System Biology on Mitochondrion Genomes

Michael G.Sadovsky
Institute of computational modelling SB RAS
660036 Krasnoyarsk, Russia
msad@icm.krasn.ru

Natalia A. Zaitseva
Siberian Federal univeristy
Svobodny prosp., 79
660041 Krasnoyarsk, Russia
zaiceva-5@g-service.ru

Yulia A. Putintseva
Siberian Federal univeristy
Svobodny prosp., 79
660041 Krasnoyarsk, Russia
kinommanka5@mail.ru

***Abstract*—Relations between the triplet composition of mitochondria genomes, and the phylogeny of their bearers is considered. It is shown that the genomes are split into several classes in the space of information values of the triplets. The classification exhibits a feasible correlation to the phylogeny attribution of the genomes. The stability of the classification, as well as the impact of various techniques of a data pretreatment is analyzed. A strong and fruitful correlation between the structure of trinucleotide composition of mitochondrion genomes, and the taxonomy of the bearers of these genomes is proven.**

***Keywords*-frequency, entropy, mutual entropy, order, phylogeny, elastic map, knowledge retrieval**

I. INTRODUCTION

A study of statistical properties of nucleotide sequences may bring a lot towards the relation between structure and function encoded in these former. A consistent and comprehensive investigation of the features and peculiarities is based on the study of frequency dictionary of a nucleotide sequence [1], [2], [3]. Such approach answers the questions concerning the statistical and information properties of DNA sequences. A frequency dictionary, whatever one understands for it, is rather multidimensional entity.

In particular, a relation between a structure (i. e., oligonucleotide composition and their frequency), and the taxonomy of the bearers of DNA sequences is of great importance. Here we studied this relation for the set of mitochondrion genomes. They exhibit a significant violation of the second Chargaff's rule. Such violation may provide another opportunity for knowledge retrieval from the statistical properties of them [7], [8].

Consider a continuous symbol sequence from four-letter alphabet $\{\mathsf{A}, \mathsf{C}, \mathsf{G}, \mathsf{T}\}$ of the length N. No other symbols or gaps in a sequence are supposed to take place. Any coherent string $\omega = \nu_1\nu_2 \dots \nu_q$ of the length q is a word. A set of all the words occurred within a sequence makes the support of that latter. Counting the numbers of copies n_ω of the words, one gets a finite dictionary; changing the numbers for the frequency

$$f_\omega = \frac{n_\omega}{N}$$

one gets the frequency dictionary W_q of the thickness q. This is the main object of our study.

Further, we shall study the triplet composition only, i. e., consider the frequency dictionaries W_3. Thus, any genome is represented as a point in 63-dimensional space. What is the pattern of the distribution of genomes in that space, and whether the distribution exhibits a correlation to a phylogeny of the genome bearers are two key questions of our study. Reciprocally, all genomes are known for a symmetry: frequencies of a couple of string composing a complimentary palindrome[1] are pretty close. This proximity is not an absolute equivalence, and the deviation varies for various genomes; mitochondria are well known to be the most variable, from that point of view [1], [4], [5].

To address these questions, we have implemented an unsupervised classification of the mitochondrion genomes, in various spaces of frequencies (or information values) of triplets. Then, the taxa composition of the classes developed due to the classification has been studied; a considerable correlation between taxa composition, and the class occupation was found. Some results of the study of the correlation of the distribution of bacterial taxa in the information value space, developed over 16S RNA are presented in [4], [5].

To study the effects of a structure peculiarities of a genome expressed in the violation of the symmetry mentioned above (for W_3, we developed similar classifications in 32-dimensional symmetrized spaces; the former is the space of differences of the frequencies of two triplets composing a complementary palindrome, and the latter is the space of differences of information values of two triplets composing a complementary palindrome.

The paper presents the evidences of the strong relation between the structure of mitochondrion genomes, and the taxonomy of their bearers. Section II describes the source of the genetic data, Section III provides a short description of the techniques of the classification and knowledge retrieval. The results of the study are present at Section IV, where the subsection IV-A provides the results of the study of the impact of a symmetry violation towards the relation between structure and taxonomy, and the subsection IV-B shows the results proving the high level of interrelation between the

[1]For example, the couples ATC $\leftrightarrow$ GAT and GGCAATC $\leftrightarrow$ GATTGCC are the complementary palindromes. One must bear in mind, that such entities are determined over a single strand!

 ISBN: 978-1-61208-007-9

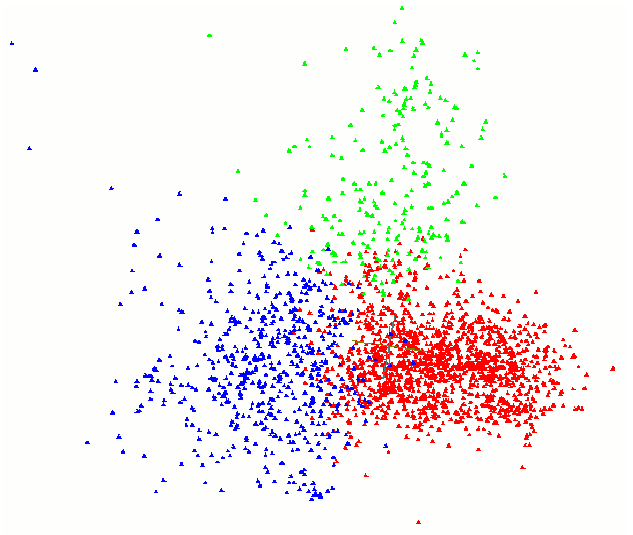

Figure 1. Unsupervised classification developed in 63-dimensional space of frequencies separates mitochondrion genomes into 3 classes; shown in principal components, raw database.

structure, and taxonomy. Finally, the biological issues, as well as some more mathematically oriented questions are discussed in Section V.

II. Databases and methods

Mitochondrion genomes were retrieved from EMBL–bank. The the list of genomes available at EMBL–bank is inhomogeneous, from the point of view of the equity of the number of species of various genders enlisted into the database. The excessive number of closely related genomes (not speaking about the strains and variants) may yield a cluster of increased density[2] that overweights other points at the space, thus resulting in a distortion of the real pattern of a genome distribution at the space of information value of words.

To eliminate the effect of the possible bias described above, we hashed the databases: a single genome from a gender was selected randomly, while the other ones were eliminated from the database. It resulted in a decrease of the number of entries in the database up to 1651 ones.

Another problem in database structure results form an abundant set of entries representing taxonomically rather high clade solely: a single genome is deciphered in a clade. The scattered nature of the database conspires the effects of the correlation between taxonomy and statistical features of the genomes. Thus, we have also hashed the database, excluding the entries which are less than 50 taxons within a class, or an order; finally, 1132 entries was gathered into the database.

[2]And they do, in reality.

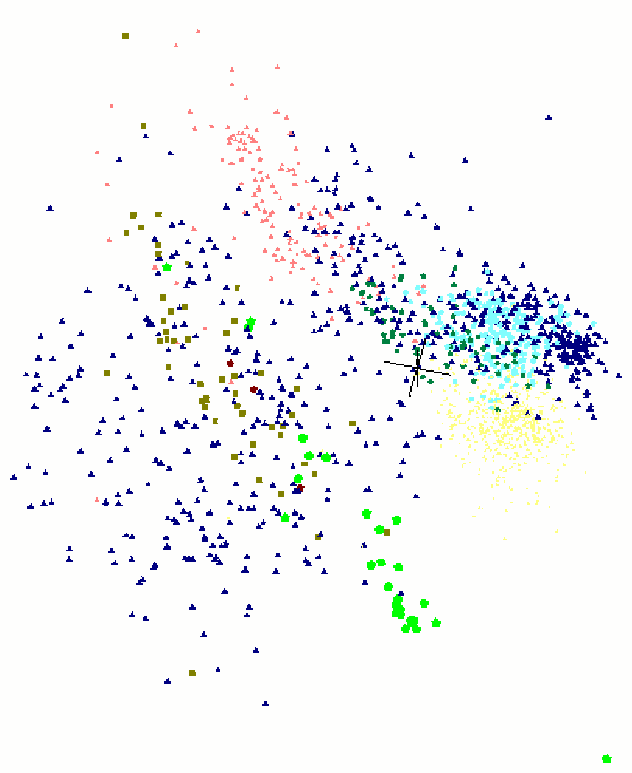

Figure 2. A distribution of seven clades in the space determined by principal components. See text for details.

III. Classification

A standard unsupervised classification technique was implemented to develop a classification of the genomes in the information value space. We used *ViDaExpert* software [9] to do that; the algorithm of the classification development see in [4], [5].

An unsupervised classification does not increases a number of classes: if no separability condition is verified, then the number remains the same. Separability of classes means that two classes are discrete with respect to a relation of a distance between their centroids, and their radii. An excess of a distance between two centroids over the sum of the radii of the relevant classes is the strongest separability condition. On the contrary, one faces the weakest separability condition, if the greater radius (among two classes to be checked for a separability) is not longer than the distance between two centroids of the classes.

We did not check formally the separability conditions, for the classification developed over the genomes. Meanwhile, a stability of the abundance of each class, and the maintenance of the class occupation were checked.

IV. Results and Discussion

Fig. 1 shows an example of an unsupervised classification of the mitochondrion genomes. This is a typical pattern of

 ISBN: 978-1-61208-007-9

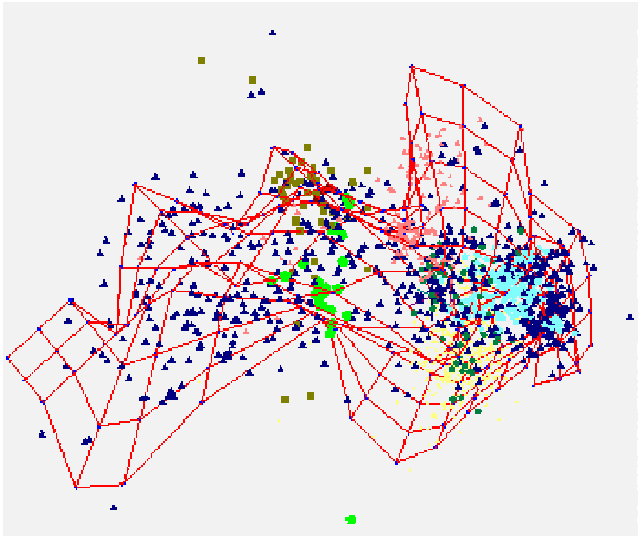

Figure 3. The data distribution around the elastic map (with high rigidity). Seven clades (see text for details) are shown in color.

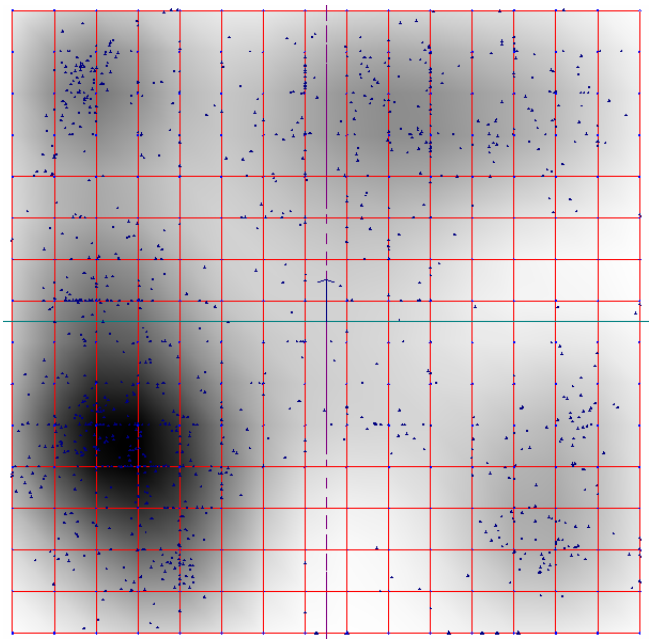

Figure 4. A distribution of 1132 genomes in 63-dimensional space of triplet frequencies. An averaged local density is shown in grey tone.

the separation of the genomes into three classes; namely, in a series of independent experiments with classification development, when a random initial distribution of the genomes takes place, the classification exhibits three outcomes: the genomes could gather into three classes, two classes, and all of them may occupy a single class (with very few exceptions).

Since the classification exhibits not so high instability in the separation of genomes into classes (the separation into

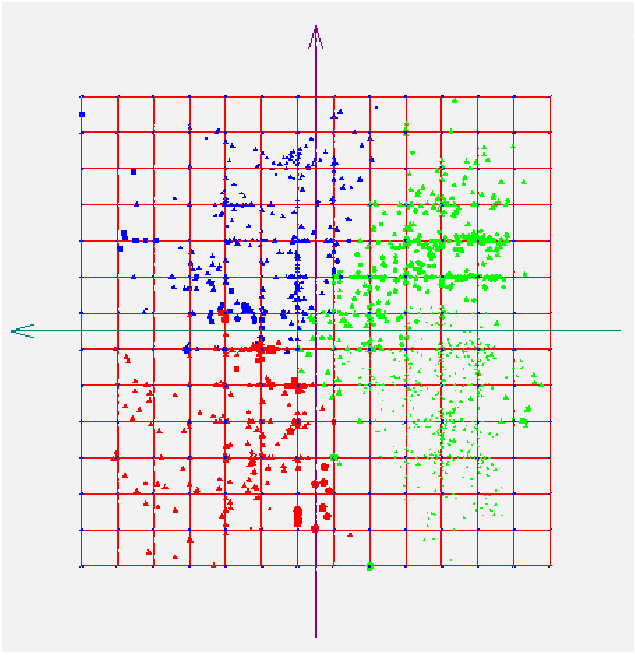

Figure 5. Unsupervised classification of genomes shown on the elastic map (inner coordinates).

three classes seems to be the most stable one), we checked it through a series of experiments. A series of independent classifications were developed, with independent and random redistribution of the genomes into the initial classes. We traced the preference in the class occupation by the same genome; indeed, the greatest majority of the genomes always occupy the same group, if any. Few genomes are quite volatile. They change the class occupation very often, with no regular pattern of the behaviour. Such genomes have been excluded from the database, total number of excluded entities was 41.

A distribution of the clades of various taxonomy level is of great interest; Fig. 2 shows the distribution of seven clades in the space determined by the principal components. This figure shows a distribution of seven (rather different in abundance) clades; moreover, these latter belong to different taxonomy levels. There are three clades of very high taxonomy level: *Fungi* (khaki, 50 entities), *Viridoplantae* (bright green, 31 entities), and *Rhodophyta* (reddish-brown, 3 entities). Four other clades are of lower taxa: fishes (*Actinopterygii*) are shown in yellow (507 entities), *Amphibia* are shown in green-blue (66 entities), and incests (*Neoptera*) are pink (143 entities). Finally, *Mammalia* are shown in light blue (212 entities). Small very dark-blue triangles show all other genomes.

To analyze the patterns of the interdependence of taxonomy and distribution in the space determined by the triplet composition of the genomes, we have implemented elastic map technique; Figure 3 shows such (rigid) map, and the

 ISBN: 978-1-61208-007-9

distribution of the genomes around it. Fig. 4 shows similar elastic map in the inner coordinates; it makes obvious a non-random pattern of the distribution of clades in the space. Moreover, the nonlinear statistical approach (i. e., elastic map implementation) identifies more than three classes: from seven to ten clusters are identified, depending on a rigidity of the map.

An unsupervised classification similar to that one shown in Fig. 1 developed for the distribution of clades projected to the elastic map presented in the inner coordinates is shown in Figure 4. A direct comparison of Figs. 4 and 5 clearly exhibits a character of the distribution of clades over these three classes obtained due to the unsupervised classification implementation. One should avoid a misconduct: the colors in these two figures have nothing to do each other; colors in Fig. 3 represent the clades, while the colors in Fig. 5 represent the classes obtained from the unsupervised classification implementation.

A. Symmetry and asymmetry impact on classification

Complementary palindromes exhibit close figure for the frequencies of the strings composing them. Similar (while less) proximity is observed for information values of the strings. This symmetry has been observed at the very beginning of the decipher of genetic sequences [7]. A violation of the second Chargaff's rule is the example of the asymmetry observed for W_1. Yet, very few attention is paid to the phenomenon of the symmetry [8].

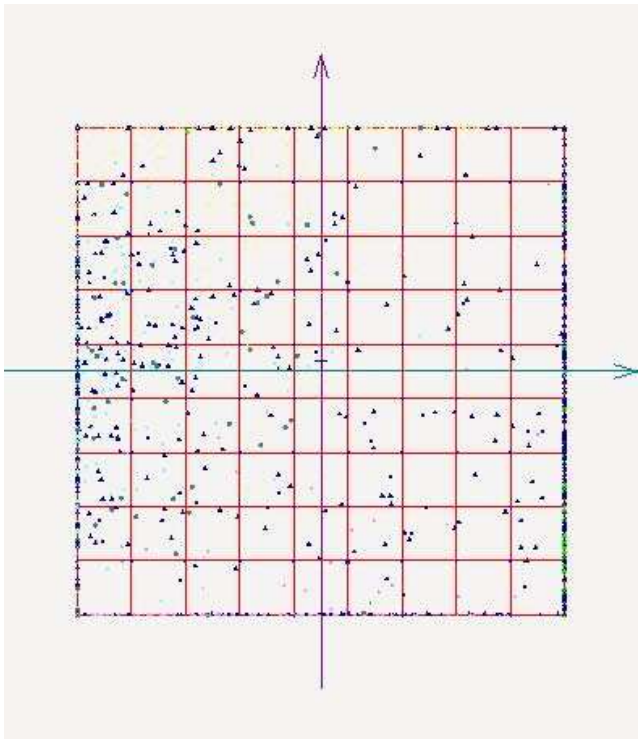

Figure 7. A distribution of 7 clades in 32-dimensional symmetrized space of information values. The distribution of shown on elastic map in internal coordinates.

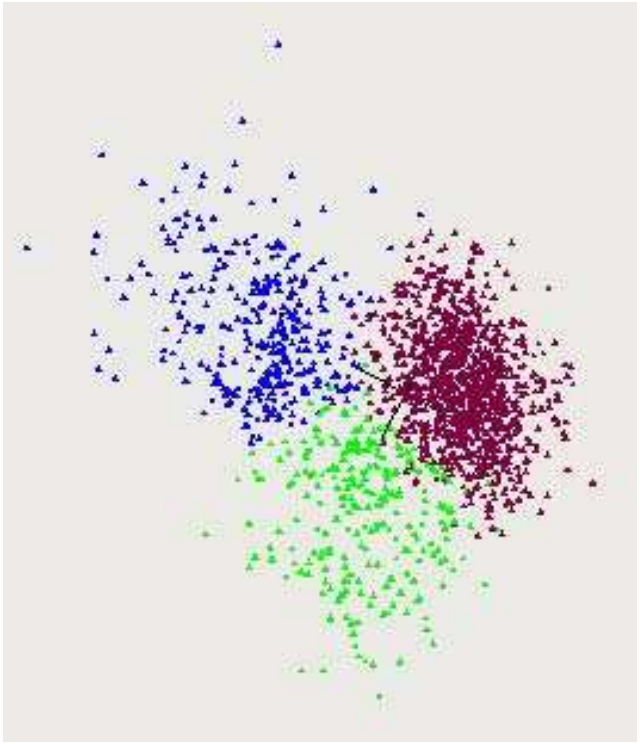

Figure 6. Unsupervised classification of genomes shown in 32-dimensional symmetrized space of information values.

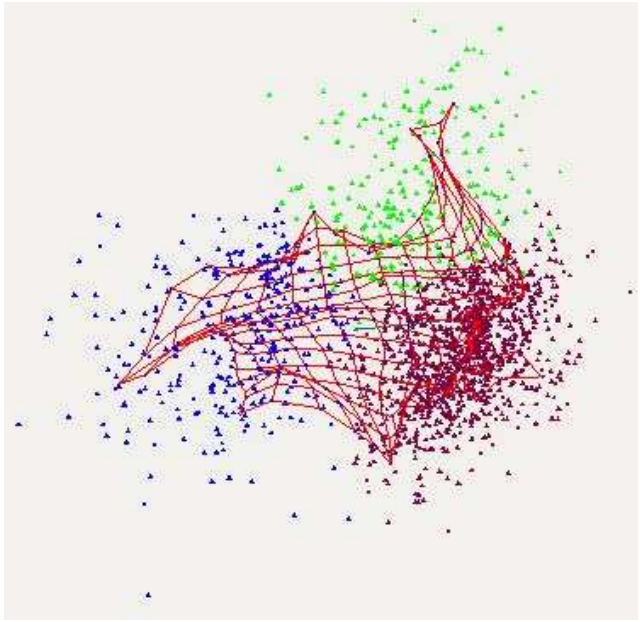

Figure 8. An unsupervised classification shown on elastic map in principal components.

 ISBN: 978-1-61208-007-9

Different genomes exhibit different level of the symmetry violation [8]; mitochondria are known for very high level of the violation. Such effect may bring additional knowledge towards the relation between structure (frequency dictionary) and taxonomy, or a function encoded at genetic entity. To test this idea, we have implemented a series of classifications both in direct, and symmetrized spaces determined by information values of triplets.

Figures 6, 8 and 7 show the distributions similar to those described above, while these latter were developed in the symmetrized 32-dimensional space of information values of triplets. It is evident, that the details of the shape of the distribution, especially the distribution of clades looks rather different. This difference means that the asymmetry in complementary triplets information values play some role in the statistical properties of mitochondrion genomes, while the details of that role are still conspired from a researcher, and await for further investigations.

B. Relation of classification and taxonomy

The relation between the classes obtained due to the separation of the genomes by various statistical techniques, and the taxonomic composition of those classes is the key issue of the paper. In general, the answer on this question is positive.

Fig. 2 shows that taxonomically close entities occupy an obviously isolated subspace, in total genomes distribution. Meanwhile, this figure does not prove an inverse statement. A direct comparison of the composition of the classes shows that there exists a strong correlation in the composition of a class determined statistically, and the taxonomy of its members.

The first class[3] contains genomes of *Actinopterygii* taxon: it contains 457 entries of that taxon, from 506 ones, totally. The second class contains *Neoptera* genomes (137 entries), 8 genomes of *Amphibia* and a genome of *Mammalia*. The third class exhibits the most complicated pattern. It bears 49 *Actinopterygii* genomes, 39 *Amphibia* genomes, the genomes of *Archosauria* and *Lepidosauria*, 97 and 87 entries, correspondingly. Also, this class incorporates almost all mammalian genomes (210 entries) and *Testudines* genomes: 24 entries (all of them are in the class). Besides, the class contains also 4 genomes of insects.

It should be said that the genomes of *Amphibia* are separated between two classes in rather non-random way: the occupation of a class is predominated by the genomes of the same taxon, of lower level. We checked, whether the classification developed over a single clade[4] yields similar pattern, and found that it works. By isolating the genomes of the same class (or other taxonomic level), and developing an unsupervised classification within that latter, one faces the similar behaviour of genetic entities: the statistically determined classes are predominantly occupied by the entities of the same taxon.

[3]Here we understand class as statistically determined cluster.

[4]That must be sufficiently abundant, of course.

V. Conclusion

We addressed the problem of the relation between a structure of DNA sequence, and the sense encoded in that latter. A triplet composition with information values of these latter was considered to be a structure. Genetically, the mitochondrion genomes are rather conserve, thus providing a good raw for knowledge extraction. The distribution of mitochondrion genes in 63-dimensional space of triplets frequency, 32-dimensional space of symmetrized frequencies, and in the relevant spaces determined by information values of triplets is far from a random one.

Statistically (i. e., with no external or additional knowledge, or assumptions), the considered genomes are gathered into several clusters. Linear analysis (unsupervised classification implementation) fails to discrete the genomes properly, providing three clusters, at best. An implementation of some techniques of the nonlinear statistical analysis improves the situation. Self-organized elastic maps clearly identify at least four clusters in the spaces mentioned above.

Reciprocally, the genomes from the came clade always occupy a single cluster; moreover, the clades yield obvious discrete groups within a cluster. It is very important, that the genomes from the same clade never share themselves among two (or more) clusters.

The patterns and relations standing behind the observed distribution of the clades over the clusters is the key issue; a the first glance, there is no simple order in the distribution that could be easily interpreted in the terms of classic systematics or traditional biology. Here are the following options in the studies of the relation between structure and taxonomy (or structure and function encoded in some genetic entities).

Comparative study of the distribution of clades (or functional groups) in frequencies vs. that one in information values: this study aimed to compare and figure out the differences in the composition of the clusters observed in two different spaces may reveal the role and significance of a symmetry observed in triplet composition;

Detailed study of the clade composition of various clusters: this study is the most essential. It is evident, that clades occupy some very peculiar areas in the spaces used to figure out a distribution of genomes. Moreover, there are no clade shared among two (or several) clusters.

Thus, the composition of those clusters is not random, or accidental. The point is that the very different clades occupy the same cluster. What makes them be close each other? Classical biologists consider them to be very far from; meanwhile, the statistics of their genomes seems to be pretty similar. A comprehensive analysis of the triplets that provide the proximity of very far species may reveal

 ISBN: 978-1-61208-007-9

the inner relations in the genetical entities observed in various taxonomy levels. Such relation does not disprove the classical taxonomy; it just provides another dimension for a study of genetic entities and their bearers;

Studying of distributions of the genomes in the space of longer oligonucleotides: this is an obvious expansion of the approach presented above. Here we presented some preliminary results of the comparison of statistical properties of trinucleotide composition of genomes, and the taxonomy of their bearers. Meanwhile, one make pose a question whether the distributions described above are stable against the growth of the strings taken into consideration. Indeed, suppose we changed triplets for octanucleotides; then, what happens with the structure of the clusters described above, and would happen with the distributions?

Exponential growth of the dimension of the relevant space is the main problem here. Indeed, even an abundant database is small enough to exceed the dimension of the space determined by the frequency dictionary W_8. Probably, the most fruitful approach here is the efficient reduction of the space of oligonucleotides. If a greater number of strings have the same frequency, then they make no contribution into the discretion of the genomes, and they might be eliminated from the analysis. In reality, such strings have very proximal, but different frequency. Thus, one needs to identify the strings providing the most significant distinctions of the genetic entities.

A dual problem reveals such strings: one should study the distribution of strings in the "space" of genetic entities, instead of the studying of the distribution of the entities in the space of frequencies. The solution of this dual problem would identify the strings that prove the greatest difference among the genetic entities under consideration. The set of such strings may be rather meager, thus providing a researcher with a good space for direct problem solution. More detailed discussion of all these approaches and options falls beyond this paper.

Another surprising point is that we have studied two, formally speaking, incompatible issues: the former is a proximity between the structures of mitochondrion genomes, and the latter is a taxonomy of the bearers determined morphologically. The point is that the evolution of the bearers, strictly speaking, is quite diverse from the evolution of mitochondrion genomes. The results shown above actually prove that the (co)evolution of mitochondrion genomes and the genetic entities of the bearers runs extremely tightly. This proof could be verified through the study of the relation between taxonomy and statistically defined structuredness of the bearer genomes.

ACKNOWLEDGMENT

The authors are thankful to Andrew Zinovyev for the promoting interest, help in *ViDaExpert* software adoption and valuable discussion.

REFERENCES

[1] M.G. Sadovsky, A.S. Shchepanovsky, and J.A. Putintzeva, *Genes, Information and Sense: Complexity and Knowledge Retrieval* // Theory in Biosciences (2008). V.**127**, pp. 69 – 78.

[2] M.G. Sadovsky, *Comparison of real frequencies of strings vs. the expected ones reveals the information capacity of macromoleculae* // J.of Biol.Physics (2003). V.**29**, pp. 23 – 38.

[3] M.G. Sadovsky, *Information capacity of nucleotide sequences and its applications* // Bulletin of Math.Biology (2006). V.**68**, pp. 156 – 178.

[4] A.N. Gorban, T.G. Popova, M.G. Sadovsky, and D.C. Wunsch, *Information content of the frequency dictionaries, reconstruction, transformation and classification of dictionaries and genetic texts*. *Intelligent Engineering Systems through Artificial Neural Networks*, 11 — *Smart Engineering System Design*, N.-Y.: ASME Press, (2001). pp. 657 – 663.

[5] A.N. Gorban, T.G. Popova, and M.G. Sadovsky, *Classification of symbol sequences over thier frequency dictionaries: towards the connection between structure and natural taxonomy* // Open Systems & Information Dyn. (2000). V.**7**(1), pp. 1 – 17.

[6] N.N. Bugaenko, A.N. Gorban, and M.G. Sadovsky, *Maximum entropy method in analysis of genetic text and measurement of its information content* // Open Systems & Information Dyn. (1998). V.**5**(3), pp. 265 – 278.

[7] G.R. Day, and R.D. Blake *Statistical significance of symmetrical and repetitive segments in DNA* // Nucl. Acids Res. (1982) V.**10**(24), pp. 8323-- 8339.

[8] D. Mitchell, and R. Bridge *A test of Chargaff's second rule* // Biochem. and Biophys. Res. Commun. (2006) V.**340**, pp. 90 – 94.

[9] *http://bioinfo-out.curie.fr/projects/vidaexpert/* March, 2011.

 ISBN: 978-1-61208-007-9