | 562 |
| 6 | 541 | 369 | 627 | 396 | 384 | 000 | 401 | 319 | 406 | 706 | 581 | 665 | 455 | 528 | 387 | 526 | 383 | 753 | 676 | 675 | 510 | 458 | 381 | 499 | 481 | 457 | 693 | 320 | 436 | 475 | 400 | 527 |
| 7 | 589 | 396 | 668 | 385 | 395 | 401 | 000 | 397 | 389 | 763 | 630 | 727 | 471 | 580 | 425 | 584 | 392 | 695 | 738 | 741 | 556 | 429 | 346 | 573 | 458 | 451 | 657 | 400 | 488 | 463 | 382 | 573 |
| 8 | 536 | 381 | 626 | 384 | 321 | 319 | 397 | 000 | 400 | 723 | 595 | 691 | 453 | 537 | 396 | 527 | 345 | 724 | 687 | 696 | 518 | 457 | 392 | 534 | 474 | 457 | 685 | 312 | 448 | 470 | 392 | 526 |
| 9 | 562 | 386 | 670 | 396 | 397 | 406 | 389 | 400 | 000 | 747 | 585 | 700 | 462 | 561 | 415 | 565 | 390 | 725 | 703 | 722 | 532 | 448 | 403 | 571 | 467 | 469 | 681 | 409 | 477 | 482 | 327 | 546 |
| 10 | 633 | 733 | 714 | 767 | 777 | 706 | 763 | 723 | 747 | 000 | 628 | 635 | 706 | 661 | 676 | 699 | 723 | 674 | 653 | 678 | 634 | 720 | 693 | 677 | 767 | 758 | 538 | 712 | 697 | 793 | 775 | 656 |
| 11 | 465 | 600 | 728 | 630 | 644 | 581 | 630 | 595 | 585 | 628 | 000 | 560 | 584 | 462 | 549 | 535 | 594 | 859 | 568 | 579 | 494 | 596 | 589 | 526 | 636 | 608 | 790 | 582 | 560 | 631 | 639 | 464 |
| 12 | 610 | 686 | 739 | 727 | 736 | 665 | 727 | 691 | 700 | 635 | 560 | 000 | 673 | 610 | 631 | 601 | 687 | 871 | 379 | 381 | 556 | 688 | 669 | 589 | 724 | 706 | 795 | 667 | 646 | 729 | 731 | 571 |
| 13 | 530 | 463 | 666 | 457 | 481 | 455 | 471 | 453 | 462 | 706 | 584 | 673 | 000 | 535 | 446 | 467 | 449 | 741 | 665 | 678 | 434 | 391 | 454 | 454 | 414 | 402 | 678 | 463 | 454 | 413 | 461 | 448 |
| 14 | 370 | 542 | 678 | 579 | 584 | 528 | 580 | 537 | 561 | 661 | 462 | 610 | 535 | 000 | 502 | 566 | 545 | 790 | 614 | 627 | 526 | 545 | 539 | 558 | 578 | 549 | 723 | 511 | 492 | 545 | 582 | 539 |
| 15 | 512 | 409 | 655 | 422 | 433 | 387 | 425 | 396 | 415 | 676 | 549 | 631 | 446 | 502 | 000 | 515 | 400 | 772 | 630 | 642 | 477 | 478 | 395 | 506 | 510 | 483 | 705 | 390 | 437 | 509 | 426 | 493 |
| 16 | 565 | 549 | 777 | 577 | 570 | 526 | 584 | 527 | 565 | 699 | 535 | 601 | 467 | 566 | 515 | 000 | 529 | 913 | 580 | 571 | 401 | 484 | 548 | 350 | 483 | 461 | 836 | 528 | 513 | 481 | 589 | 379 |
| 17 | 545 | 349 | 617 | 383 | 375 | 383 | 392 | 345 | 390 | 723 | 594 | 687 | 449 | 545 | 400 | 529 | 000 | 719 | 684 | 701 | 514 | 442 | 391 | 543 | 462 | 461 | 673 | 376 | 443 | 468 | 387 | 503 |
| 18 | 800 | 722 | 731 | 677 | 672 | 753 | 695 | 724 | 725 | 674 | 859 | 871 | 741 | 790 | 772 | 913 | 719 | 000 | 871 | 884 | 851 | 708 | 759 | 897 | 664 | 690 | 538 | 759 | 763 | 694 | 709 | 874 |
| 19 | 624 | 698 | 744 | 723 | 742 | 676 | 738 | 687 | 703 | 653 | 568 | 379 | 665 | 614 | 630 | 580 | 684 | 871 | 000 | 366 | 579 | 701 | 682 | 565 | 734 | 711 | 799 | 668 | 647 | 721 | 729 | 547 |
| 20 | 640 | 708 | 760 | 733 | 751 | 675 | 741 | 696 | 722 | 678 | 579 | 381 | 678 | 627 | 642 | 571 | 701 | 884 | 366 | 000 | 585 | 717 | 688 | 567 | 752 | 718 | 806 | 679 | 656 | 729 | 739 | 551 |
| 21 | 520 | 515 | 737 | 546 | 554 | 510 | 556 | 518 | 532 | 634 | 494 | 556 | 434 | 526 | 477 | 401 | 514 | 851 | 579 | 585 | 000 | 446 | 515 | 386 | 469 | 462 | 787 | 508 | 498 | 485 | 549 | 344 |
| 22 | 556 | 440 | 661 | 447 | 451 | 458 | 429 | 457 | 448 | 720 | 596 | 688 | 391 | 545 | 478 | 484 | 442 | 708 | 701 | 717 | 446 | 000 | 438 | 473 | 377 | 369 | 644 | 465 | 471 | 379 | 451 | 469 |
| 23 | 548 | 401 | 663 | 411 | 421 | 381 | 346 | 392 | 403 | 693 | 589 | 669 | 454 | 539 | 395 | 548 | 391 | 759 | 682 | 688 | 515 | 438 | 000 | 539 | 492 | 478 | 705 | 380 | 451 | 490 | 416 | 528 |
| 24 | 562 | 543 | 767 | 571 | 579 | 499 | 573 | 534 | 571 | 677 | 526 | 589 | 454 | 558 | 506 | 350 | 543 | 897 | 565 | 567 | 386 | 473 | 539 | 000 | 503 | 479 | 822 | 522 | 509 | 465 | 569 | 372 |
| 25 | 515 | 462 | 692 | 442 | 429 | 481 | 458 | 474 | 467 | 767 | 636 | 724 | 414 | 578 | 510 | 483 | 462 | 664 | 734 | 752 | 469 | 377 | 492 | 503 | 000 | 346 | 627 | 484 | 486 | 344 | 467 | 486 |
| 26 | 570 | 455 | 680 | 434 | 444 | 457 | 451 | 457 | 469 | 758 | 608 | 706 | 402 | 549 | 483 | 461 | 461 | 690 | 711 | 718 | 462 | 369 | 478 | 479 | 346 | 000 | 621 | 460 | 453 | 366 | 451 | 471 |
| 27 | 726 | 681 | 690 | 637 | 650 | 693 | 657 | 685 | 681 | 538 | 790 | 795 | 678 | 723 | 705 | 836 | 673 | 538 | 799 | 806 | 787 | 644 | 705 | 822 | 627 | 621 | 000 | 694 | 699 | 634 | 663 | 805 |
| 28 | 524 | 388 | 646 | 403 | 418 | 320 | 400 | 312 | 409 | 712 | 582 | 667 | 463 | 511 | 390 | 528 | 376 | 759 | 668 | 679 | 508 | 465 | 380 | 522 | 484 | 460 | 694 | 000 | 389 | 478 | 409 | 525 |
| 29 | 511 | 452 | 648 | 474 | 498 | 436 | 488 | 448 | 477 | 697 | 560 | 646 | 454 | 492 | 437 | 513 | 443 | 763 | 647 | 656 | 498 | 471 | 451 | 509 | 486 | 453 | 699 | 389 | 000 | 476 | 488 | 500 |
| 30 | 589 | 464 | 710 | 447 | 453 | 475 | 463 | 470 | 482 | 793 | 631 | 729 | 413 | 545 | 509 | 481 | 468 | 694 | 721 | 729 | 485 | 379 | 490 | 465 | 344 | 366 | 634 | 478 | 476 | 000 | 466 | 479 |
| 31 | 589 | 383 | 661 | 378 | 393 | 400 | 382 | 392 | 327 | 775 | 639 | 731 | 461 | 582 | 426 | 589 | 387 | 709 | 729 | 739 | 549 | 451 | 416 | 569 | 467 | 451 | 663 | 409 | 488 | 466 | 000 | 541 |
| 32 | 540 | 532 | 753 | 568 | 562 | 527 | 573 | 526 | 546 | 656 | 464 | 571 | 448 | 539 | 493 | 379 | 503 | 874 | 547 | 551 | 344 | 469 | 528 | 372 | 486 | 471 | 805 | 525 | 500 | 479 | 541 | 000 |

Average badness $\delta = 0.2429$;

it approximately corresponds to the triangle with the sides 10.5, 9.5, and 8.5.

TABLE V
THE MATRIX OBTAINED BY APPLYING THE NEEDLEMAN – WUNSCH'S ALGORITHM
TO 32 SPECIES OF MONKEYS (NO MORE THAN ONE SPECIES FROM EACH GENUS)

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 | 31 | 32 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 000 | 250 | 375 | 260 | 253 | 256 | 283 | 253 | 277 | 156 | 143 | 192 | 197 | 157 | 274 | 216 | 253 | 477 | 206 | 204 | 154 | 187 | 284 | 161 | 188 | 192 | 381 | 263 | 256 | 192 | 281 | 153 |
| 2 | 250 | 000 | 293 | 184 | 168 | 168 | 267 | 167 | 265 | 250 | 253 | 287 | 258 | 256 | 264 | 240 | 123 | 473 | 289 | 289 | 246 | 247 | 275 | 254 | 245 | 249 | 473 | 180 | 179 | 251 | 267 | 249 |
| 3 | 375 | 293 | 000 | 322 | 323 | 320 | 371 | 320 | 368 | 373 | 377 | 476 | 380 | 375 | 384 | 375 | 286 | 476 | 474 | 476 | 374 | 372 | 383 | 378 | 377 | 376 | 296 | 327 | 329 | 381 | 372 | 374 |
| 4 | 260 | 184 | 322 | 000 | 191 | 191 | 271 | 189 | 270 | 258 | 263 | 295 | 264 | 264 | 271 | 258 | 182 | 476 | 297 | 298 | 257 | 258 | 278 | 265 | 258 | 259 | 405 | 196 | 199 | 259 | 270 | 261 |
| 5 | 253 | 168 | 323 | 191 | 000 | 146 | 268 | 145 | 265 | 255 | 258 | 289 | 259 | 259 | 272 | 251 | 169 | 474 | 292 | 293 | 253 | 253 | 276 | 260 | 250 | 250 | 472 | 169 | 184 | 251 | 269 | 256 |
| 6 | 256 | 168 | 320 | 191 | 146 | 000 | 276 | 091 | 271 | 254 | 254 | 286 | 261 | 259 | 271 | 249 | 165 | 477 | 286 | 287 | 252 | 253 | 274 | 255 | 255 | 255 | 474 | 163 | 180 | 256 | 273 | 253 |
| 7 | 283 | 267 | 371 | 272 | 268 | 276 | 000 | 272 | 152 | 283 | 287 | 319 | 286 | 285 | 255 | 279 | 266 | 477 | 319 | 320 | 281 | 277 | 253 | 289 | 276 | 275 | 406 | 273 | 278 | 282 | 177 | 281 |
| 8 | 253 | 167 | 320 | 189 | 145 | 091 | 272 | 000 | 266 | 251 | 253 | 286 | 257 | 257 | 269 | 247 | 165 | 474 | 289 | 288 | 251 | 254 | 275 | 256 | 253 | 253 | 471 | 163 | 181 | 255 | 269 | 254 |
| 9 | 277 | 265 | 368 | 270 | 266 | 271 | 152 | 266 | 000 | 275 | 277 | 312 | 279 | 276 | 250 | 272 | 263 | 477 | 311 | 313 | 275 | 271 | 250 | 282 | 272 | 271 | 402 | 270 | 273 | 275 | 172 | 275 |
| 10 | 156 | 250 | 373 | 258 | 255 | 254 | 283 | 251 | 275 | 000 | 159 | 201 | 202 | 173 | 275 | 212 | 249 | 477 | 191 | 191 | 084 | 190 | 279 | 153 | 191 | 196 | 377 | 260 | 253 | 192 | 279 | 148 |
| 11 | 143 | 253 | 377 | 263 | 258 | 254 | 287 | 253 | 277 | 159 | 000 | 174 | 202 | 140 | 273 | 215 | 251 | 480 | 205 | 202 | 153 | 191 | 280 | 162 | 191 | 193 | 384 | 264 | 258 | 194 | 283 | 156 |
| 12 | 192 | 287 | 476 | 295 | 289 | 286 | 319 | 286 | 312 | 201 | 174 | 000 | 244 | 193 | 301 | 246 | 285 | 479 | 160 | 157 | 201 | 236 | 312 | 203 | 235 | 234 | 478 | 291 | 287 | 237 | 316 | 200 |
| 13 | 197 | 258 | 380 | 264 | 259 | 261 | 286 | 257 | 279 | 202 | 202 | 244 | 000 | 209 | 281 | 227 | 256 | 478 | 246 | 245 | 197 | 167 | 284 | 200 | 176 | 174 | 363 | 266 | 267 | 174 | 283 | 197 |
| 14 | 157 | 256 | 375 | 264 | 259 | 259 | 285 | 257 | 276 | 173 | 140 | 193 | 209 | 000 | 278 | 225 | 258 | 478 | 221 | 219 | 169 | 200 | 287 | 179 | 200 | 206 | 378 | 270 | 265 | 207 | 282 | 173 |
| 15 | 274 | 264 | 384 | 271 | 272 | 271 | 255 | 269 | 250 | 275 | 273 | 301 | 281 | 278 | 000 | 267 | 266 | 476 | 301 | 301 | 274 | 273 | 202 | 277 | 273 | 273 | 476 | 271 | 275 | 278 | 249 | 272 |
| 16 | 216 | 240 | 375 | 258 | 251 | 249 | 279 | 247 | 272 | 212 | 215 | 246 | 227 | 225 | 267 | 000 | 244 | 481 | 245 | 245 | 208 | 217 | 275 | 216 | 219 | 222 | 399 | 254 | 250 | 222 | 275 | 210 |
| 17 | 253 | 123 | 286 | 182 | 169 | 165 | 266 | 165 | 263 | 249 | 251 | 285 | 256 | 259 | 266 | 245 | 000 | 473 | 288 | 289 | 247 | 248 | 272 | 252 | 252 | 250 | 407 | 179 | 179 | 251 | 267 | 247 |
| 18 | 477 | 472 | 476 | 476 | 474 | 477 | 476 | 474 | 477 | 477 | 480 | 479 | 478 | 478 | 476 | 482 | 473 | 000 | 480 | 481 | 478 | 479 | 477 | 478 | 475 | 476 | 476 | 477 | 477 | 479 | 474 | 479 |
| 19 | 206 | 289 | 474 | 297 | 292 | 286 | 319 | 289 | 311 | 191 | 205 | 160 | 246 | 221 | 301 | 245 | 288 | 480 | 000 | 077 | 189 | 234 | 311 | 200 | 237 | 237 | 477 | 296 | 290 | 239 | 317 | 199 |
| 20 | 204 | 289 | 475 | 298 | 293 | 287 | 320 | 288 | 313 | 191 | 202 | 157 | 245 | 219 | 301 | 245 | 289 | 481 | 077 | 000 | 189 | 236 | 312 | 196 | 236 | 236 | 478 | 293 | 288 | 241 | 316 | 195 |
| 21 | 154 | 246 | 374 | 257 | 253 | 252 | 281 | 251 | 275 | 084 | 153 | 201 | 197 | 169 | 274 | 208 | 247 | 477 | 189 | 189 | 000 | 185 | 281 | 146 | 187 | 190 | 379 | 256 | 253 | 190 | 276 | 141 |
| 22 | 187 | 247 | 372 | 258 | 254 | 253 | 277 | 254 | 271 | 190 | 191 | 236 | 167 | 200 | 273 | 217 | 248 | 479 | 234 | 236 | 185 | 000 | 279 | 193 | 142 | 129 | 336 | 264 | 257 | 145 | 271 | 187 |
| 23 | 284 | 275 | 383 | 278 | 276 | 274 | 253 | 275 | 250 | 279 | 280 | 312 | 284 | 287 | 202 | 275 | 272 | 477 | 311 | 312 | 281 | 279 | 000 | 287 | 282 | 281 | 476 | 276 | 279 | 284 | 253 | 282 |
| 24 | 161 | 254 | 378 | 265 | 260 | 255 | 289 | 256 | 282 | 153 | 162 | 203 | 200 | 179 | 277 | 216 | 252 | 479 | 200 | 196 | 146 | 193 | 286 | 000 | 199 | 197 | 382 | 264 | 260 | 196 | 286 | 095 |
| 25 | 188 | 245 | 377 | 258 | 250 | 255 | 276 | 253 | 272 | 191 | 192 | 235 | 176 | 200 | 273 | 219 | 252 | 474 | 237 | 236 | 187 | 142 | 282 | 199 | 000 | 148 | 348 | 267 | 256 | 148 | 272 | 192 |
| 26 | 192 | 249 | 376 | 259 | 250 | 255 | 275 | 253 | 271 | 196 | 193 | 234 | 174 | 206 | 273 | 222 | 250 | 477 | 237 | 236 | 190 | 129 | 281 | 197 | 148 | 000 | 339 | 264 | 256 | 153 | 276 | 192 |
| 27 | 381 | 473 | 296 | 405 | 472 | 474 | 406 | 471 | 402 | 377 | 384 | 478 | 363 | 378 | 475 | 399 | 407 | 476 | 477 | 478 | 379 | 336 | 476 | 382 | 348 | 339 | 000 | 477 | 471 | 352 | 403 | 380 |
| 28 | 263 | 180 | 327 | 196 | 169 | 163 | 273 | 163 | 270 | 260 | 264 | 291 | 266 | 270 | 270 | 254 | 179 | 477 | 296 | 293 | 256 | 264 | 276 | 264 | 267 | 264 | 477 | 000 | 190 | 265 | 273 | 259 |
| 29 | 256 | 179 | 329 | 199 | 184 | 180 | 278 | 181 | 273 | 253 | 258 | 286 | 267 | 265 | 275 | 250 | 179 | 477 | 290 | 288 | 253 | 257 | 279 | 260 | 256 | 256 | 472 | 190 | 000 | 261 | 275 | 255 |
| 30 | 192 | 251 | 380 | 259 | 251 | 256 | 282 | 254 | 275 | 192 | 194 | 237 | 174 | 207 | 278 | 222 | 251 | 480 | 239 | 241 | 190 | 145 | 284 | 196 | 148 | 153 | 352 | 265 | 261 | 000 | 279 | 195 |
| 31 | 281 | 267 | 372 | 270 | 269 | 273 | 177 | 269 | 172 | 280 | 283 | 316 | 283 | 282 | 249 | 275 | 267 | 475 | 317 | 316 | 276 | 272 | 253 | 286 | 272 | 276 | 403 | 273 | 275 | 279 | 000 | 279 |
| 32 | 153 | 249 | 374 | 261 | 256 | 253 | 281 | 254 | 275 | 148 | 156 | 200 | 197 | 173 | 272 | 210 | 247 | 479 | 199 | 195 | 141 | 187 | 282 | 095 | 192 | 192 | 380 | 259 | 255 | 195 | 279 | 000 |

Average badness $\delta = 0.2233$;

it approximately corresponds to the triangle with the sides 11, 10, and 9.

It is important that $X_i \leqslant X_j$ and $Y_i \leqslant Y_j$ (otherwise, we change *its order*, *changing* also the sign of the answer), and $X_j - X_i \leqslant Y_j - Y_i$ (otherwise, we change *the names*, *not changing* the sign of the answer). The answer is

$$R = \frac{\delta_A \cdot S}{\delta_B \cdot (S+1)}, \quad \text{where} \quad S = \frac{\delta_A^2}{2\delta_\delta} \quad \text{and} \quad \delta_\delta = \delta_B - \delta_A;$$

two other values are shown on the figure. This mini-algorithm is also shown in C++ on the following Fig. 6 [11].

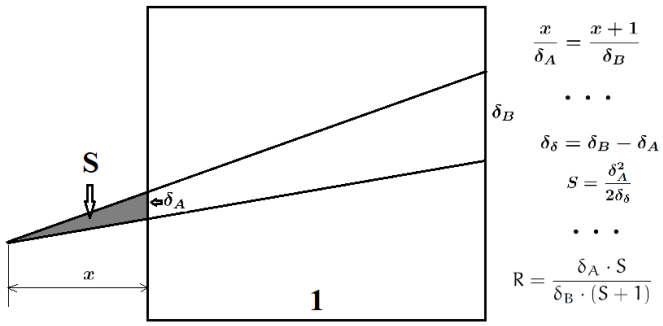



Fig. 5. The proposed calculation of the pair correlation

```
bool bOrder = true; // by default, the correct order is in both pairs
double A1 = pairOne.GetA(), B1 = pairOne.GetB(),
       A2 = pairTwo.GetA(), B2 = pairTwo.GetB();
if (A1<A2) { Swap(A1,A2); Swap(B1,B2); bOrder = !bOrder; }
if (B1<B2) { Swap(B1,B2); bOrder = !bOrder; }
// we obtained A1>=A2, B1>=B2,
// and if !bOrder then we make the negative answer
double deltaA = A1-A2, deltaB = B1-B2;
if (deltaA>deltaB) { Swap(A1,B1); Swap(A2,B2); Swap(deltaA,deltaB); }
// we obtained deltaA<=deltaB,
// but we do not change bOrder here!
if (::IsNulla(deltaA)) return (bOrder ? deltaB : -deltaB);
double deltadelta = deltaB-deltaA;
if (::IsNulla(deltadelta)) return 0.0;
double double Return = (deltaA*S)/deltaB*(S+1.0);
return (bOrder ? Return : -Return);
```

Fig. 6. The part of the text of the function for the proposed calculation of the pair correlation

Here are examples of our version of pair correlation for some specific pairs of value pairs. The captions to the above figures show whether we observe a strong, medium or small correlation value, including figures for degenerate cases.

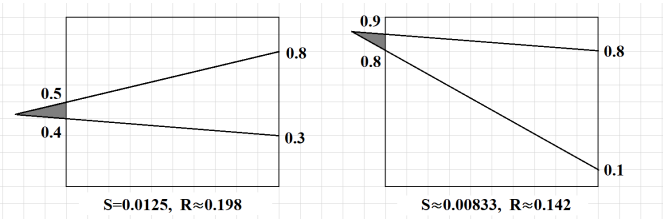



Fig. 7. Examples of calculating values for the observed "small" correlation

[11] Note that in the preliminary versions of the calculations, we tested a much simpler formula, exactly $R = \frac{\delta_A}{\delta_B}$ for the same case: $A_1 > A_2$ and $B_1 > B_2$. However, after some further calculations, we came to the conclusion that the formula considered in this paper is some better. The details may be of interest, and perhaps we shall discuss this thing in one of the following publications.

Firstly, consider Fig. 7. Both examples correspond to the same order of elements in pairs (as well as all further drawings, otherwise we change the sign of the answer), but at the same time in one of the sequences [12], the difference in the values of the elements is much smaller than in the other. As expected, the correlation value is positive, but very small.

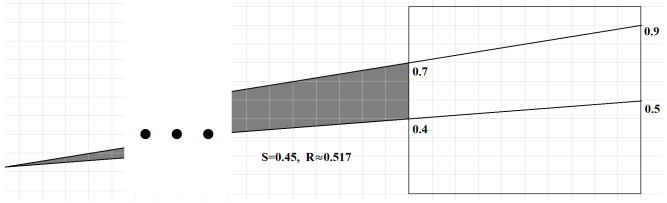



Fig. 8. Example of calculating value for the observed "big" correlation

Secondly, consider Fig. 8. It corresponds to the case, when the difference of the same values is much more. As expected, the correlation value is more than 0.5.

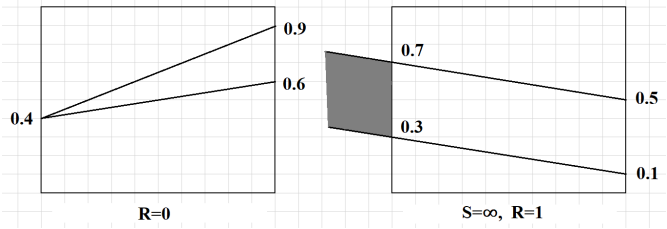



Fig. 9. Example of calculating value for the degenerate cases

Thirdly, consider two extreme cases, Fig. 9.

At the end of reviewing these examples, we note the following. In all the examples (excluding the left degenerate case, see the left part of Fig. 9), it makes sense to consider only the methods of calculating the correlation (4) and (0) (see Section II); the other methods, i.e. (1), (2) and (3), are not meaningless, but make some sense only when considering more than one pairs of values. Thus, each time, we can use the above formulas to calculate the usual value of the correlation coefficient $R_{(0)} = 0.5$. We consider the values we receive to be closer to the truth.

(Let us remark that we can not count it: we understand from the statistics course that each time this value turns out to be equal $R_{(0)} = 0.5$, excluding the left degenerate case only.)

Let us also repeat that we are averaging the values in all pairs. Thus, in the examples considered in the paper, the dimensions of the matrices are 32. As already noted, two sequences of badness are formed, each of which consists of 4960 values. Therefore, there are

$$\frac{4960 \cdot 4959}{2} = 12\,298\,320$$

pairs of such values in total for averaging.

[12] Not "in one of the pairs", those are different things.

## 6. Some Results of Computational Experiments and Some Discussion of Them

The work on comparing algorithms of Jaro – Winkler and Needleman – Wunsch was carried out by us due to the fact that the correlation between these algorithms was not visually visible. This is exactly what we got as a result of the calculations done, and it was our variant (4) that showed the result closest to 0.

In general, all the calculation results are shown in the following Table VI; in the second line ("with"), we used normalization, and in the first line ("without"), did not use it. As usual, normalization is what we call the linear mapping of all the received data into any segment; as a rule, a specific variant of the segment is indifferent, for example, it may be $[0, 1]$.

The columns are certainly the methods of calculation of the pair correlation (not the badness).

TABLE VI
THE RESULTS OF COMPUTATIONAL EXPERIMENTS

| **Option** | (0) | (1) | (2) | (3) | (4) |
|---|---|---|---|---|---|
| without | 0.0817 | 0.136 | 0.0742 | 0.0909 | $\approx 10^{-4}$ |
| with | 0.0817 | 0.136 | 0.139 | 0.0909 | $\approx 10^{-5}$ |

Certainly, most of the results do not depend on the possible use of normalization; this can be also simply obtained as a consequence of the description of the algorithms used to calculate the pair correlation.

And, of course, the above tables $32 \times 32$ can be also considered the results of the calculations obtained, especially since in this paper we used new set of species for the Jaro – Winkler algorithm.

## 7. Conclusion

We believe that the presented article has three main results, and we cannot yet say which one is more important.

First, we have presented *a new possible method for calculating pair correlation*, which was not specified in previous monographs and papers, [5] etc.

The second result is a description of the application of pair correlation (any variant of it, not necessarily considered by us) *to the comparison of various algorithms for calculating distances between genomes*.

Even more interesting is another result obtained in the work, which can be called *a small connection between two well-known algorithms* for determining the distances between genomes, namely, algorithms of Jaro – Winkler and Needleman – Wunsch.

Let us formulate some direction of the future work.

First, linear regression algorithms are not needed for future calculations, since, according to the meaning of the problem, good algorithms should ideally give a pair correlation value close to 1. In practice, this is not the case, so for application it is necessary to choose one of the algorithms for calculating the distances between genomes. By and large, this is the choice that most of our works on this topic are devoted to.

Secondly, in the near future it is necessary to improve the formula itself for calculating the correlation value for two pairs consisting of pairs of elements of two random variables.

Thirdly, it is desirable to strictly formulate the principles of constructing auxiliary algorithms for calculating correlation, so that the algorithms given in Section V fall within these principles.

## Acknowledgment

The work of the first author was partially supported by a grant from the scientific program of Chinese universities "Program to support the stability of Higher Education" (section "Shenzhen 2022 – Commission on Science, Technology and Innovation of the Shenzhen Municipality").

### *References*

[1] B. Melnikov, S. Pivneva, and M. Trifonov. Various algorithms, calculating distances of DNA sequences, and some computational recommendations for use such algorithms. *Conference: Information Technology and Nanotechnology*, 2017, p. 43–50, DOI: 10.18287/1613-0073-2017-1902-43-50.

[2] B. Melnikov, M. Trenina, and E. Melnikova. Different approaches to solving the problem of reconstructing the distance matrix between DNA chains. *Communications in Computer and Information Science, CCIS-1201*, 2020, p. 211–223.

[3] M. Abramyan, B. Melnikov, and Y. Zhang. Some more on restoring distance matrices between DNA chains: reliability coefficients. *Cybernetics and Physics*, Vol. 12, No. 4, 2023, p. 237–251.

[4] B. Melnikov. Algorithms for calculating DNA distances based on the pair correlation. *Journal of Harbin Engineering University*, Vol. 44, No. 12, 2023, pp. 1462–1466.

[5] M. Lagutin. *Visual mathematical statistics*. Moscow: Binom, 2012, 472 p. (in Russian).

[6] N. Polavarapu, G. Arora, V. Mittal, and J. McDonald, Characterization and potential functional significance of human-chimpanzee large INDEL variation. *Mob DNA*, Oct 25:2:13, 2011.

[7] S. Needleman and Ch. Wunsch, A general method is applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, Vol. 48, No. 3, 1970, pp. 443–453.

[8] Y. Munekawa, F. Ino, and K. Hagihara. Design and implementation of the Smith-Waterman algorithm on the CUDA-compatible GPU. *8th IEEE International Conference on BioInformatics and BioEngineering, BIBE-2008*, DOI: 10.1109/BIBE.2008.4696721.

[9] V. Levenshtein. Binary codes capable of correcting. Deletions, insertions, and reversals. *Soviet Physics Doklady*, Vol. 10, 1966, pp. 707–710.

[10] B. Melnikov and A. Panin. Parallel implementation of the multiheuristic approach in the task of comparing genetic sequences. *Vector of science of Tolyatti State University*, Vol. 22, No. 4, 2012, pp. 83–86 (in Russian).

[11] W. Winkler. String comparator metrics and enhanced decision rules in the Fellegi-Sunter model of record linkage. *Proceedings of the Survey Research Methods Sections, American Statistical Association*, 1990, pp. 354–359.

[12] M. van der Loo. The string-dist package for approximate string matching. *The R Journal*. Vol. 6, 2014, pp. 111–122.

**Contribution of Individual Authors to the Creation of a Scientific Article (Ghostwriting Policy)**

The authors equally contributed in the present research, at all stages from the formulation of the problem to the final findings and solution.

**Sources of Funding for Research Presented in a Scientific Article or Scientific Article Itself**

No funding was received for conducting this study.

**Conflict of Interest**

The authors have no conflicts of interest to declare that are relevant to the content of this article.