


## Article

# 4D-Dynamic Representation of DNA/RNA Sequences: Studies on Genetic Diversity of *Echinococcus multilocularis* in Red Foxes in Poland

Dorota Bielińska-Wąż<sup>1,\*</sup>, Piotr Wąż<sup>2</sup>, Anna Lass<sup>3</sup> and Jacek Karamon<sup>4</sup> <sup>1</sup> Department of Radiological Informatics and Statistics, Medical University of Gdańsk, 80-210 Gdańsk, Poland<sup>2</sup> Department of Nuclear Medicine, Medical University of Gdańsk, 80-210 Gdańsk, Poland;

phwaz@gumed.edu.pl

<sup>3</sup> Department of Tropical Parasitology, Medical University of Gdańsk, 81-519 Gdynia, Poland;

anna.lass@gumed.edu.pl

<sup>4</sup> Department of Parasitology and Invasive Diseases, National Veterinary Research Institute,

24-100 Puławy, Poland; j.karamon@piwet.pulawy.pl

\* Correspondence: djwaz@gumed.edu.pl

**Abstract:** The 4D-Dynamic Representation of DNA/RNA Sequences, an alignment-free bioinformatics method recently developed by us, has been used to study the genetic diversity of *Echinococcus multilocularis* in red foxes in Poland. Sequences of three mitochondrial genes, i.e., NADH dehydrogenase subunit 2 (*nad2*), cytochrome b (*cob*), and cytochrome c oxidase subunit 1 (*cox1*), are analyzed. The sequences are represented by sets of material points in a 4D space, i.e., 4D-dynamic graphs. As a visualization of the sequences, projections of the graphs into 3D space are shown. The differences between 3D graphs corresponding to European, Asian, and American haplotypes are small. Numerical characteristics (*sequence descriptors*) applied in the studies can recognize the differences. The concept of creating descriptors of 4D-dynamic graphs has been borrowed from classical dynamics; these are coordinates of the centers or mass and moments of inertia of 4D-dynamic graphs. Based on these descriptors, classification maps are constructed. The concentrations of points in the maps indicate one Polish haplotype (EmPL9) of Asian origin.

**Keywords:** data analysis; bioinformatics; alignment-free methods; moments of inertia



**Citation:** Bielińska-Wąż, D.; Wąż, P.; Lass, A.; Karamon, J. 4D-Dynamic Representation of DNA/RNA Sequences: Studies on Genetic Diversity of *Echinococcus multilocularis* in Red Foxes in Poland. *Life* **2022**, *12*, 877. <https://doi.org/10.3390/life12060877>

Academic Editor: Koichiro Tamura

Received: 25 April 2022

Accepted: 8 June 2022

Published: 10 June 2022

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Copyright:** © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

Recently, a rapid growth of the experimental data in nucleotide databases can be observed, which stimulated the development of mathematical methods to describe these large and complex objects. One group of approaches is formed by the so-called alignment-free bioinformatics methods. For reviews of alignment-free methods, see [1,2]. They are an alternative to standard, alignment-based sequence analysis approaches, e.g., ClustalW [3], Blast [4], Needleman–Wunsch algorithm [5], or T-Coffee [6]. Alignment-free methods are usually computationally simple and there are no sequence limitations. They are particularly useful for Big Data analysis and research on various aspects of similarity between the biological (DNA, RNA, protein) sequences.

Similarity of complex objects is not unique. Multi-dimensional objects can be similar in one aspect/property and very different if other characteristics are taken into account. Different aspects of similarity may be relevant to different problems. Let us take a model example and consider two different pairs of DNA sequences:

1. G G T T  
G G A A
2. G T G T  
G A G A

In both cases, the similarity value is 50%, but non-zero contributions to the final result come from different positions of G in the sequences. In the first case, G are cumulative at the beginning of the sequences, and in the second one the distributions of G are symmetric. The same results are also obtained if, for example, G is replaced by C. Different structures give the same result in standard alignment methods. The degree of non-uniqueness increases with the lengths of the sequences. Advantages of non-standard (alignment-free) methods include the suitability for Big Data analysis with no restrictions for the sequences, as already mentioned, as well as a variety of derived information. In non-standard methods, we obtain a series of values characterizing different properties of a single sequence. The similarity of these properties can be studied separately using non-standard methods and may be correlated with different biological consequences. Therefore, the creation of new methods is very important to reveal some hidden properties of the sequences.

Similarity/dissimilarity analysis is strictly related to classification studies, which is an interdisciplinary problem [7,8]. For example, we obtained information about different types of objects by examining their similarity in the quality of life research [9,10], or in bioinformatics [11,12].

In bioinformatics, there are many different alignment-free methods. For example, Zhou et al. constructed a complex network for similarity/dissimilarity analysis of DNA sequences [13]. We represented the protein sequence as a set of material points in a 20D space [14]. Saw et al. analyzed the similarity of DNA sequences using the fuzzy integral with a Markov chain [15]. Lichtblau applied frequency chaos game representation and signal processing for genomic sequence comparison [16]. He et al. introduced a numerical representation of a DNA sequence, called the Subsequence Natural Vector, and applied it for HIV-1 subtype classification [17].

A subgroup within *alignment-free* bioinformatics methods is formed by the so-called *Graphical Representations of Biological Sequences*, applicable to both graphical and numerical similarity/dissimilarity analysis of biological sequences. It is not obvious how to represent graphically multidimensional objects in two or three dimensions to reveal the most important features without losing information. A variety of approaches have been developed, bringing together ideas from different fields of science, and each of them focuses on various aspects of similarity. Method names are often associated with some properties or ideas applied to the construction of graphs or numerical characteristics describing the graphs. The first graphical representation methods were based on *walks* in three [18,19] and two [20–22] dimensions. Since then, there has been a dynamic development of the graphical bioinformatics branch observed (for reviews see [23,24]). Let us just mention the last few methods of graphical representation: in the “Spider representation of DNA sequences”, the graphs resemble a spider’s web [25]; in a method called by us “Spectral-dynamic representation of DNA sequences”, the plots resemble atomic, molecular, or stellar spectra composed of sequences of sharp spectral lines [11]. For the numerical characterization of these plots, we applied some ideas used in classical dynamics. Hu et al. applied fractal interpolation in their graphical representation of protein sequences [26]. Graphical representations of protein sequences based on physiochemical properties may be found in works by Mahmoodi-Reihani et al. [27], or by Xie and Zhao [28]. A graphical representation of DNA sequences proposed by Xie et al. is based on trigonometric functions [29]. The 2D graphic representation of the DNA sequence proposed by Liu is based on the horizon lines [30]. Another 2D graphical representation of DNA sequences proposed by Wu et al. is based on variant map [31]. The goal is to create approaches in which both graphs and numerical characteristics, often referred to as *sequence descriptors*, represent a biological sequence in a unique (i.e., *degeneracy-free*) way. The first sequence descriptors related to graphical representation of sequences were designed by Raychaudhury and Nandy [32] and by Randić et al. [33]. Since then, many approaches have been created, e.g., spectral moments in the sequence similarity studies were considered by Agüero-Chapin et al. [34] (for review see [35]).

In the present work, we apply 4D-Dynamic Representations of DNA/RNA Sequences created by us [12]. This is a multidimensional alignment-free bioinformatics method, but it also offers some kind of visualization (for details see subsequent section). We applied this method for a characterization of SARS-CoV-2 and Zika viruses. In the present work, we perform analogous studies on genetic diversity of *Echinococcus multilocularis* in red foxes in Poland. Alveolar echinococcosis is a serious parasitic zoonosis caused by *Echinococcus multilocularis*, Leuckart 1863. *E. multilocularis* was found in Poland in relatively high percentages in red foxes; in some regions, the prevalence reached up to approximately 50% [36]. More than one hundred cases of human alveolar echinococcosis were described before 2013 [37]. The present study is a continuation of our previous work in which the results have been obtained using a standard ClustalW method [38].

## 2. Materials and Methods

In the present studies, we apply the 4D-Dynamic Representation of DNA/RNA Sequences—an alignment-free bioinformatics method proposed by us [12]. In this approach, the DNA/RNA sequence is represented as a set of material points in a 4D space, called the *4D-dynamic graph*. The distribution of the points in the space is characteristic for the sequence. A 4D-dynamic graph is created using a method of shifts (*walk*) starting from the point with coordinates  $(0, 0, 0, 0)$ . The first shift is performed according to the unit vector representing the first nucleobase in the sequence. Starting from the end of this vector, the second shift is performed according to the unit vector representing the second nucleobase in the sequence. The process continues until the last nucleobase in the sequence. At the end of each vector, a material point is located with the mass  $m_i = 1$ . Then, the total mass of the 4D-dynamic graph is the length of the sequence ( $N$ ):

$$N = \sum_{i=1}^N m_i. \quad (1)$$

We represent the nucleobases by the following unit vectors: adenine by the vector  $A = (1, 0, 0, 0)$ , cytosine by  $C = (0, 1, 0, 0)$ , guanine by  $G = (0, 0, 1, 0)$ , and thymine/uracil by  $T/U = (0, 0, 0, 1)$ . The final similarity relations between the sequences are the same for different assignments of particular unit vectors to the nucleobases. Choosing the mass different from 1, the final relative similarity relations also remain the same. The mass of each material point and the unit vectors representing particular nucleobases should be the same for all the sequences.

An example of the construction of the 4D-dynamic graph for a model sequence AUGAC is given in [12].

As a visualization of the 4D-dynamic graphs, we apply their projections into 2D or 3D spaces. For example, if we put  $x_i^1$  and  $x_i^2$  coordinates equal to zero, then we obtain a 2D projection, i.e.,  $x^3x^4$ -graph. The distributions of the material points in the 3D or 2D spaces give some information about the locations of three or two nucleobases along the sequences.

As the numerical characteristics of the 4D-dynamic graphs (*sequence descriptors*), we apply values analogous to the ones used in the classical dynamics. One kind of such sequence descriptors are the coordinates of the center of mass of the 4D-dynamic graph:

$$\mu^k = \frac{\sum_{i=1}^N m_i x_i^k}{\sum_{i=1}^N m_i} = \frac{1}{N} \sum_{i=1}^N x_i^k. \quad (2)$$

$x_i^k$  are the coordinates of the mass  $m_i$  in the 4D space and  $k = 1, 2, 3, 4$ .

Another kind of value analogous to the one used in the classical dynamics is the tensor of the moment of inertia of 4D-dynamic graph. It is given by the matrix

$$\hat{I} = \begin{pmatrix} I_{11} & I_{12} & I_{13} & I_{14} \\ I_{21} & I_{22} & I_{23} & I_{24} \\ I_{31} & I_{32} & I_{33} & I_{34} \\ I_{41} & I_{42} & I_{43} & I_{44} \end{pmatrix} \quad (3)$$

with the elements:

$$I_{jj} = \sum_{i=1}^N m_i \sum_{k=1}^4 [\hat{x}_i^k (1 - \delta_{jk})]^2, \quad (4)$$

$$I_{jk} = I_{kj} = - \sum_{i=1}^N m_i \hat{x}_i^j \hat{x}_i^k, \quad (5)$$

where

$$\delta_{jk} = \begin{cases} 1 & j = k, \\ 0 & j \neq k \end{cases}$$

is the Kronecker delta.  $\hat{x}_i^k$  are the coordinates of  $m_i$  in the Cartesian coordinate system for which the origin has been selected at the center of mass:

$$\hat{x}_i^k = x_i^k - \mu^k. \quad (6)$$

The eigenvalue problem of the tensor of inertia is defined as:

$$\hat{I}\omega_k = I_k\omega_k, \quad k = 1, 2, 3, 4, \quad (7)$$

where  $I_k$  are the eigenvalues and  $\omega_k$  are the eigenvectors. The eigenvalues are obtained by solving the fourth-order secular equation:

$$\det(\hat{I} - I\hat{E}) = 0, \quad (8)$$

where  $\hat{E}$  is  $4 \times 4$  unit matrix. The eigenvalues  $I_k$  are called the *principal moments of inertia*.

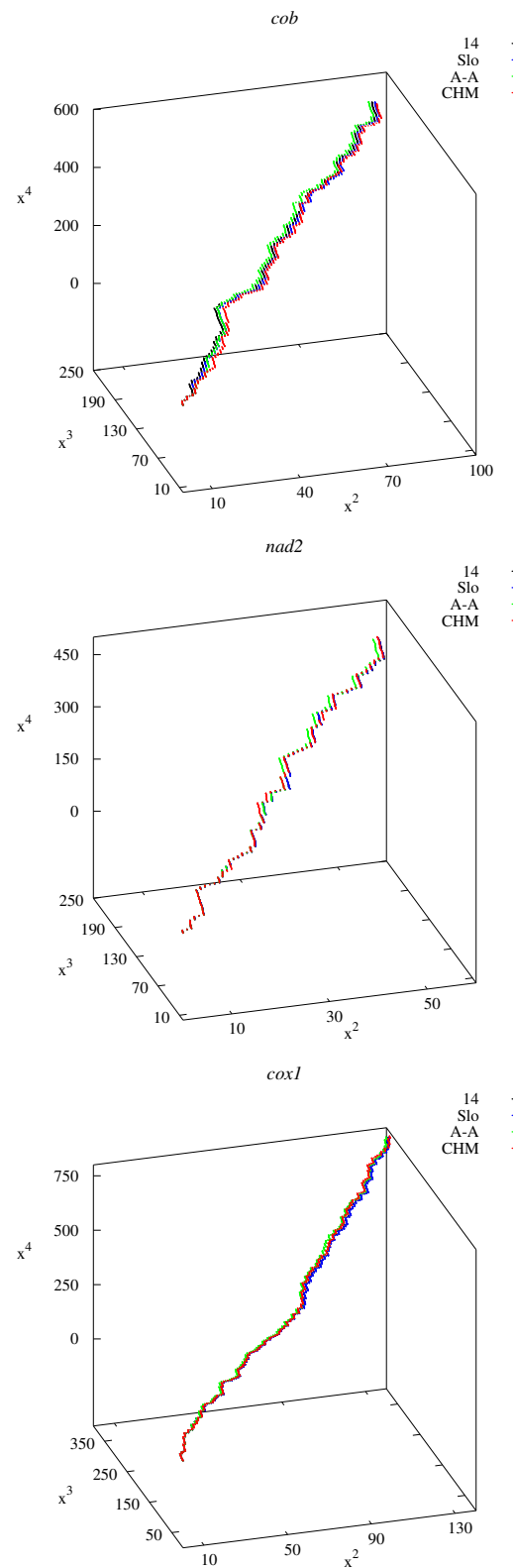
As the sequence descriptors, we apply the normalized principal moments of inertia:

$$r_k^{4D} = \sqrt{\frac{I_k}{N}}, \quad k = 1, 2, 3, 4. \quad (9)$$

The presented method is applied to estimate the genetic diversity of the cestode *Echinococcus multilocularis*, Leuckart 1863, in Poland based on sequence analysis of the mitochondrial genes of worms isolated using the sedimentation and counting technique [39] from the intestines of red foxes *Vulpes vulpes* (Linnaeus). More details concerning the isolation of parasites, sample preparation, polymerase chain reactions (PCRs), and sequencing were described earlier [38]. The nucleotide sequence data used for the calculations are available in GenBank. Sequences of three mitochondrial genes, i.e., NADH dehydrogenase subunit 2 (*nad2*), cytochrome b (*cob*), and cytochrome c oxidase subunit 1 (*cox1*), are analyzed (for the accession numbers see subsequent section) [38,40].

### 3. Results and Discussion

Figure 1 shows examples of projections of the 4D-dynamic graphs to 3D space:  $x^2x^3x^4$ -graphs. The differences between the graphs representing the sequences for different countries (Poland, Slovakia, USA, China) are small. The corresponding principal moments of inertia for all the sequences used in the calculations are shown in Tables 1–6.



**Figure 1.**  $x^2x^3x^4$ -graphs representing *cob* (top panel), *nad2* (middle panel), and *cox1* (bottom panel) genes. Notations: 14—Polish haplotype (sequences No. 14 in Tables 1–3); Slo—Slovakia; A-A—USA, Alaska (St. Lawrence Island); CHM—China (Inner Mongolia).

**Table 1.** Principal moments of inertia of 4D-dynamic graphs representing the *cob* gene for Poland ( $N = 1068$ ).

No.	Accession	Polish Haplotype	$I_1/10^5$	$I_2/10^5$	$I_3/10^5$	$I_4/10^2$
1	KY205662	EmPL1 cob_A	335.9	335.8	335.7	325.5
2	KY205663	EmPL2 cob_G	335.9	335.8	335.7	325.5
3	KY205664	EmPL3 cob_E	336.0	335.9	335.8	325.7
4	KY205665	EmPL4 cob_A	335.9	335.8	335.7	325.5
5	KY205666	EmPL5 cob_D	336.7	336.6	336.5	313.5
6	KY205667	EmPL6 cob_A	335.9	335.8	335.7	325.5
7	KY205668	EmPL7 cob_H	335.4	335.4	335.3	313.6
8	KY205669	EmPL8 cob_A	335.9	335.8	335.7	325.5
9	KY205670	EmPL9 cob_B	338.5	338.4	338.3	341.9
10	KY205671	EmPL10 cob_C	335.9	335.8	335.7	333.7
11	KY205672	EmPL11 cob_F	336.0	335.9	335.8	316.7
12	KY205673	EmPL12 cob_A	335.9	335.8	335.7	325.5
13	KY205674	EmPL13 cob_A	335.9	335.8	335.7	325.5
14	KY205675	EmPL14 cob_J	336.2	336.1	336.0	332.8
15	KY205676	EmPL15 cob_I	335.4	335.4	335.3	313.6

**Table 2.** Principal moments of inertia of 4D-dynamic graphs representing the *nad2* gene for Poland ( $N = 882$ ).

No.	Accession	Polish Haplotype	$I_1/10^5$	$I_2/10^5$	$I_3/10^5$	$I_4/10^2$
1	KY205692	EmPL1 nad_A	207.7	207.7	207.6	181.5
2	KY205693	EmPL2 nad_A	207.7	207.7	207.6	181.5
3	KY205694	EmPL3 nad_D	207.8	207.8	207.7	189.3
4	KY205695	EmPL4 nad_A	207.7	207.7	207.6	181.5
5	KY205696	EmPL5 nad_D	207.8	207.8	207.7	189.3
6	KY205697	EmPL6 nad_A	207.7	207.7	207.6	181.5
7	KY205698	EmPL7 nad_A	207.7	207.7	207.6	181.5
8	KY205699	EmPL8 nad_C	207.7	207.6	207.6	178.2
9	KY205700	EmPL9 nad_B	208.4	208.4	208.3	176.9
10	KY205701	EmPL10 nad_A	207.7	207.7	207.6	181.5
11	KY205702	EmPL11 nad_D	207.8	207.8	207.7	189.3
12	KY205703	EmPL12 nad_C	207.7	207.6	207.6	178.2
13	KY205704	EmPL13 nad_C	207.7	207.6	207.6	178.2
14	KY205705	EmPL14 nad_D	207.8	207.8	207.7	189.3
15	KY205706	EmPL15 nad_A	207.7	207.7	207.6	181.5

**Table 3.** Principal moments of inertia of 4D-dynamic graphs representing the *cox1* gene for Poland ( $N = 1608$ ).

No.	Accession	Polish Haplotype	$I_1/10^6$	$I_2/10^6$	$I_3/10^6$	$I_4/10^3$
1	KY205677	EmPL1 cox_A	117.1	117.1	117.0	132.2
2	KY205678	EmPL2 cox_B	117.1	117.1	117.0	134.1
3	KY205679	EmPL3 cox_B	117.1	117.1	117.0	134.1
4	KY205680	EmPL4 cox_B	117.1	117.1	117.0	134.1
5	KY205681	EmPL5 cox_B	117.1	117.1	117.0	134.1
6	KY205682	EmPL6 cox_C	117.0	117.0	116.9	131.1
7	KY205683	EmPL7 cox_A	117.1	117.1	117.0	132.2
8	KY205684	EmPL8 cox_D	117.1	117.0	117.0	134.6
9	KY205685	EmPL9 cox_E	117.3	117.3	117.2	140.4
10	KY205686	EmPL10 cox_B	117.1	117.1	117.0	134.1
11	KY205687	EmPL11 cox_B	117.1	117.1	117.0	134.1
12	KY205688	EmPL12 cox_B	117.1	117.1	117.0	134.1
13	KY205689	EmPL13 cox_F	117.2	117.2	117.1	138.9
14	KY205690	EmPL14 cox_G	117.1	117.1	117.0	134.1
15	KY205691	EmPL15 cox_B	117.1	117.1	117.0	134.1

**Table 4.** Principal moments of inertia of 4D-dynamic graphs representing the *cob* gene for different countries ( $N = 1068$ ).

No.	Accession	Country	$I_1/10^5$	$I_2/10^5$	$I_3/10^5$	$I_4/10^2$
1.	AB461395	Austria	335.9	335.8	335.7	325.5
2.	AB461396	France	336.6	336.5	336.4	313.2
3.	AB461397	Slovakia	336.0	335.9	335.8	325.7
4.	AB461398	Kazakhstan	338.5	338.4	338.3	341.9
5.	AB461399	Japan (Hokkaido)	338.6	338.5	338.4	359.2
6.	AB461400	USA (Alaska)	338.4	338.4	338.3	328.7
7.	AB461401	USA (Indiana)	342.1	342.1	342.0	281.8
8.	AB461402	China (Mongolia)	338.5	338.5	338.4	298.7

**Table 5.** Principal moments of inertia of 4D-dynamic graphs representing the *nad2* gene for different countries ( $N = 882$ ).

No.	Accession	Country	$I_1/10^5$	$I_2/10^5$	$I_3/10^5$	$I_4/10^2$
1.	AB461403	Austria	207.7	207.7	207.6	164.6
2.	AB461404	France	207.8	207.7	207.7	167.3
3.	AB461405	Slovakia	207.9	207.8	207.8	174.7
4.	AB461406	Kazakhstan	208.4	208.4	208.3	163.2
5.	AB461407	Japan (Hokkaido)	208.9	208.9	208.8	171.8
6.	AB461408	China (Sichuan)	209.0	208.9	208.8	162.1
7.	AB461409	USA (Alaska)	206.8	206.8	206.7	157.7
8.	AB461410	USA (Indiana)	207.7	207.6	207.6	161.7
9.	AB461411	China (Mongolia)	208.0	208.0	207.9	146.7

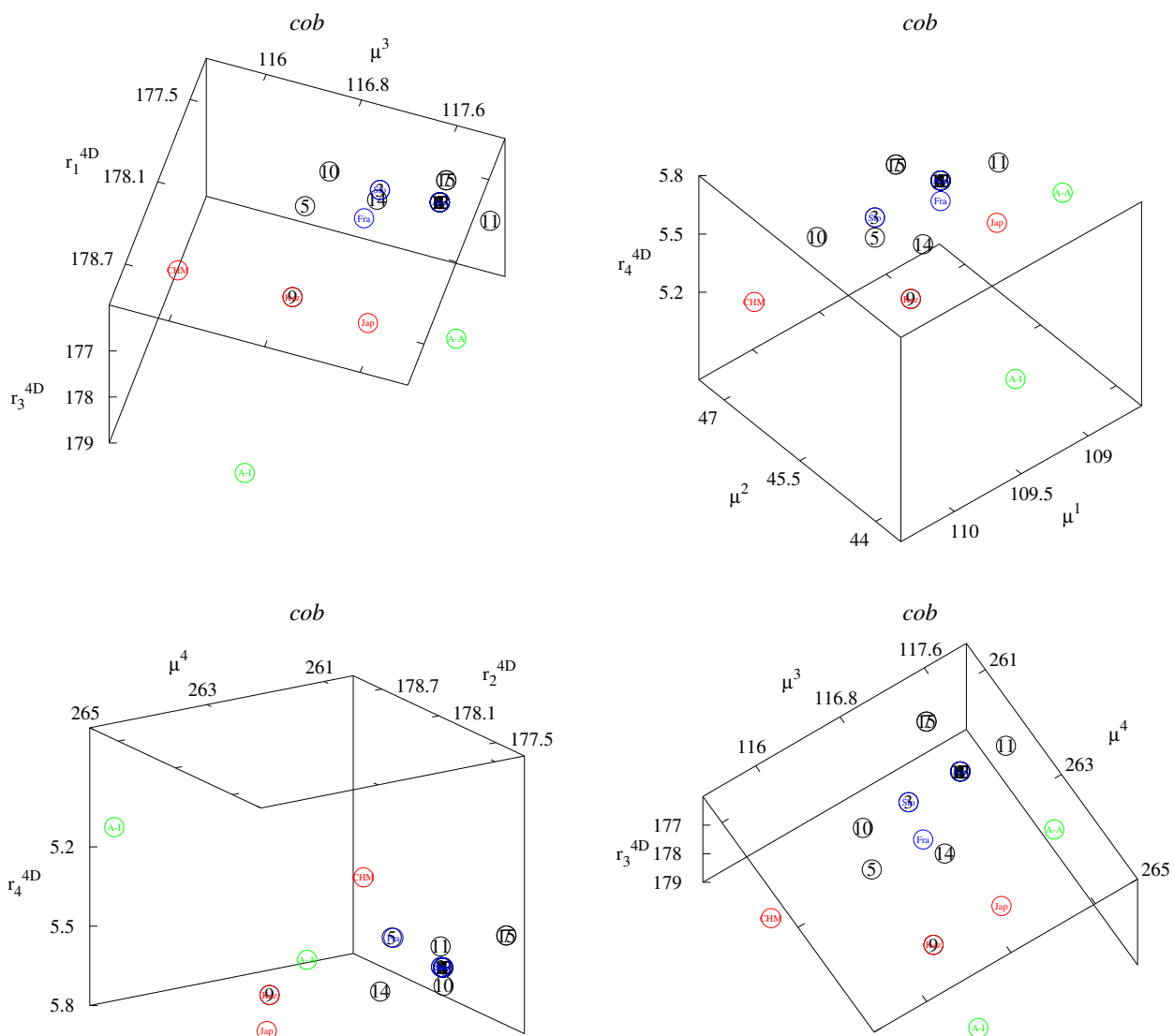
**Table 6.** Principal moments of inertia of 4D-dynamic graphs representing the *cox1* gene for different countries ( $N = 1608$ ).

No.	Accession	Country	$I_1/10^6$	$I_2/10^6$	$I_3/10^6$	$I_4/10^3$
1.	AB461412	Austria	117.0	117.0	116.9	139.0
2.	AB461413	France	117.3	117.3	117.2	134.3
3.	AB461414	Slovakia	117.1	117.1	117.0	134.1
4.	AB461415	Kazakhstan	117.4	117.4	117.3	142.8
5.	AB461416	Japan (Hokkaido)	117.3	117.3	117.2	138.8
6.	AB461417	China (Sichuan)	117.3	117.3	117.2	140.4
7.	AB461418	USA (Alaska)	117.3	117.3	117.2	145.2
8.	AB461419	USA (Indiana)	117.4	117.4	117.3	146.2
9.	AB461420	China (Mongolia)	117.1	117.1	117.0	137.9

In our previous work, combined sequence analysis of three genes (*cob*, *nad2*, *cox1*) exhibited fifteen Polish haplotypes (EmPL1–EmPL15). Separate analyzes within individual genes showed less differentiation. The number of haplotypes is smaller for *cob*, *nad2*, and *cox1* genes. They are denoted by the letters A–J (see Tables 1–3) [38]. As a consequence, in some cases, the sequence descriptors are the same. For example, the descriptors of sequences No. 1 and No. 7 in Table 3 (haplotypes A) are the same. All the values for particular genes are similar. For example, the principal moments of inertia are similar for sequences No. 6 (EmPL6 *cox\_C*) and No. 7 (EmPL7 *cox\_A*) (Table 3). They are equal to 117.0, 117.0, 116.9, and 131.1 for sequence No. 6 and to 117.1, 117.1, 117.0, and 132.2 for sequence No. 7.

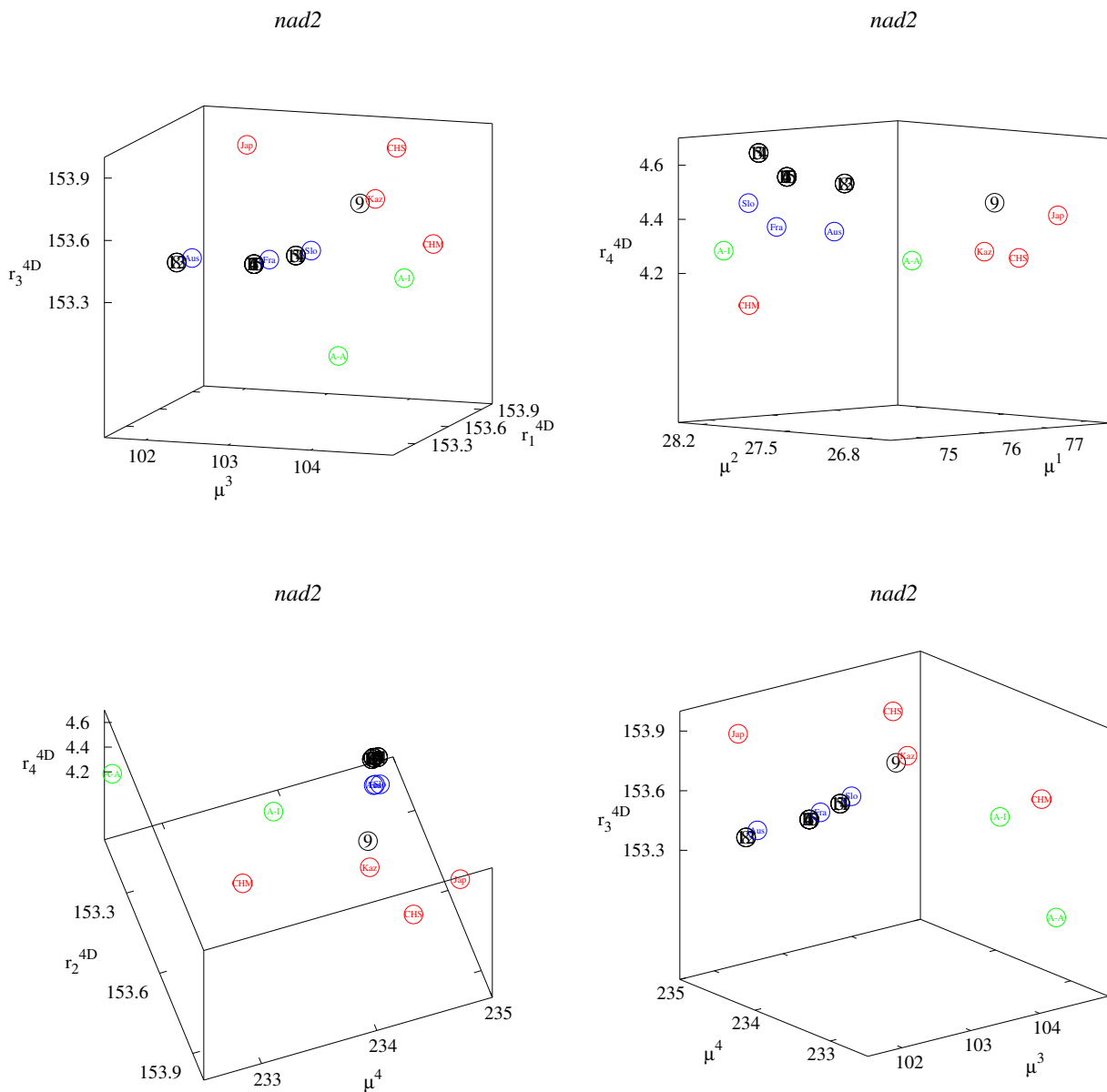
These small differences can be better observed in the classification maps (Figures 2–6). Figures 2–4 show  $r_k^{4D} - r_l^{4D} - \mu^m$  and  $\mu^k - \mu^l - r_m^{4D}$  classification maps. Figure 2 represents the *cob* gene, Figure 3 represents the *nad2* gene, and Figure 4 represents the *cox1* gene. Figure 5 shows  $\mu^1 - \mu^2 - \mu^4$  classification maps for all three genes. Figure 6 shows  $\mu^2 - \mu^3 - \mu^4$  also for all three genes. The points in the maps corresponding to fourteen Polish haplotypes EmPL1, EmPL2, . . . EmPL8 and EmPL10, EmPL11, . . . EmPL15 are concentrated

close to the ones representing European clades. Several Polish haplotypes nearly overlap with some European clades (for example with Austria in Figure 2 or with Slovakia in Figure 4). The exception is the Polish haplotype EmPL9. The points representing this sequence are concentrated close to the points representing Asian clades. In particular, Kazakhstan is the closest point to EmPL9 in: Figure 2 (all panels), Figure 3 (all panels), Figure 5 (panels top, middle), and Figure 6 (panels top, middle). This means that the largest similarities between EMPL9 and Kazakhstan are observed for *cob* and *nad2* genes in all the aspects considered. Figure 4, Figure 5 (bottom panel), and Figure 6 (bottom panel) show the classification maps for the *cox1* gene. In these cases, China (Sichuan) and Japan (Hokkaido) are the closest points to EMPL9.



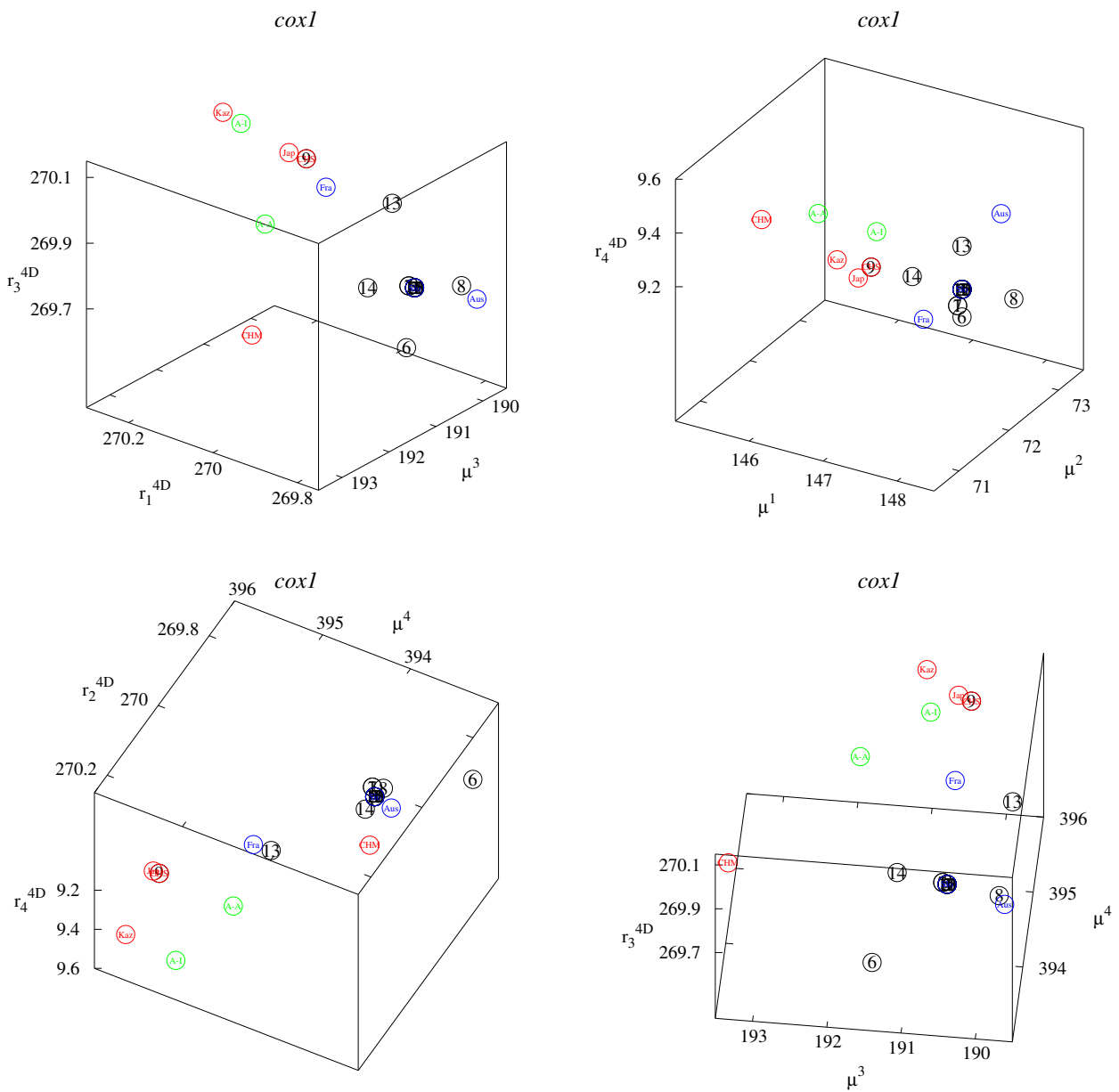
**Figure 2.** Classification maps for *cob* gene:  $r_k^{AD} - r_l^{AD} - \mu^m$  (left panel) and  $\mu^k - \mu^l - r_m^{AD}$  (right panel);  $k, l, m = 1, 2, 3, 4$ . Colors: blue—Europe excluding Poland; red—Asia; green—America; black—Poland. Detailed notations: 1, 2, . . . 15—Polish haplotypes (Table 1); A-A—USA, Alaska (St. Lawrence Island); A-I—USA, Indiana; Aus—Austria; CHM—China (Inner Mongolia); Fra—France; Jap—Japan (Hokkaido); Kaz—Kazakhstan; Slo—Slovakia.





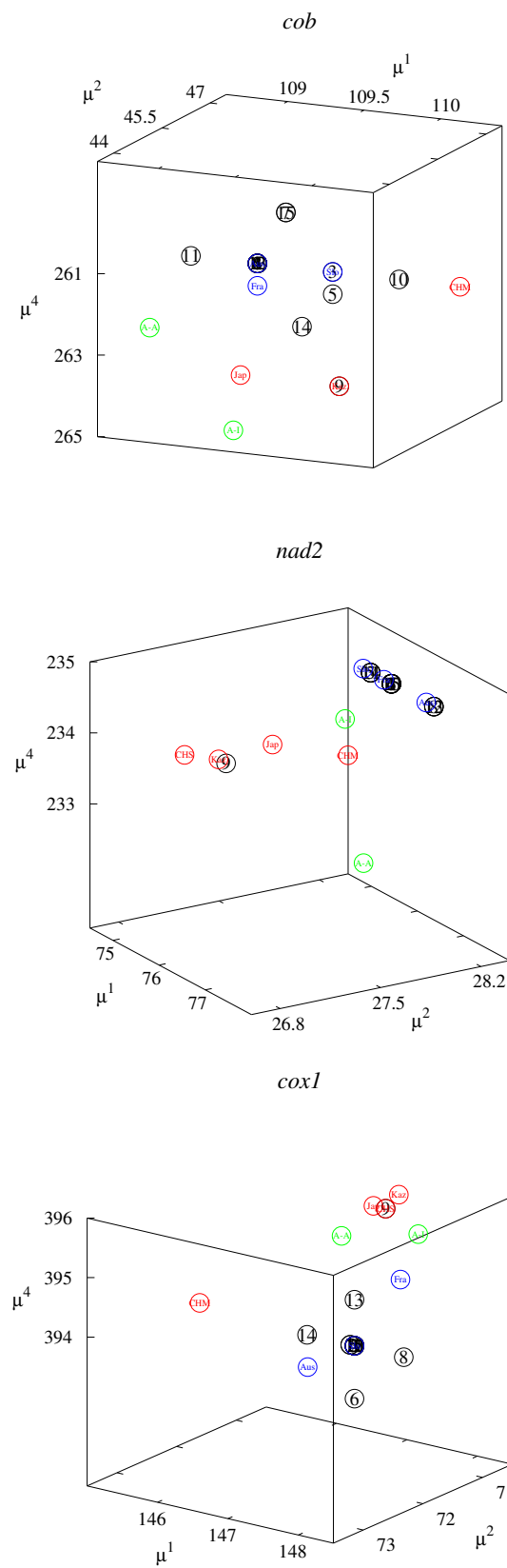
**Figure 3.** Classification maps for *nad2* gene:  $r_k^{AD} - r_l^{AD} - \mu^m$  (left panel) and  $\mu^k - \mu^l - r_m^{AD}$  (right panel);  $k, l, m = 1, 2, 3, 4$ . Colors: blue—Europe excluding Poland; red—Asia; green—America; black—Poland. Detailed notations: 1, 2, . . . 15—Polish haplotypes (Table 2); A-A—USA, Alaska (St. Lawrence Island); A-I—USA (Indiana); Aus—Austria; CHM—China (Inner Mongolia); CHS—China (Sichuan); Fra—France; Jap—Japan (Hokkaido); Kaz—Kazakhstan; Slo—Slovakia.

The results coming from our method can be also presented in a form similar to phylogenetic trees of the standard methods. Figure 7 shows cluster dendrogram for the *cob* gene using  $r_3^{AD}$ ,  $r_1^{AD}$ ,  $\mu^3$ , and the Euclidean distance measure. This dendrogram is another representation of the results of the calculations shown in the top left panel of Figure 2.



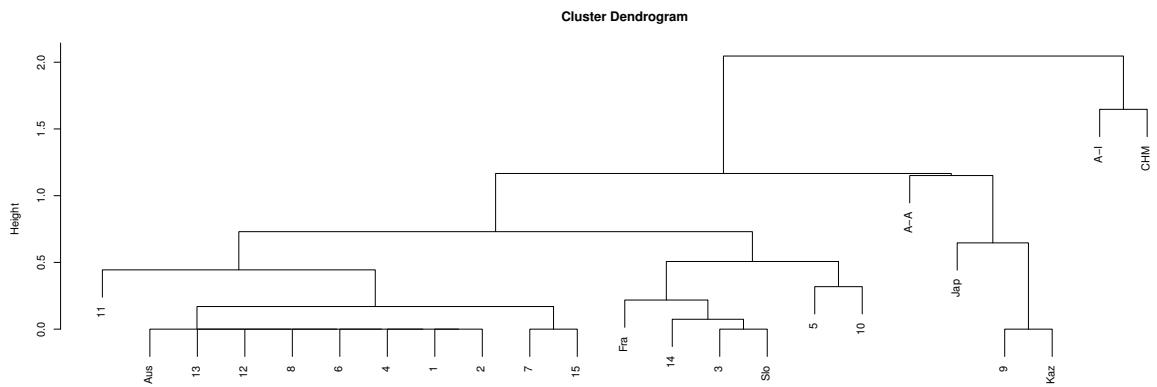
**Figure 4.** Classification maps for the *cox1* gene:  $r_k^{AD} - r_l^{AD} - \mu^m$  (left panel) and  $\mu^k - \mu^l - r_m^{AD}$  (right panel);  $k, l, m = 1, 2, 3, 4$ . Colors: blue—Europe excluding Poland; red—Asia; green—America; black—Poland. Detailed notations: 1, 2, . . . 15—Polish haplotypes (Table 3); A-A—USA, Alaska (St. Lawrence Island); A-I—USA, Indiana; Aus—Austria; CHM—China (Inner Mongolia); CHS—China (Sichuan); Fra—France; Jap—Japan (Hokkaido); Kaz—Kazakhstan; Slo—Slovakia.

The method has no restriction as far as the lengths of the sequences are concerned. Within this method, it is also possible to compensate the information coming from three genes separately into one sequence. Figure 8 shows the  $x^2x^3x^4$ -graphs for combined long sequences *cob*, *nad2*, and *cox1* genes. The same four examples (14; Slo; A-A; CHM) are displayed as in Figure 1. Analogous calculations of the descriptors, as the ones shown in Figures 2–7, can be performed for these concatenated data from the three mitochondrial genes.

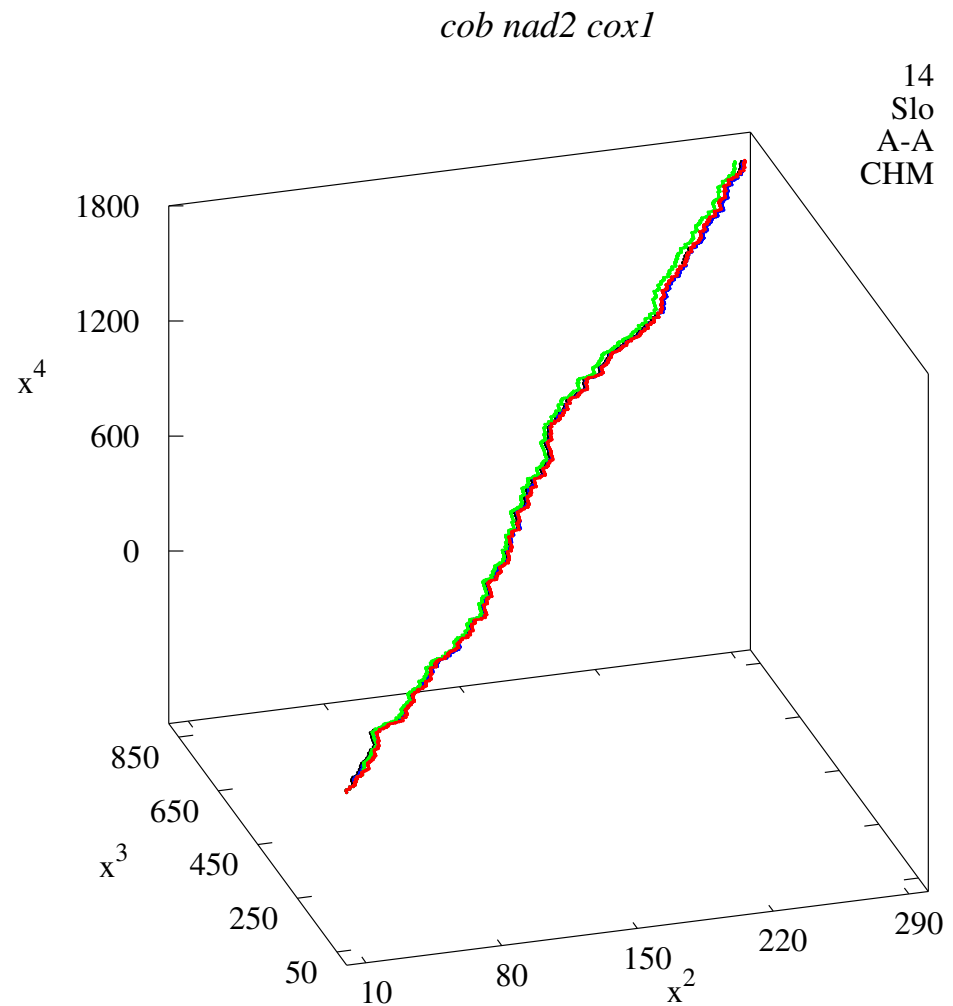


**Figure 5.** Classification maps  $\mu^1 - \mu^2 - \mu^4$  for *cob* (top panel), *nad2* (middle panel), and *cox1* (bottom panel) genes. The colors and the detailed notations are the same as in Figures 2–4.





**Figure 7.** Cluster dendrogram obtained using Euclidean distance measure and  $r_3^{4D}$ ,  $r_1^{4D}$ , and  $\mu^3$  for the *cob* gene (top left panel of Figure 2).



**Figure 8.**  $x^2x^3x^4$ -graphs representing *cob nad2 cox1* genes.

#### 4. Conclusions

In the present work, non-standard bioinformatics studies on the genetic diversity of the cestode *Echinococcus multilocularis* in red foxes in Poland are performed. The 4D-Dynamic Representation of DNA/RNA Sequences, an alignment-free method proposed by us, has been applied [12].

Visualization of multidimensional method is restricted, but some aspects (appropriate projections into 3D space) are shown. The sequences corresponding to European, Asian, and American haplotypes are similar to each other, so the corresponding 3D projections nearly overlap [Figure 1 all panels (sequences No. 14 in Tables 1–3; No. 3, 6, and 8 in Table 4; and No. 3, 7, and 9 in Tables 5 and 6), and Figure 8].

We observed much larger differences for coronaviruses in our previous study [12]. Our studies have shown that the distribution of clusters of points which emerged in the classification maps supports the hypothesis that SARS-CoV-2 may have originated in bat and in pangolin [12].

The considered sequence descriptors are sensitive enough to study the differences for *Echinococcus multilocularis*. Our first report based on the standard bioinformatics method indicated one Polish haplotype (EmPL9 found only in northeast Poland) of probable Asian origin [38]. The present studies indicate aspects of similarities (descriptors related to some properties of the sequences represented in the axes of the maps), in which Polish haplotypes are similar to sequences for different countries. By analyzing the clusters of points in the classification maps (Figures 2–6), the Asian origin of one Polish haplotype (EmPL9) is confirmed.

In summary, by choosing the descriptors, we can reveal different properties of the sequences. In particular, the principal moments of inertia (the values used in the classical dynamics) are equal to the moments of inertia associated with the rotations around the principal axes. The moment of inertia of an object around a rotational axis describes how difficult it is to induce the rotation of the object around this axis. If the mass is concentrated far away from the axis, it is difficult to accelerate into spinning fast and the moment of inertia is large. As a consequence, the descriptors based on moments of inertia reflect the concentrations of masses of the 4D-dynamic graphs around the axes. This way, we can compare the shapes of the graphs representing the sequences.

The correct interpretation of biological and medical data strongly depends on the accuracy of the mathematical models used. Because the accuracy of the presented method is very high (the descriptors used in this method can recognize a difference by a single nucleobase in the compared sequences) the medical importance of the presented approach is significant.

An attractive application of this approach in our future research is predicting the development of viral sequences. Building a predictive model can be crucial in dealing with the future epidemics. Pilot calculations for the Zika virus showed that such an approach could be used to describe the time evolution of the viral genome sequences [12].

**Author Contributions:** Conceptualization, D.B.-W., P.W., A.L., and J.K.; methodology, D.B.-W. and P.W.; software, P.W. and D.B.-W., formal analysis, D.B.-W. and P.W.; writing—original draft preparation, D.B.-W.; visualization, P.W. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by National Science Centre, Poland (grant no. 2020/37/B/NZ7/03934).

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The nucleotide sequence data used for the calculations are available in GenBank.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Vinga, S.; Almeida, J. Alignment-free sequence comparison—a review. *Bioinformatics* **2003**, *19*, 513–523. [[CrossRef](#)] [[PubMed](#)]
2. Jin, X.; Jiang, Q.; Chen, Y.; Lee, S.J.; Nie, R.; Yao, S.; Zhou, D.; He, K. Similarity/dissimilarity calculation methods of DNA sequences: A survey. *J. Mol. Graph. Model.* **2017**, *76*, 342–355. [[CrossRef](#)] [[PubMed](#)]
3. Chenna, R.; Sugawara, H.; Koike, T.; Lopez, R.; Gibson, T.J.; Higgins, D.G.; Thompson, J.D. Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Res.* **2003**, *31*, 3497–3500. [[CrossRef](#)] [[PubMed](#)]
4. Altschul, S.; Gish, W.; Miller, W.; Myers, E.; Lipman, D. Basic local alignment search tool. *J. Mol. Biol.* **1990**, *215*, 403–410. [[CrossRef](#)]
5. Needleman, S.B.; Wunsch, C.D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* **1970**, *48*, 443–453. [[CrossRef](#)]
6. Notredame, C.; Higgins, D.G.; Heringa, J. T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.* **2000**, *302*, 205–217. [[CrossRef](#)] [[PubMed](#)]
7. Bielińska, A.; Majkiewicz, M.; Bielińska-Wąż, D.; Wąż, P. Classification Studies in Various Areas of Science. In *Numerical Methods and Applications*; Nikolov, G., Kolkovska, N., Georgiev, K., Eds.; Conference Proceedings NMA 2018, Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2019; Volume 11189, pp. 326–333.
8. Bielińska, A.; Majkiewicz, M.; Wąż, P.; Bielińska-Wąż, D. Mathematical Modeling: Interdisciplinary Similarity Studies. In *Numerical Methods and Applications*; Nikolov, G., Kolkovska, N., Georgiev, K., Eds.; Conference Proceedings NMA 2018, Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2019; Volume 11189, pp. 334–341.
9. Bielińska, A.; Bielińska-Wąż, D.; Wąż, P. Classification Maps in Studies on the Retirement Threshold. *Appl. Sci.* **2020**, *10*, 1282. [[CrossRef](#)]
10. Bielińska, A.; Wąż, P.; Bielińska-Wąż, D. A Computational Model of Similarity Analysis in Quality of Life Research: An Example of Studies in Poland. *Life* **2022**, *12*, 56. [[CrossRef](#)]
11. Bielińska-Wąż, D.; Wąż, P. Spectral-dynamic representation of DNA sequences. *J. Biomed. Inform.* **2017**, *72*, 1–7. [[CrossRef](#)]
12. Bielińska-Wąż, D.; Wąż, P. Non-standard bioinformatics characterization of SARS-CoV-2. *Comput. Biol. Med.* **2021**, *131*, 104247. [[CrossRef](#)]
13. Zhou, J.; Zhong, P.Y.; Zhang, T.H. A Novel Method for Alignment-free DNA Sequence Similarity Analysis Based on the Characterization of Complex Networks. *Evol. Bioinform.* **2016**, *12*, 229–235. [[CrossRef](#)] [[PubMed](#)]
14. Czerniecka, A.; Bielińska-Wąż, D.; Wąż, P.; Clark, T. 20D-dynamic representation of protein sequences. *Genomics* **2016**, *107*, 16–23. [[CrossRef](#)] [[PubMed](#)]
15. Saw, A.K.; Raj, G.; Das, M.; Talukdar, N.C.; Tripathy, B.C.; Nandi, S. Alignment-free method for DNA sequence clustering using Fuzzy integral similarity. *Sci. Rep.* **2019**, *9*, 3753. [[CrossRef](#)] [[PubMed](#)]
16. Lichtblau, D. Alignment-free genomic sequence comparison using FCGR and signal processing. *BMC Bioinformatics* **2019**, *20*, 742. [[CrossRef](#)]
17. He, L.L.; Dong, R.; He, R.L.; Yau, S.S.T. A novel alignment-free method for HIV-1 subtype classification. *Infect. Genet. Evol.* **2020**, *77*, 104080. [[CrossRef](#)]
18. Hamori, E.; Ruskin, J.H. Curves, a novel method of representation of nucleotide series especially suited for long DNA sequences. *J. Biol. Chem.* **1983**, *258*, 1318–1327. [[CrossRef](#)]
19. Hamori, E. Novel DNA sequence representations. *Nature* **1985**, *314*, 585–586. [[CrossRef](#)]
20. Gates, M.A. Simpler DNA sequence representations. *Nature* **1985**, *316*, 219. [[CrossRef](#)]
21. Nandy, A. A new graphical representation and analysis of DNA sequence structure. I: Methodology and application to globin genes. *Curr. Sci.* **1994**, *66*, 309–314.
22. Leong, P.M.; Morgenthaler, S. Random walk and gap plots of DNA sequences. *Comput. Appl. Biosci.* **1995**, *11*, 503–507. [[CrossRef](#)]
23. Randić, M.; Novič, M.; Plavšić, D. Milestones in graphical bioinformatics. *Int. J. Quant. Chem.* **2013**, *113*, 2413–2446. [[CrossRef](#)]
24. Mizuta, S. Graphical Representation of Biological Sequences. In *Bioinformatics in the Era of Post Genomics and Big Data*; Abdurakhmonov, I.Y., Ed.; IntechOpen: London, UK, 2018.
25. Aram, V.; Iranmanesh, A.; Majid, Z. Spider representation of DNA sequences, *J. Comput. Theor. Nanos.* **2014**, *11*, 418–420. [[CrossRef](#)]
26. Hu, H.; Li, Z.; Dong, H.; Zhou, T. Graphical Representation and Similarity Analysis of Protein Sequences Based on Fractal Interpolation. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2017**, *14*, 182–192. [[CrossRef](#)]
27. Mahmoodi-Reihani, M.; Abbasitabar, F.; Zare-Shahabadi, V. A novel graphical representation and similarity analysis of protein sequences based on physicochemical properties. *Phys. A* **2018**, *510*, 477–485. [[CrossRef](#)]
28. Xie, X.L.; Zhao, Y.X. A 2D Non-degeneracy Graphical Representation of Protein Sequence and Its Applications. *Curr. Bioinform.* **2020**, *15*, 758–766. [[CrossRef](#)]
29. Xie, G.S.; Jin, X.B.; Yang, C.L.; Pu, J.X.; Mo, Z.X. Graphical Representation and Similarity Analysis of DNA Sequences Based on Trigonometric Functions. *Acta Biotheor.* **2018**, *66*, 113–133. [[CrossRef](#)] [[PubMed](#)]
30. Liu, H.L. 2D graphical representation of dna sequence based on horizon lines from a probabilistic view. *Biosci. J.* **2018**, *34*, 744–750. [[CrossRef](#)]
31. Wu, R.X.; Liu, W.J.; Mao, Y.Y.; Zheng, J. 2D Graphical Representation of DNA Sequences Based on Variant Map. *IEEE Access* **2020**, *8*, 173755–173765. [[CrossRef](#)]

32. Raychaudhury, C.; Nandy, A. Indexing scheme and similarity measures for macromolecular sequences. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 243–247. [[CrossRef](#)]
33. Randić, M.; Vračko, M.; Nandy, A.; Basak, S.C. On 3-D graphical representation of DNA primary sequences and their numerical characterization. *J. Chem. Inf. Comp. Sci.* **2000**, *40*, 1235–1244. [[CrossRef](#)]
34. Agüero-Chapin, G.; Sánchez-Rodríguez, A.; Hidalgo-Yanes, P.I.; Pérez-Castillo, Y.; Molina-Ruiz, R.; Marchal, K.; Vasconcelos, V.; Antunes, A. An alignment-free approach for eukaryotic ITS2 annotation and phylogenetic inference. *PLoS ONE* **2011**, *6*, e26638. [[CrossRef](#)]
35. Agüero-Chapin, G.; Galpert, D.; Molina-Ruiz, R.; Ancede-Gallardo, E.; Pérez-Machado, G.; De la Riva, G.A.; Antunes, A. Graph Theory-Based Sequence Descriptors as Remote Homology Predictors. *Biomolecules* **2020**, *10*, 26. [[CrossRef](#)]
36. Karamon, J.; Kochanowski, M.; Dąbrowska, J.; Sroka, J.; Różycki, M.; Bilska-Zajac, E.; Cencek, T. Dynamics of *Echinococcus multilocularis* infection in red fox populations with high and low prevalence of this parasite in Poland (2007–2014). *J. Vet. Res.* **2015**, *59*, 213–217.
37. Nahorski, W.L.; Knap, J.P.; Pawłowski, Z.S.; Krawczyk, M.; Polański, J.; Stefaniak, J.; Patkowski, W.; Szostakowska, B.; Pietkiewicz, H.; Grzeszczuk, A.; et al. Human alveolar echinococcosis in Poland: 1990–2011. *PLoS Negl. Trop. Dis.* **2013**, *7*, e1986. [[CrossRef](#)]
38. Karamon, J.; Stojek, K.; Samorek-Pieróg, M.; Bilska-Zajac, E.; Różycki, M.; Chmurzyńska, E.; Sroka, J.; Zdybel, J.; Cencek, T. Genetic diversity of *Echinococcus multilocularis* in red foxes in Poland: The first report of a haplotype of probable Asian origin. *Folia Parasitol.* **2017**, *64*, 007. [[CrossRef](#)] [[PubMed](#)]
39. Hofer, S.; Gloor, S.; Muller, U.; Mathis, A.; Heggin, D.; Deplazes, P. High prevalence of echinococcus multilocularis in urban red foxes (*Vulpes vulpes*) and voles (*Arvicola terrestris*) in the city of Zurich, Switzerland. *Parasitology* **2000**, *120*, 135–142. [[CrossRef](#)] [[PubMed](#)]
40. Nakao, M.; Xiao, N.; Okamoto, M.; Yanagida, T.; Sako, Y.; Ito, A. Geographic pattern of genetic variation in the fox tapeworm *Echinococcus multilocularis*. *Parasitol. Int.* **2009**, *58*, 384–389. [[CrossRef](#)]