MATHEMATICAL PHYLOGENETICS OF SPECIES

Makarov L.M.[1], Ivanov D.O.[2], Pozdnyakov A.V.[3] (Russian Federation)
Email: Makarov457@scientifictext.ru

[1]Makarov Leonid Mikhailovich - Candidate of Technical Sciences, Professor,
DEPARTMENT AUTOMATION OF COMMUNICATION ENTERPRISES,
ST. PETERSBURG STATE UNIVERSITY OF TELECOMMUNICATIONS NAMED AFTER PROF. M.A. BONCH-BRUEVICH;
[2]Ivanov Dmitriy Olegovich - Doctor of Medical Sciences, Professor, Rector;
[3]Pozdnyakov Alexander Vladimirovich - Doctor of Medical Sciences, Professor,
DEPARTMENT OF MEDICAL BIOPHYSICS,
ST. PETERSBURG STATE PEDIATRIC MEDICAL UNIVERSITY,
ST. PETERSBURG

Abstract: *the systematic approach of classification of species of living organisms in terms and concepts of phylogenetics is presented. The possibility of using information technologies to create a graphic image of a living organism based on the genome has been established. Procedures for analyzing evolutionary processes in terms of cladistics have been created. A source of NCBI genomes was used to create computer formation procedures. The metric display space of the symbol of different organisms. The formal definition of the species in terms of metric space is co-defined using the individual metric of the nucleotide set.*
Keywords: *phylogenetics, taxonomic system, computer analysis.*

МАТЕМАТИЧЕСКАЯ ФИЛОГЕНЕНТИКА ВИДОВ

Макаров Л.М.[1], Иванов Д.О.[2], Поздняков А.В.[3] (Российская Федерация)

[1]Макаров Леонид Михайлович – кандидат технических наук, профессор,
кафедра автоматизации предприятий связи,
Санкт-Петербургский государственный университет телекоммуникаций им. проф. М.А. Бонч-Бруевича;
[2]Иванов Дмитрий Олегович - доктор медицинских наук, профессор, ректор;
[3]Поздняков Александр Владимирович – доктор медицинских наук, профессор,
кафедра медицинской биофизики,
Санкт-Петербургский государственный педиатрический медицинский университет,
г. Санкт-Петербург

Аннотация: *представлен системный подход классификации видов живых организмов в терминах и понятиях филогенетики. Установлена возможность использования информационных технологий создания графического образа живого организма на основе генома. Созданы процедуры анализа эволюционных процессов в терминах кладистики. Использован источник NCBI геномов, для создания компьютерных процедур формирования метрического пространства отображения графического образа различных организмов. Создано формальное определение вида в терминах метрического пространства с использованием индивидуальной метрики набора нуклеотидов.*
Ключевые слова: *филогенетика, таксономическая система, компьютерный анализ.*

UDC 57.06+519.7

Understanding of natural processes of Nature is formed on the basis of observations and systematization of knowledge. The creation of a large number of information sources of knowledge about living organisms on the Internet contributes to the development of phylogenetic research. Information technology, placed on the field of phylogenetics, introduced new knowledge about the evolution and peculiarities of the structure of different living organisms.

The presence of a large variety of living organisms in Nature initializes the creation of a scientific concept for the systematization of species, implemented by means of mathematics, providing the possibility of a formalized comparison of different organisms. The most progressive work on the systematization of species was prepared at the beginning of the 18th century by K. Linnaeus, who created the general principles of the systemic description of organisms [1]. The system of description of living organisms was based on the categories of genus and species of the organism. The modern system of description of living organisms is represented by five kingdoms (Fig. 1).

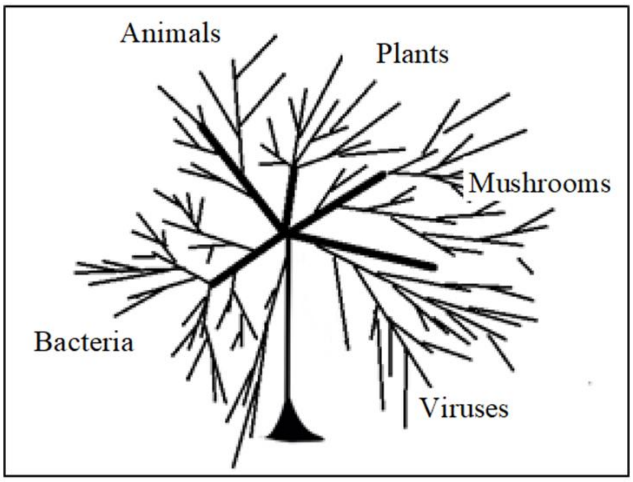

Fig. 1. Kingdoms of living organisms

Living Nature, with a wide variety of organisms, provides the opportunity to construct logical schemes of true judgments about the presence of a system of mutual communication in the evolution of different organisms. The relevance of this problem to phylogenetics is obvious. This thesis is strengthened in numerous scientific works declaring the presence in the distant past of the only ancestor of all organisms.

A systematic approach to full-scale research allows for the wide application of expert estimates of the similarity/similarity of living organisms. Biological taxonomy uses many methods for creating expert assessments, which are the basis for building a system or the organic world of Nature. The basic concepts of biological systematics are believed to be the presence of a hierarchical structure (scheme) of living Priroda, in which homogeneous groups are distinguished - taxa and ranks.

The modern classification of living organisms reproduces that hierarchical scheme, in which the upper and lower level of the hierarchy (rank) is distinguished: kingdom, type, class, order, family, genus and species. Using expert assessments, it is believed that any particular organism has a kinship at all levels of the hierarchical system.

This is a traditional judgment scheme based on a morphometric method of describing organisms. The morphological similarity of organisms as a species is recognized as the main argument in the creation of a taxonomical scheme.

The term - species is correlated with a group of closely related, by morphological features, organisms. In the general understanding of this term, the concept of the combination of several organisms with common morphological features is invested, and most importantly, propagating within the isolated group. These simple concepts created the basis for the formation of the concept of the natural evolution of living material forms [2]. Following these ideas, we note that the theory of evolution proposed by C. Darwin (*Charles Robert Darwin*) well complements the theory of classification of living organisms of C. Linnaeus (*Carolus Linnaeus)*. In this conceptual thesis, the evolution and modification of living forms of matter is recognized by the phenomenon of Nature, which is realized by the genotype.

Understanding the substance of this problem, instrumental research methods are actively involved. In particular, methods of genetic research. By the middle of the twentieth century, an evolutionary concept was formed for all organisms, the basis of which was genetics. Possessing the material of instrumentally obtained data on the genotype of a large number of organisms, it becomes possible to operate with information concepts and definitions. A new qualitative description of the evolution of species is created in terms of the scientific discipline of cariology, using the concepts of the nucleosome and nucleotides.

The established scientific postulates on evolution are well consistent with the theory of E. Major [3] in which three axioms are identified as the basis:

1. species created by Nature based on guanine (G), thymine (T), cytosine (C) and amine (A);
2. species are represented by individual individuals;
3. populations of one species are reproductively isolated

The isolation of the thesis of four nucleotides determining the uniqueness of a living organism as a species is the basis for genetics [4]. This basic concept for genetics initializes the launch of information technologies for the extraction of new knowledge on the field of numerous experimental materials of phylogenetics. In the theory of genetics, these nucleotides occupy a leading position that determines the individual information program of the development of the body, and therefore a place in the taxonomic scheme.

On this basis, key judgments are formulated that all species are represented by populations that have a common genetic program historically formed in the process of evolution. In this understanding, the definition of the species of an organism is created on the representation of a group, as a population within which reproduction processes are actually reproduced, and isolated from other groups.

In the historical process of evolution of the group, a natural modification of the set of nucleotides occurs, which appears to be a non-empty set of X (x, y). We fix this aspect of judgment in terms of mathematics. A set of elements of the same origin is called a metric space if the metric pi (x, y) is entered, for which three conditions [5] are valid:

$$1).\ \rho(x,y)\geq 0;\quad 2).\ \rho(x,y)=0\ \text{ if }\ x=y\ ;\quad 3).\ \rho(x,y)=\rho(y,x) \qquad (1)$$

In such a case, the metric space is positioned by a pair (X, p) where X is not an empty set of nucleotides.

In terms of biological informatics, this clearly indicates the possibility of unique positioning of sets of nucleotides in metric space. Add to this concept.

In terms of a mathematical topology that generalizes the properties of a metric space, two figures, or in other words, two spaces, are called homeomorphic if there is a mutual unambiguously continuously mapping one shape to another. In this formulation, the concept of space and some geometric figure are logically combined. Speaking of space, it was implicit that there was some metric related to a geometric figure whose position could be considered in a system of two or three coordinates. The reverse statement is also true: space is compared with a certain way - a figure and, of course, with an area bounded by this figure. Some geometric figures have homeomorphism.

In mathematics, the concept of a topological space is a generalization of a metric space in which geometric images are placed. Topological spaces are naturally formed in the process of analyzing numerous indicators of processes and objects. So, for example, this procedure, implemented according to the numerical indicators of the set of nucleotides, forms a graphic image - a figure.

Moving to the discussion of the geometric image formed in metric space, the idea is clearly manifested not only visualization of the image, but also the possibility of comparing several images obtained on different nucleosomes.

Let's build a geometric image of space, which has numerous elements that form a set. Let X = Rm be a nonempty set, where m is the dimension of the arithmetic space, the elements of which are represented by quantitative indicators of the nucleotide set: A, T, G, C. This rule, created in terms of mathematics, applies to all species of organisms.

In this case, for any two elements of the set X there is always a non-negative number, which is a metric ($\boldsymbol{\rho}$). For metric $\boldsymbol{\rho}$, the following relations are performed:

$$X = R, \qquad \forall A, T, G, C \in X \qquad (2)$$

and

$$\begin{aligned}
\rho_1(A,T) &= |A-T| = \sqrt{(A-T)^2}\ ;\\
\rho_2(A,G) &= |A-G| = \sqrt{(A-T)^2}\ ;\\
\rho_3(A,C) &= |A-C| = \sqrt{(A-C)^2}|;\\
\rho_4(T,G) &= |T-G| = \sqrt{(T-G)^2};\\
\rho_5(T,C) &= |T-C| = \sqrt{(T-C)^2};\\
\rho_6(G,C) &= |G-C| = \sqrt{(G-C)^2};
\end{aligned} \qquad (3)$$

Enter axiomatic rules:

1. any nucleosome is created on the basis of four nucleotides, the numerical indicators of which, named in the prescribed order, identify the coordinates of elements in the metric space (2);

2. the metric in the four nucleotide basis is created as a paired combination of element coordinates;

3. a paired combination of four nucleotide coordinates forms a sweep track of six events (3);

4. track of scanning events on a plane forms a nucleosome image.

Taking these positions as a basis, we will create an image of a biological species on a taxonomic scheme [1,4]. The taxonomic scheme is represented by three levels: 1). root (first level - base), 2). node (the second level is the place of the first addition of the number of species); 3). leaves (the third level is the place of multiple addition of species). In the modern taxonomic scheme, there is a clear selection of the positioning levels of living organisms, but there is no symbol of organisms.

By means of mathematical analysis, we create a formal criterion for the equivalence of images. We use the NCBI information resource, which contains sets of genomes of organisms [6]. Genome - an array of nucleotides of a particular organism: A; T; G; C. The genome array is co-written in FASTA format [6] and represents text with an alphabet of four nucleotides. The text string contains up to 80 characters. Consider the genome array as text. Put two arrays for comparison. Let's highlight the first row of the array M1 - a vector in the form:

$$Q_{M1}^{1} = q_A^1\ ; q_T^1;\ q_G^1;\ q_C^1 \qquad (4)$$

Where q^i_M is the total value of the selected nucleotide in the array line; the upper index (i) is the row number of the array; subscript (M) is a nucleotide from the set [A, T, G, C]. The vector of the second and subsequent rows (i) of the M_1 array is defined as:

$$R^i_{M1} = q^i_A\,; q^i_T;\ q^i_G;\ q^i_C \quad \text{where } 1 < i < N \tag{5}$$

We define the scalar product of vectors in the form:

$$Q^1_{M1}{}^T * R^i_{M1} = F^i_{M1} \quad \text{where } 1 < i < N \tag{6}$$

By analogy, we create the vector W^1_{M2}, S^i_{M2} for the array M_2

$$W^1_{M2} = q^1_A\,; q^1_T;\ q^1_G;\ q^1_C \tag{7}$$

and

$$S^i_{M2} = q^i_A\,; q^i_T;\ q^i_G;\ q^i_C \quad \text{where } 1 < i < N \tag{8}$$

We define the scalar product of vectors in the form:

$$W^1_{M2}{}^T * S^i_{M2} = H^i_{M2} \quad \text{where } 1 < i < N \tag{9}$$

We state that the scalar product of vectors is a number. A numerical set created from a series of scalar products of vectors is a set of numbers $\{F^i_{M_1}\}$, $\{W^i_{M_2}\}$. Such a numerical set characterizes a series of amplitude values of events created sequentially in each array.

It is obvious that a series of events presented by a track of M_1 array can be correlated to a track of events of M_2 array. We will compare the tracks in metric space with the metric $\boldsymbol{\rho}$:

$$\rho_p = \sqrt{\left(F^i_{M1} - H^i_{M2}\right)^2} \tag{10}$$

Then we have:

1. tautology - arrays are equivalent and compared objects are identical if the condition is met:

$$\rho_p = \sqrt{\left(F^i_{M1} - H^i_{M2}\right)^2} = 0; \quad 1 < i < N \tag{11}$$

2. objects are different if the condition is met:

$$\rho_p = \sqrt{\left(F^i_{M1} - H^i_{M2}\right)^2} \neq 0; \quad 1 < i < N \tag{12}$$

The formalized criteria (11) and (12) provide the necessary and sufficient conditions for matching nucleotide arrays, and hence images of living organisms.

We illustrate the possibility of building an organism image from the kingdom of viruses (Vira). A typical modern classification of viruses is created on the identification of nucleic acids that form the genome. On this basis, RNA viruses and DNA viruses are isolated. Refer to NCBI information source [3].

Let us choose the typical virus Abaca Bunchy top - pathogenic plant virus of the family Nanoviridae [7].

List item and number of species in NCBI source: Viruses (21045); Monodnaviria (2077); Shotokuvirae (1572); Cressdnaviricota (1048); Arfiviricetes (333); Mulpavirales (13); Nanoviridae (13); Babuvirus (3); Abaca bunchy top virus (1). The considered type of virus occupies the first position in the list (root), and therefore is represented in a single instance (figure in parentheses). Consider the dataset for the Abaca bunchy top virus (Table 1).

Table 1. Genomes data at two Abaca bunchy top virus levels

Virus. Kingdom: Viruses; Subgroup: Nanoviridae	Analyzed virus - view	Level in View List	Number of virus instances in the subgroup	METRICS: {**ρ**1; **ρ**2; **ρ**3; **ρ**4; **ρ**5; **ρ**6} / evaluated by expression (3)/
Abaca Bunchy top	Abaca Bunchy top	1 / root /	1	295; 504; 938; 231; 643; 434
Babuvirus	Cardamom bushy dwarf virus	2	3	885; 690; 1377; 372; 641; 808

- For Abaca Bunchy top and Cardamom bushy dwarf virus indicator $\boldsymbol{\rho}_P \neq 0$

Create images of viruses (Fig. 2).

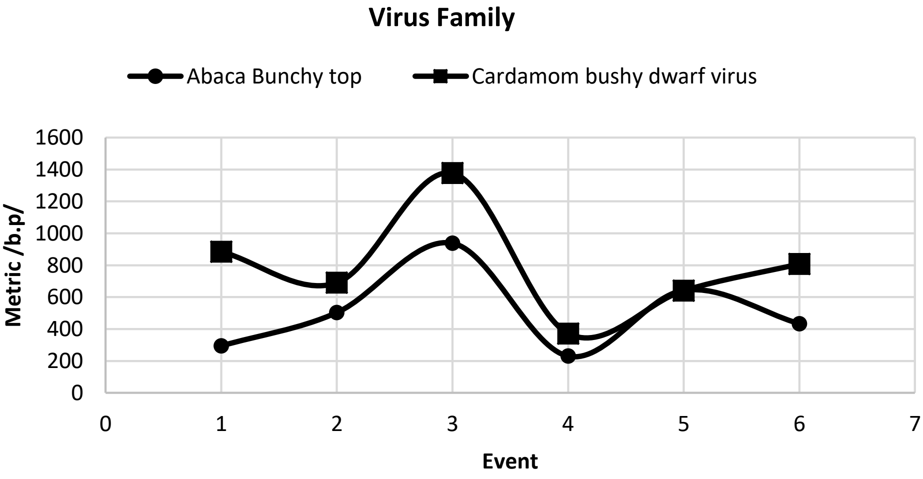

Fig. 2. Abaca Bunchy Top and Cardamom bushy dwarf virus images

We note: two different organisms of viruses, which are closely related, have similarities/similarities of visual images.

Consider another virus, from the source of NCBI: Abutilon mosaic Brazil virus, a pathogenic plant virus of the Geminiviridae family [7].

List item and number of views in NCBI source: Viruses (21044); Monodnaviria (2077); Shotokuvirae (1572); Cressdnaviricota (1048); Repensiviricetes (714); Geplafuvirales (714); Geminiviridae (563); Begomoviral (485); Abutilon mosaic Brazil virus (1). We create a set of indicators for the Abutilon mosaic Brazil virus (Table 2).

Table 2. Genomes data at two levels of Abutilon mosaic Brazil virus

Virus. Kingdom: Viruses; Subgroup: Geminiviridae	Analyzed virus - view	Level in View List	Number of virus instances in the subgroup	METRICS: {$\boldsymbol{\rho}$1; $\boldsymbol{\rho}$2; $\boldsymbol{\rho}$3; $\boldsymbol{\rho}$4; $\boldsymbol{\rho}$5; $\boldsymbol{\rho}$6} / evaluated by expression (3)/
Abutilon mosaic Brazil	Abutilon mosaic Brazil	1 / root /	1	154; 45; 198; 181; 352; 171
Begomovirus	Sweet potato mosaic virus - Brasilia1	2	563	108; 68; 149; 176; 257; 81

- For Abutilon mosaic Brazil and Sweet potato mosaic-Brasilia1 indicator $\boldsymbol{\rho}_{P} \neq 0$

Create images of viruses (Fig. 3).

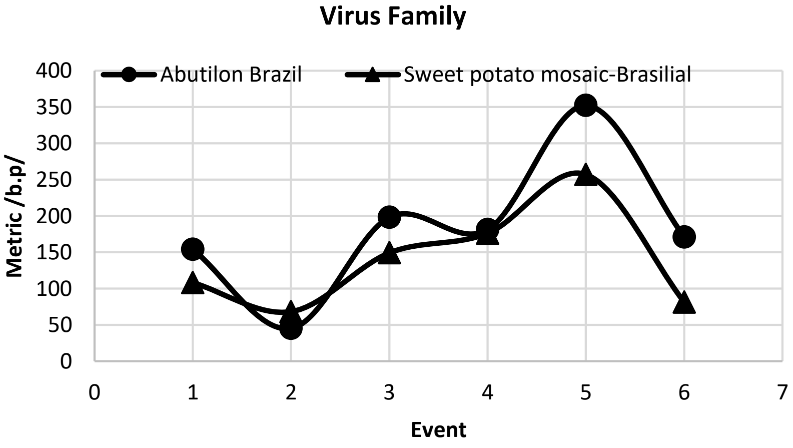

Fig. 3. Abutilon Brazil and Sweet potato mosaic virus images - Brasilial

We note: two different organisms of viruses, which are closely related, have similarities/similarities of visual images.

Modern biological taxonomy, traditionally using many information sources, seeks to create a practical system for classifying living organisms. Mathematical phylogenetics, which implements the principles of systemic analysis of genotype data, allows you to create the basis for a final judgment on the similarity and evolutionary development of different species of organisms. In the materials of the work, this thesis is illustrated by the example of the synthesis of images in two different families of viruses.

The obtained results of synthesis of images of different origin of viruses demonstrate the presence of differences. The manifestation of visual similarity among viruses of the same family is well consistent with the general hypothesis of C. Linnaeus about the similarity of related species. It is meant that they are located on the same evolutionary branch and neighboring levels. Extending these ideas, it should be noted that the similarity of images is only partially manifested. And this also corresponds to the general hypothesis of forming a taxonomic scheme. The further away from the "root species," the less similarity of images will be observed. At levels remote from the root, a large number of isolated branches with different types of viruses should be expected. In the basis of the selected viruses, this number of viruses is several thousand individuals. However, in this case, it is possible to conduct a computer mathematical analysis of the similarity of images of various types of viruses. It is natural to believe that such an analysis will require the use of information technology and the results will be identified by experts. Having the instrumental ability to calculate the compactness of a group of identical individuals makes it possible to enter a formalized definition of the form as the main unit of the taxonomic scheme.

Biological species is a collection of individuals with nucleotide sets, the quantitative indicators of which are reproduced in metric space by a series of images with close metric values.

In metrological terms, such an understanding of the biological species can be considered as a development of the main idea of C. Linnaeus about systematization of living organisms. Declaration of this thesis initializes several projects on digital identification of objects taxonomic scheme. It should be recognized that the generation of an event change track, positioned as an image of an object (virus), is a typical method of constructing a geometric image in a metric space. The creation of a computer image of any species of organisms contributes to the development of phylogenetics.

Thus, using mathematical methods for analyzing a set of nucleotides placed in the public domain of NCBI, the materials of the work present information technology for creating a compact image of a large array of data - the genotype of any living organism.

References / Список литературы

1. *Hadorn E., Werner R.* General zoology, 1989.
2. *Glazko V.I.* Evolution: new about the accident and need "Chemistry and Life - XXI Century". № 10, 2012.
3. *Major E.* Populations, species and evolution, "World," 1974.
4. *Bagotsky S.V.* Revolution in the taxonomy "Chemistry and Life". № 6, 2010.
5. *Viro O.Y., Ivanov O.A.* Elementary topology, 2010.
6. [Electronic Resource]. URL: https://www.ncbi.nlm.nih.gov/genome/gdv/ (date of access: 08.02.2021).
7. [Electronic Resource]. URL: https://www.ncbi.nlm.nih.gov/genome/browse/#!/viruses/ (date of access: 08.02.2021).