Methods of Revealing Evolutionary Factors Based on the Analysis of Exon-Intron Structure in Genes

Alexander Kaplunovsky

A THESIS SUBMITTED FOR THE DEGREE

"DOCTOR OF PHILOSOPHY"

University of Haifa
Faculty of Natural Sciences
Department of Evolutionary and Environmental Biology

February, 2012

Methods of Revealing Evolutionary Factors

Based on the Analysis of Exon-Intron Structure in Genes

by: Alexander Kaplunovsky

Supervised by: Prof. Alexander Bolshoy
Prof. Eduard Yakubov

A THESIS SUBMITTED FOR THE DEGREE

"DOCTOR OF PHILOSOPHY"

University of Haifa
Faculty of Natural Sciences
Department of Evolutionary and Environmental Biology

February, 2012

Recommended by: ______________________ Date: 27.2.2012

(Advisor)

Recommended by: E. Yakuf ______________________ Date: 28.2.2012

Approved by: ______________________ Date: ____________

(Chairman of PhD Committee)

ACKNOWLEDGEMENT

I would like to express my gratitude to all those who gave me the possibility to complete this thesis.

This dissertation could not have been written without Prof. Alexander Bolshoy from Haifa University who not only served as my supervisor but also encouraged and challenged me throughout my academic program and was abundantly helpful and offered invaluable assistance, support and guidance. I can't say *thank you* enough for his tremendous support and help.

I want to thank the Faculty of Sciences of Holon Institute of Technology giving me the financial support throughout my academic program and its head Prof. Eduard Yakubov for stimulating suggestions and encouragement helped me in all the time of research for and writing of this thesis.

I have furthermore to thank Prof. Anatoly Ivashchenko from Al-Farabi National University of Kazakhstan for his insightful comments and discussion during the development of this work.

Especially, I would like to give my special thanks to my wife Elvira whose patient love enabled me to complete this work.

CONTENT

Methods of Revealing Evolutionary Factors

Based on the Analysis of Exon-Intron Structure in Genes

Alexander Kaplunovsky

Abstract

Length of introns varies from tens to tens of thousands nucleotides. The wide variety of lengths of introns and exons in genomes correlates with some of their functions and may be caused by evolutionary factors.

The goal of this research is to determine the most appropriate approach to classify eukaryotic chromosomes, according to simple exon-intron statistics. The exon-intron structures of eukaryotes genes are quite different from each other, and the evolution of such structures raises many problematical questions. As a preliminary attempt to address some of these questions we performed statistical analysis of gene exon-intron structures. Taking whole genomes of eukaryotes, we went through all the protein-coding genes in each chromosome separately and calculated the portion of intron-containing genes and average values of the net length of all the exons in a gene, the number of the exons, and the average length of an exon. Comparing those chromosomal and genomic averages, we have developed a technique of clustering based on characteristics of the exon-intron structure. This technique of clustering separates different species, grouping them according to eukaryotes taxonomy. Our conclusion is that the best approach is based on distances among four principal components obtained by Factor analysis and following by application of Neighbor Joining clustering algorithm.

1. Introduction

1.1. Exons and Introns

In the 1960s non-bacterial (eukaryotic) ribosomal RNAs (rRNAs) were found to be synthesized as a long precursor RNA which was subsequently processed by the removal of apparently functionless internal "spacer" sequences. A similar processing was found to apply to eukaryotic precursor messenger RNAs (pre-mRNAs; Scherrer *et al.* 1970). In the mid 1970s it was found that the some of the internal sequences interrupted the protein-encoding part of the corresponding mRNAs. The internal sequences, which were removed were named "introns", and what remained in the processed mRNA constituted the "exons".

The discovery of introns and exons occurred independently in 1977 by American molecular biologists Richard Roberts and Phillip Sharp. The two scientists ran experiments which attempted to identify DNA from the resulting mRNA (Gelinas and Roberts, 1977; Donoghue and Sharp, 1977). It was assumed then that the mRNA would have the same base sequence as the DNA from which it was transcribed. This, however, was not the case. Roberts and Sharp found stretches of DNA sequences that were not part of the mRNA. Further, these sequences were interspaced between coding sequences thereby interrupting the code. These data led to the description of exons, the coding DNA, and introns, the interrupting DNA. For their work, Roberts and Sharp shared in the 1993 Nobel Prize in physiology.

In all known living systems coding genes are made up of nucleic acid sequences which are transcribed into pre-mRNAs and translated into proteins. In eukaryotic cells, the interrupted genes may be divided in four different regions including an upstream (regulatory) region, exons, intron(s) and a downstream (stop) region. The proportion of interrupted genes varies with each organism. Simple organisms such as yeasts mainly have intronless genes. In more complex organisms like mammals, almost every gene has at least one intron.

While each gene region plays its role, the exons are the sequences that actually code for proteins. The number of exons that code for a protein vary. Some proteins may have three or four exons but others can have 30 or 40. These differences are found in most species.

The introns are interspaced between the exons in an alternating fashion. Their nucleic acid sequence is highly variable and can be as short as a few dozens bases or as long as a few hundreds. However, the sequences at the beginning and end of the intron are highly conserved. These capping sequences are called splicing junctions and their location defines exon-intron structure.

When a gene is transcribed into RNA, the entire sequence, including the introns, is copied. This primary transcript of RNA is further processed to produce the protein-coding mRNA. In this process, the exons are spliced together by a series of enzymes. First, the ends of exons are brought together. Then the introns are removed and the exons are chemically bonded.

The intron-containing genes are mostly coding for proteins though introns are also found in tRNA and rRNA genes (Singer and Berg, 1991; Lewin, 2000). If introns could be dispensed with in bacteria, then perhaps they had no function. Alternatively, whatever function introns had, either was not necessary in bacteria, or might be achieved in other ways by bacteria. Since members of many bacterial species appeared to be under intense pressure to streamline their genomes to facilitate rapid replication, if it were possible they would have dispensed with any pre-existing introns and/or would have been reluctant to acquire them. On the other hand, if introns played a role and/or did not present too great a selective burden, eukaryotes would have tended to retain pre-existing introns, or could have acquired them (Raible *et al*. 2005).

Knowing the function of introns seemed critical for sorting out these issues. There were many ingenious suggestions. Some thought introns were just another example of the apparently non-utile "junk" DNA which littered the DNA of many eukaryotes. However, some principles to guide investigation of a possible error-checking role were presented, and there is now growing evidence that introns play such a role (Forsdyke, 1981; 1995), although the mechanism may be somewhat different to that originally proposed (Liebovitch *et al*., 1996). It appears that the order of bases in nucleic acids might have been under evolutionary pressure to develop the potential to form stem-loop structures which would facilitate "in-series" or "in-parallel" error-correction by recombination.

In mitochondrial, chloroplast, and bacterial genomes a small number of intron-containing genes is determined and therefore, statistical analysis of intron size is problematic (Umesono *et al.,* 1988; Turmel et al, 1999; Odintsova and Yurina, 2002). The genomes of multicellular eukaryotes contain a substation proportion of introns (Singer and Berg, 1991; Lewin, 2000; Venter, 2001). In fungal, plant, and insect genomes, genes with introns contain also exons widely varying in length. The average intron length significantly increases with increasing size of eukaryotic genome. It is important also to determine whether the length of introns and exons depends on the number of introns per gene, because this fact may have an effect on the splicing time and the rate of gene expression.

1.2. Distribution of exon sizes in protein-coding genes

One of the greatest enigmas of eukaryotic genome evolution is the widespread existence of introns. Introns have been detected in genes of both lower and higher eukaryotes, and also of their viruses, chloroplasts and mitochondria. There are several types of introns, and this study focuses on the most important type: the spliceosomal introns of nuclear-encoded protein genes. We study properties of exon-intron structure of these genes in selected eukaryotic genomes.

A putative link between the biological role of introns and the distribution of exon sizes in protein-coding genes was established soon after intron discovery (Naora and Deacon, 1982). Since then many studies – including statistical analysis – of the exon-intron structures of higher and lower eukaryote genes were performed (Deutsch and Long, 1999; Roy and Penny, 2007; Hawkins, 1988; Kriventseva and Gelfand, 1999; Ivashchenko and Atambayeva, 2004; Atambayeva *et al*., 2008). The problem of intron length variability has a long history (Atambayeva *et al*., 2008; Ivashchenko *et al*., 2009), and it remains unsolved. We still do not know why intron lengths are so widely variable, both between different organisms and between different genes of the same organism.

In our work, we focus on the exon features rather than those of introns. We study relations between the exon lengths, the protein lengths, the average exon sizes, and the numbers of exons per gene (exon densities). There is an interesting observation regarding distributions of exon lengths in different eukaryotes: exon sizes follow a lognormal distribution typical of a random Kolmogorov fractioning process (Gudlaugsdottir *et al*., 2007; Ryabov and Gribskov, 2008). The evolutionary mechanisms of exon-intron structure formation are rather controversial. A theory suggesting that introns appeared as a result of insertion of transposons (Cho and Doolittle, 1997; Roy, 2004) is currently quite popular. Frequently, this point of view implicitly assumes that longer genes possess a higher probability of splitting since they are larger targets for transposons. Some of the present authors have showed (Ivashchenko *et al*., 2009) that the exon–intron organizations in Arabidopsis thaliana, in Caenorhabditis elegans, and in Homo sapiens have much in common. In particular, the net length of all exons in a gene correlates with the number of exons, while the average length of an exon decreases: there are fewer long exons (over 400 nucleotides) and more short exons (80 to 140 nucleotides). This observation seems to support the transposon hypothesis: longer exons appear as larger targets for insertion of mobile elements. Gudlaugsdottir *et al.*, (2007) found some arguments supporting both the intron-early theory (Gilbert, 1987) and the intron-late theory (Cavalier-Smith, 1985; Logsdon and Palmer, 1994) and proposed a mixture model. There is still much controversy and research on newly sequenced genomes should be continued. Here, we apply our efforts mainly for better visualization of new and old results, and application of clustering techniques to strengthen specific genomic properties of common exon-intron organization.

To avoid possible misunderstandings, we would like to clarify our terminology. By "gene" we mean a sequence of DNA nucleotides, which occupies a specific location along a chromosome and determines a particular characteristic in an organism. The structure of a typical protein-coding gene consists of a promoter, a transcription initiation site, a coding region including exons and introns, the polyadenylation signal, and a termination site. Exons are gene fragments that are transcribed in the functional mRNA. All coding sequences are either internal exons or parts of the first or the last exon, while there are non-coding exons, or partially non-coding exons. Introns are non-coding sequences. Some eukaryotic genes have no introns (intronless genes). There are structurally simple genes (two exons separated by one intervening sequence), and there are extremely complex genes whereby a very large number of exons form the final mRNA. For instance, the dystrophin gene comprises at least 70 exons and its length is more than one million base pairs of DNA.

1.3. Genome-specific features of the exon-intron organization in various eukaryotes

Our study focuses on the most important type of introns: the spliceosomal introns of nuclear-encoded protein genes. Here we survey some of the properties of the exon-intron structure of these genes in practically all completely sequenced eukaryotic genomes. Net and averaged exonic lengths are among the attributes considered in this study.

The exon and intron lengths vary within a broad range (Kupfer et. al., 2004; Deutsch and Long 1999; Wendel *et al.*, 2002; Sakharkar *et al.*, 2004; Roy and Penny, 2007). Statistical analyses of the exon and intron lengths were performed several times on different sets of eukaryotes (Naora and Deacon, 1982; Hawkins, 1988; Deutsch and Long

1999; Kriventseva and Gelfand, 1999; Ivashchenko and Atambayeva, 2004; Roy and Penny, 2007; Atambayeva *et al*., 2008; Ivashchenko *et al*., 2009; Kaplunovsky *et al*., 2009-2011).

Previously, we have shown some genome-specific features of the exon-intron organization of eukaryotic genes using a limited set of genomes of different kingdoms (Kaplunovsky *et al*., 2009). We have shown that the most general feature found in all genomes is the positive correlation between the number of introns in a gene and the corresponding protein's length (equivalently, the net length of all the exons of the gene). In addition, we have shown that the average exon length negatively correlates with the average number of exons. Recently, analyses of patterns of exon-intron architecture variation brought Zhu and co-authors to the same conclusions (Zhu *et al*., 2009). One of their main conclusions was a decrease of average exon length as the total exon numbers in a gene increased. While the laws of exon-intron statistics appeared to be quite general, nevertheless, many of the correlation parameters were genome-specific.

Intron density, which is an average number of introns per gene, is an evolutionary riddle as well. At first, it was thought that one can simply predict intron density from organismal complexity. Initial studies supported this hypothesis: *Homo sapiens* had 8.1 introns per gene on average (Collins *et al*., 2004), *Caenorhabditis elegans* – 4.7 (Schwarz *et al*., 2006), *Drosophila melanogaster* – 3.4 (Drysdale and Crosby, 2005), and *Arabidopsis thaliana* – 4.4 (Haas *et. al*., 2005); by contrast, unicellular species were found to have less introns per gene (Logsdon *et al*., 1998). However, further studies found significantly high intron densities in many unicellular species (Archibald *et al*., 2002; Ivashchenko *et al*., 2009), and intron densities in basidiomycetes and zygomycete fungi appeared to be among the highest known for eukaryotes (4-6 per gene), (Loftus *et al*., 2005; Martinez *et al*., 2008). Diversity in intron densities among fungal genomes makes them extremely attractive for exploring possible answers to the questions of exon-intron structure evolution. Indeed, fungi display a wide diversity of gene structures, ranging from far less than one intron per gene for yeasts, to approximately 1–2 introns per gene on average for many recently sequenced lower fungi (including the organisms in this study), and to roughly 5.5 introns per gene on average for some basidiomycetes (e.g., Cryptococcus).

Following the genome sequencing of several lower eukaryotes, it has become possible to examine exon–intron statistics with sufficiently large samples of genes. The purpose of our recent publication (Kaplunovsky *et al*., 2010) was to determine the most appropriate approach to classify fungal chromosomes according to simple exon-intron statistics. We tested a few clustering techniques measuring distances among the chromosomes in different ways. As a result of our analysis, we commented on the consistent similarity of the partitions, which resulted from rather different clustering methods. Clustering results (Kaplunovsky *et al*., 2010) obtained with scaled and normalized Euclidean distances appeared to be sufficiently similar. A Principal Components (PC) based clustering method, the Principal Directions Divisive Partitioning (PDDP) method, and the Neighbor joining (NJ) algorithm produced very similar clustering results. Therefore, we propose techniques of clustering that are able to distinguish between chromosomes of different species with satisfactory success. The addition of regression parameters to averaged chromosomal parameters improved the resolution of clustering.

There is a mixture of different chromosomal characters of exon-intron organization. Here, in this study, similarly to our previous publications, we chose to limit ourselves to consider only pure exonic properties and, additionally, proportions of intron-containing genes among all protein coding genes. We calculated and compared such exonic properties as exon densities, average exon lengths, and average net exon lengths. In this study we would like to check correlation between the number of exons in a gene (exon density) and the corresponding protein's length; to compare intragenomic variation with intergenomic variance of exon densities, average exon lengths, and average net exon lengths; to compare genomic trees obtained with different approaches of clustering based on exonic parameters; and to pave a road to further evolutionary in silico research of exon-intron structure, its origin and development.

2. Data and Methods

2.1. Data set

Nucleotide sequences have been obtained from the database of Eukaryotic Genome Sequencing Projects http://www.ncbi.nlm.nih.gov/genomes/leuks.cgi. Gene annotations were used to calculate genic statistical properties.

A standard gene annotation looks like the following annotation of a randomly chosen gene NCU08052.1 of *Neurospora crassa*

```
gene    <25457..>26451                                            .
mRNA    join(<25457..25690,25755..26055,26117..>26451)            .
CDS     join(25457..25690,25755..26055,26117..26451)              .
```

The annotation means the first exon of this gene starts somewhere upstream of the position 25457, and the last exon of the gene ends somewhere downstream of the position 26451. (Everywhere in this study, when referring to "exons", we mean "coding parts of exons". In other words, only those introns within coding sequences and exons without UTR (untranslated regions) were used for analysis). The data related to coding parts of exons are taken from CDS (coding sequence) lines. For example, the CDS of NCU08052.1 consists of three "exons" [25457:25690], [25755:26055], [26117:26451] with lengths of 234bp, 301bp, and 335bp. The length of the gene is larger than 995 bp, the number of exons is equal to 3, the net length of the exons (the protein size in bp) is equal to 870, and the average exon length is equal to 290.

2.2. Exon-intron structure - statistical parameters

Each gene was assigned three gene-related exonic values: the net length L_{ex} of all its exons, the number N_{ex} of those exons, and an average exon length A_{ex}: $A_{ex} = \frac{L_{ex}}{N_{ex}}$.

For each chromosome of each genome, several absolute and averaged chromosomal characters were calculated. In addition to the three averaged characteristics of exons - the average net length l_{ex} of all the exons in a gene per chromosome, the average number n_{ex} of the exons in a gene per chromosome, and the average exon length a_{ex} per chromosome - the proportion of intron-containing genes (p_c) as a relevant attribute was taken as well. It should be mentioned that a_{ex} is the mean of the A_{ex} values of individual genes per chromosome, $a_{ex} = \frac{1}{n}\sum_{1}^{n} A_{ex}$, where *n* denotes a number of genes in the chromosome here. The measure a_{ex} defined in this is different from the average length $\bar{a}_{ex}$ of all the exons in the chromosome, regardless to which gene(s) they belong. The $\bar{a}_{ex}$, is calculated as the total length of all exons in a chromosome divided by the total number of all exons in a chromosome (see Sakharkar *et al.*, 2004). The a_{ex} usually have significantly larger values than the $\bar{a}_{ex}$ because an average length of *i*-th exon exponentially decreases with an index *i* (Gudlaugsdottir *et al.*, 2007).

We also calculated species-averaged exon parameters: N_g (total number of genes per genome), AN_{ex} (average number of exons in a gene per genome), AL_{ex} (average net length of all exons in a gene per genome), AA_{ex} (average exon length in a gene per genome), $AN1_{ex}$ (average number of exons in an intron-containing gene per genome), $AL0_{ex}$ (average length of an intronless gene per genome), $AL1_{ex}$ (average net length of all exons in an intron-containing gene per genome), and P_g (proportion of intron-containing genes in a genome in percent).

2.3. Distances between pairs of genomes

One of our goals was to cluster genomes using exon-intron structure parameters. We used distance-based methods of clustering; therefore, we had to define a method for distance measuring. The distance between a pair of genomes was calculated as the distance between vectors constructed from several standardized parameters defined above. The vector $\bar{x}_r$ of genomic parameters related to genome *r* consists of (AN_{ex}, AL_{ex}, AA_{ex}, $AN1_{ex}$, $L1_{ex}$, $AL0_{ex}$), and is equal to

$$\bar{x}_r = \left\{ \frac{j_{ex,r} - \mu_j}{\sigma_j} \right\},\ j \in \{AN_{ex}, AL_{ex}, AA_{ex}, AN1_{ex}, L1_{ex}, AL0_{ex}\},$$

where μ_j is the mean value of a genomic parameter *j* and σ_j is its standard deviation.

After having extracted parameters, our next task was to find an appropriate dissimilarity measure *d* such that $d(x_r, x_s)$ is small if and only if x_r and x_s are close. The simplest dissimilarity measure is a normalized (standardized) Euclidean distance:

$$d\ (x_r, x_s) = \sqrt{\sum_{k=1}^{K} (\bar{x}_{r,k} - \bar{x}_{s,k})^2}$$

2.4. Clustering of genomes

A few popular algorithms were used to cluster all 32 genomes. First of all, the well-known Neighbor Joining algorithm (Saitou and Nei, 1987) was used. Using NJ, a tree that does not assume an evolutionary clock was constructed, and therefore, in effect, an unrooted tree results. We used the program *Neighbor* of *Phylip* Package (University of Washington) http://evolution.genetics.washington.edu/phylip/doc/neighbor.html, which is an implementation of NJ. Matrices of standardized distances between all pairs of chromosomes were exported to the program *Neighbor*. The output file was drawn by the program *TreeView* of Prof. Rod Page http://taxonomy.zoology.gla.ac.uk/rod/treeview.html.

2.5. Analyses of the structural-functional organization of the system

One-way ANOVA statistical method was used to test for differences in the exon-intron structure between several groups of species. We also used Factor Analysis (FA) as an integral statistical method, giving the opportunity to define and to evaluate the structural-functional organization of the system. We chose the Principal Components Analysis (PCA) as one of the techniques of FA. The method produces a set of eigenvectors calculated from the matrix of correlations between parameters where each set represents a causal connection of elements. It is important to note that, by using the technique of PCA, all factors become orthogonal and are caused by different properties of the system.

2.6. Factor analysis

One of essential problems of these works is the establishing ***factors***, especially factors identifying major trends of environmental complexity which are based on the analysis of introns and exons of genomes, investigation of relationship between intron number in genes, exon and intron lengths, and a gene density of DNA all chromosomes in different organisms.

Since genome by its very nature is multivariate, it is necessary to analyze this data with multivariate statistical techniques. For identifying major trends and factors will be used a technique of factor analysis, as important statistical instrument of investigation in modern science, being an adequate tool to investigate the principles of interaction of components and their integration into a system (Ahmavaara and Markkonen, 1954; Kaplunovsky, 1971-2007; Bundzen *et. al*, 1975; Verhoog, 1993). This approach to the study of the form of organization, called integratism, proposes the dismembering the system to correlated elements, analyzing their cross-relations and picking out system-forming elements, their relations and hierarchy (Engelgardt, 1970, Verhoog, 1993).

Nevertheless using a method of principal components for the revealing the fundamental, significant and eventually system-forming elements (factors), especially in absence of correlation between them, requests a specific care in the interpretation and the scientific proof (Bartholomew *et al.,* 2002). Under a contemporary tendency of dissemination of the sphere of factor analysis applications, the most serious attention must be paid to the interpretation of results, especially in analyzing the reasons, causing the interrelations of components, having in mind goals and problems of the investigation to order to estimate their corresponding to the obtained factor model (Järveläinen, 1971; Forni and Lippi, 2000).

However should not be forgotten that factor analysis do not always gives a possibility to the pithy interpretation of factors. The interpretation must be based on the data of a nature and properties of elements of the system, obtained by other methods. Factor analysis in this sense is only the link among the other stages of investigation; the connection with these links must be always maintained, and only the whole chain can lead to the solution of a problem. Only the breadth of erudition of researchers, knowledge of principles of the functional integration of investigated systems is able to create a necessary basis for the objective interpretation of revealing factors (Reuchlin, 2003).

Methods of *factor analysis* (principal components) have been used by the author in his works on multi-unit activity of neurophysiology (human brain activity), and a few new developments of factor analysis have been developed (see the separate list of authors publications). Factor analysis has already been used to identify major trends in several fields of genetic (Yee Leng Yap *et al*., 2003; Kliman *et al*., 2005) and we assume that this tool will assist us in searching for further results in this field.

3. Results (published papers)

Following are the thesis results, attached as published papers

3.1. Kaplunovsky, A., Khailenko, V.A., Bolshoy, A., Atambayeva, S.A., and Ivashchenko, A.T. (2009).

Statistics of exon lengths in animals, plants, fungi, and protists.
Proceedings of World Academy of Science, Engineering and Technology **52**: 17-22;
International Journal of Biological and Life Sciences **1** (3): 139-144.

3.2. Kaplunovsky, A., Zabrodsky, D., Volkovich, Z., Ivashchenko, A.T., and Bolshoy, A. (2010).

Statistics of exon lengths in fungi.
The Open Bioinformatics Journal **4**: 31-40.

3.3. Kaplunovsky, A., Ivashchenko, A.T., and Bolshoy, A. (2011).

Statistical analysis of exon lengths in various eukaryotes.
Open Access Bioinformatics **3:** 1-15.

Preface to section 3.1

Kaplunovsky, A., Khailenko, V.A., Bolshoy, A., Atambayeva, S.A., and Ivashchenko, A.T. (2009).

Statistics of exon lengths in animals, plants, fungi, and protists.
Proceedings of World Academy of Science, Engineering and Technology **52**: 17-22;
International Journal of Biological and Life Sciences **1** (3): 139-144.

The main goals of this article are to draw attention to the statistical properties of exon size distributions, and to visualize both the general laws of exon-intron organizations of genes and the genome-specific features. The exon-intron structures of different eukaryotic species are quite different from each other, and the evolution of such structures raises many questions. We try to address some of these questions using statistical analysis of whole genomes. We go through all the protein-coding genes in a genome and study correlations between the net length of all the exons in a gene, the number of the exons, and the average length of an exon. We also take average values of these features for each chromosome and study correlations between those averages on the chromosomal level. Our data show universal features of exon-intron structures common to animals, plants, and protists (specifically, *Arabidopsis thaliana*, *Caenorhabditis elegans*, *Drosophila melanogaster*, *Cryptococcus neoformans*, *Homo sapiens*, *Mus musculus*, *Oryza sativa,* and *Plasmodium falciparum).* We have verified linear correlation between the number of exons in a gene and the length of a protein coded by the gene, while the protein length increases in proportion to the number of exons. On the other hand, the average length of an exon always decreases with the number of exons. Finally, chromosome clustering based on average chromosome properties and parameters of linear regression between the number of exons in a gene and the net length of those exons demonstrates that these average chromosome properties are genome-specific features.

Statistics of Exon Lengths in Animals, Plants, Fungi, and Protists

Alexander Kaplunovsky, Vladimir Khailenko, Alexander Bolshoy, Shara Atambayeva, and Anatoliy Ivashchenko

Abstract**—**Eukaryotic protein-coding genes are interrupted by spliceosomal introns, which are removed from the RNA transcripts before translation into a protein. The exon-intron structures of different eukaryotic species are quite different from each other, and the evolution of such structures raises many questions. We try to address some of these questions using statistical analysis of whole genomes. We go through all the protein-coding genes in a genome and study correlations between the net length of all the exons in a gene, the number of the exons, and the average length of an exon. We also take average values of these features for each chromosome and study correlations between those averages on the chromosomal level. Our data show universal features of exon-intron structures common to animals, plants, and protists (specifically, *Arabidopsis thaliana, Caenorhabditis elegans, Drosophila melanogaster, Cryptococcus neoformans, Homo sapiens, Mus musculus, Oryza sativa,* and *Plasmodium falciparum).* We have verified linear correlation between the number of exons in a gene and the length of a protein coded by the gene, while the protein length increases in proportion to the number of exons. On the other hand, the average length of an exon always decreases with the number of exons. Finally, chromosome clustering based on average chromosome properties and parameters of linear regression between the number of exons in a gene and the net length of those exons demonstrates that these average chromosome properties are genome-specific features.

Keywords**—**Comparative genomics, exon-intron structure, eukaryotic clustering, linear regression.

Abbreviations—N_{ex} = number of exons in a gene; L_{ex} = net length of all exons in a gene; A_{ex} = average exon length in a gene; n_{ex} = average (over a chromosome) number of exons in a gene; l_{ex} = average (over a chromosome) net length of all exons in a gene; a_{ex} = average (over a chromosome) of the average exon length in a gene.

I. Introduction

ONE of the greatest enigmas of eukaryotic genome evolution is the widespread existence of introns. The introns have been detected in genes of both lower and higher eukaryotes, and also of their viruses, chloroplasts and mitochondria. There are several types of introns, and this study focuses on the most important type: the spliceosomal introns of nuclear-encoded protein genes. We study properties of exon-intron structure of these genes in selected eukaryotic genomes.

A putative link between the biological role of introns and the distribution of exon sizes in protein-coding genes was established soon after intron discovery [1]. Since then many studies – including statistical analysis – of the exon-intron structures of higher and lower eukaryote genes were performed [2-9]. The problem of intron length variability has a long history [8, 9], and it remains unsolved. We still do not know why intron lengths are so widely variable, both between different organisms and between different genes of the same organism.

Likewise, we do not understand the distribution of the intron densities (average numbers of introns per gene). At first, the intron density was thought to be related to the organismal complexity. The initial studies supported this hypothesis: *Homo sapiens* has 8.1 introns per gene in average [10], *Caenorhabditis elegans* – 4.7 [11], *Drosophila melanogaster* – 3.4 [12], and *Arabidopsis thaliana* – 4.4 [13]; while, by contrast, unicellular species were found to have less introns per gene [14]. However, further studies found pretty high intron densities in many single-celled species [15, 16], and intron densities in basidiomycete and zygomycete fungi are among the highest known among eukaryotes (4-6 per gene) [17, 18].

In this article, we focus on the exon features rather than those of introns. We study relations between the exon lengths, the protein lengths, the average exon sizes, and the numbers of exons per gene (exon densities). There is an interesting observation regarding distributions of exon lengths in different eukaryotes: exon sizes follow a lognormal distribution typical of a random Kolmogorov fractioning process [19, 20]. The evolutionary mechanisms of exon-intron structure formation are rather controversial. A theory suggesting that introns appeared as a result of insertion of transposons [21, 22] is currently quite popular. Frequently,

A. Kaplunovsky is with the Department of Evolutionary and Environmental Biology and Genome Diversity Center at the Institute of Evolution, University of Haifa, Israel and with the Department of Sciences, Holon Institute of Technology, Holon, Israel.

V. Khailenko is with the Department of Biotechnology, Biochemistry, and Plant Physiology at the Al-Farabi Kazakh National University, Kazakhstan.

A. Bolshoy is with the Department of Evolutionary and Environmental Biology and Genome Diversity Center at the Institute of Evolution, University of Haifa, Israel (corresponding author, phone/fax: +972-48240382; e-mail: bolshoy@research.haifa.ac.il).

S. Atambayeva is with the Department of Biotechnology, Biochemistry, and Plant Physiology at the Al-Farabi Kazakh National University, Kazakhstan.

A. Ivashchenko is with the Department of Biotechnology, Biochemistry, and Plant Physiology at the Al-Farabi Kazakh National University, Kazakhstan (corresponding author, phone/fax: +77272 490 164; e-mail: a_ivashchenko@mail.ru).

this point of view implicitly assumes that longer genes possess a higher probability of splitting since they are larger targets for transposons. Some of the present authors have showed [9] that the exon–intron organizations in *Arabidopsis thaliana,* in *Caenorhabditis elegans*, and in *Homo sapiens* have much in common. In particular, the net length of all exons in a gene correlates with the number of exons, while the average length of an exon decreases: there are fewer long exons (over 400 nucleotides) and more short exons (80 to 140 nucleotides). This observation seems to support the transposon hypothesis: longer exons appear as larger targets for insertion of mobile elements. Gudlaugsdottir *et al.* [19] found some arguments supporting both the intron-early theory [23] and the intron-late theory [24, 25] and proposed a mixture model. There is still much controversy and research on newly sequenced genomes should be continued. Here, we apply our efforts mainly for better visualization of new and old results, and application of clustering techniques to strengthen specific genomic properties of common exon-intron organization.

To avoid possible misunderstandings, we would like to clarify our terminology. By "gene" we mean a sequence of DNA nucleotides, which occupies a specific location along a chromosome and determines a particular characteristic in an organism. The structure of a typical protein-coding gene consists of a promoter, a transcription initiation site, a coding region including exons and introns, the polyadenylation signal, and a termination site. Exons are gene fragments that are transcribed in the functional mRNA. All coding sequences are either internal exons or parts of the first or the last exon, while there are non-coding exons, or partially non-coding exons. Introns are non-coding sequences. Some eukaryotic genes have no introns (intronless genes). There are structurally simple genes (two exons separated by one intervening sequence), and there are extremely complex genes whereby a very large number of exons form the final mRNA. For instance, the dystrophin gene comprises at least 70 exons and its length is more than one million base pairs of DNA.

II. Data and Methods

Nucleotide sequences of 76 chromosomes of 8 species (Table I) containing 5 chromosomes of *Arabidopsis thaliana* (AD), 6 chromosomes of *Caenorhabditis elegans* (CE), 5 chromosomes of *Drosophila melanogaster* (DM), 14 chromosomes of *Cryptococcus neoformans* (CN), 10 chromosomes of *Homo sapiens* (HS), 10 chromosomes of *Mus musculus* (M), 12 chromosomes of *Oryza sativa* (OS), and 14 chromosomes of *Plasmodium falciparum* (PF) have been obtained from GenBank http://www.ncbi.nlm.nih.gov.

Each gene was assigned 3 numbers: the net length L_{ex} of all its exons, the number N_{ex} of those exons, and an average exon length

$$A_{ex} = \frac{L_{ex}}{N_{ex}}$$

Linear regression for the number of exons in a gene as a function of the gene's net exon length $N_{ex}=a+b\cdot L_{ex}$ was performed using the program SPSS for every chromosome. For every chromosome, we also calculated the average net length l_{ex} of all the exons in a gene, the average number n_{ex} of such exons, and the average exon length a_{ex} - which is the mean of the A_{ex} values of individual genes,

$$a_{ex} = \frac{1}{n}\sum_{i=1}^{n} A_{ex}\,,$$

where n is a number of genes in the chromosome. Note that the a_{ex} is different from the average length $\bar{a}_{ex}$ of all the exons in the chromosome, regardless of which gene(s) they belong to. (The $\bar{a}_{ex}$, is calculated as a total length of all exons in a chromosome divided by a total number of all exons in a chromosome, see [26]). The a_{ex} usually have significantly larger values than the $\bar{a}_{ex}$ because an average length of i-th exon exponentially decreases with an i (see [19]).

We also considered regression parameters a and b and a parameter of explained variation R^2. These data are compiled in Supplementary Material. Distance between each pair of chromosomes has based on these six parameters standardized in the interval [−1 ÷ +1], and was calculated by SPSS as a Euclidean distance in a six-dimension space.

A matrix of distances for all 76 chromosomes was exported to the program *Neighbor* of *Phylip Package* (the University of Washington) http://evolution.genetics.washington.edu/phylip/doc/neighbor.html using Neighbor Joining Algorithm. Output file was viewed and drawn by the program *TreeView* of Prof. Rod Page http://taxonomy.zoology.gla.ac.uk/rod/treeview.html.

TABLE I
List of Processed Species and their Chromosomes

N	Name of the organism	Kingdom	Number of chromosomes	Processed chromosomes
1	*Arabidopsis thaliana*	Plant	5	1-5
2	*Caenorhabditis elegans*	Animal	6	1-6
3	*Cryptococcus neoformans*	Fungi	14	1-14
4	*Drosophila melanogaster*	Animal	4+X	2L,2R,3L,3R,X
5	*Homo sapiens*	Animal	22+XY	1-10
6	*Mus musculus*	Animal	19+XY	1-10
7	*Oryza sativa*	Plant	12	1-12
8	*Plasmodium falciparum*	Protists	14	1-14

III. Results and Discussion

A. Average Numbers of Exons and Net Exon Lengths in Different Chromosomes

For each of 76 chromosomes of eight species, we have calculated the average parameters l_{ex} (net length of gene's exons), n_{ex} (number of exons in a gene) and a_{ex} (average exon length). These averages turned out to be pretty similar for different chromosomes of the same species but rather distant for different species. Fig. 1 presents a scatter plot of the l_{ex} *vs* n_{ex}; it shows clear clustering of the chromosomes by species.

It also shows a wide separation between PF – a protist – and the other species (animals, fungi, and plants). The PF chromosomes have much longer average proteins (l_{ex}) and much lower exon density (n_{ex}) than all the other eukaryote chromosomes we have studied. Moreover, all species except PF have rather similar ranges of the l_{ex} parameter, but the n_{ex} fall into quite distinct regions on the plot for the DM (*D. melanogaster*) and CN, and more doubtful areas for plants (AD and OS) and mammals (*H. sapiens* and *M. musculus*).

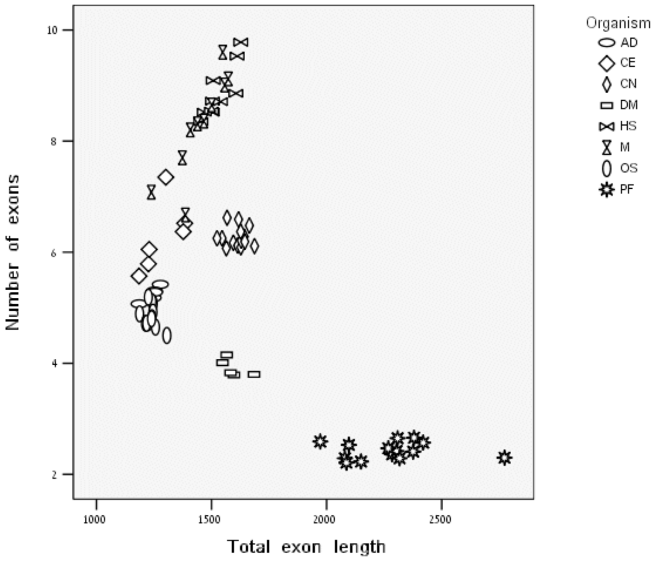

Fig. 1 Scatter-plot of the average net exon length per gene l_{ex} (x-axis) *vs* the average number of exons per gene n_{ex} (y-axis), for all 76 processed chromosomes of eight species

Fig. 2 is a scatter-plot of the average exon length a_{ex} *vs* the average number of exons in a gene n_{ex}; the outlier chromosomes of PF are not shown. This plot shows much better grouping of chromosomes belonging to the same species than Fig. 1 – all kingdoms are grouped separately. Still, the resolution is not sufficient and there is a slight overlapping between species from the same kingdom (M and HS, AD and OS). In addition, *C. elegans* chromosomes may be characterized by relatively short exons in average and rather big variation in intron density. To improve the resolution between the species, we are going to take a closer look at the relation between the average exon number and the average net exon length of a gene.

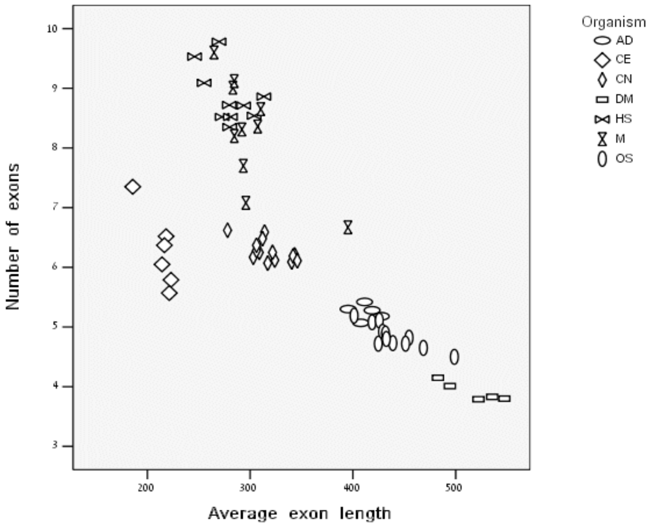

Fig. 2 Scatter-plot of the average exon length a_{ex} (x-axis) *vs* the number of exons n_{ex} (y-axis), for 62 processed chromosomes of seven species

B. Relations between the Average Exon Number and the Average Net Length of Exons in a Gene

It was already shown [8] that the average exon length in *A. thaliana*, *O. sativa*, *C. elegans*, and *Homo sapiens* genes decreases with an increasing number of introns. In addition, positive linear correlation was observed between the sum of exon lengths and the number of exons [8]. Fig. 3 shows the relation between the net length of exons and the number of exons in 12156 genes on ten chromosomes of *H. sapiens*. Parameters of linear regression $N_{ex}=a+b{\cdot}L_{ex}$ are a=1.118 and b = 0.005028. Explained variation of the regression R²=0.666, significance p< 0.001.

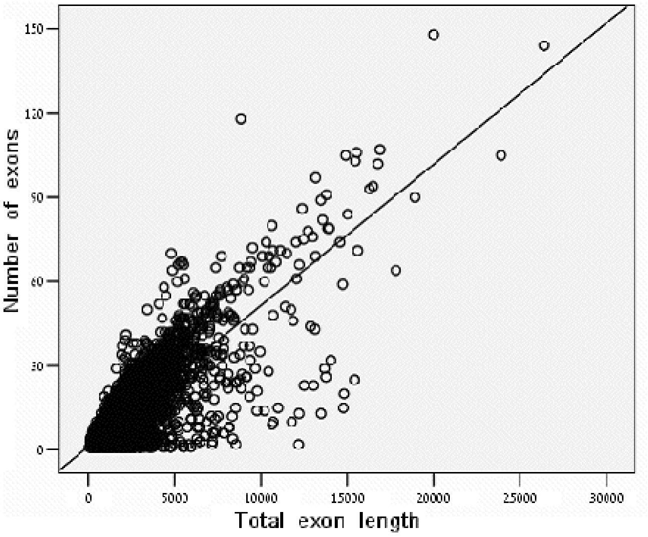

Fig. 3 Linear regression between the net length of exons of a gene (L_{ex}, x-axis) and the number of exons (N_{ex}, y-axis) in genes on all processed chromosomes of *H. sapiens*

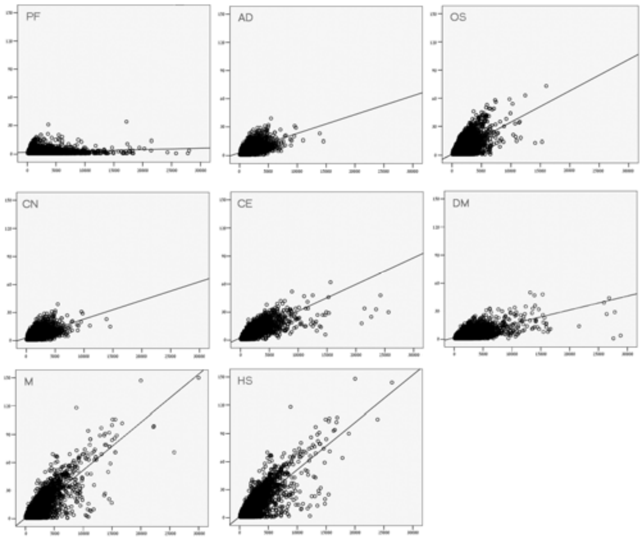

Fig. 4 Scatter plots of L_{ex} (x-axes) *vs* N_{ex} (y-axes) and lines of linear regression for chromosomes of *P. falciparum* (PF), *A. thaliana* (AD), *O. sativa* (OS), *C. neoformans* (CN), *C. elegans* (CE), *D. melanogaster*, *M. musculus* (M), and *H. sapiens* (HS)

Fig. 4 presents similar plots for all eight species. Each species is represented by a scatter-plot of L_{ex} *vs* N_{ex} with a linear regression. There are dramatic differences between average and maximal values of L_{ex} and N_{ex} for animals, plants, fungi, and protists, and especially between parameters *a* and *b* of the linear regression equation $y=a+bx$. In light of these differences, we decided to check if the regression parameters could be used in classification of genomes by their exon properties. We have calculated the linear regressions for all 76 processed chromosomes of all eight genomes. Our results show significant correlations between the protein lengths and the numbers of exons in all eight studied genomes. The Supplementary Material tabulates the parameters *a* and *b* of the linear regression $N_{ex}=a+b\cdot L_{ex}$; their values testify to high reliability of the correlation.

Fig. 5 presents the scatter-plot of parameters *a* and *b* of the linear regression $N_{ex}=a+b\cdot L_{ex}$ for all the processed chromosomes. One can recognize five clusters in the figure: (i) PF, (ii) DM, (iii) plants (AD + OS), (iv) mammals (M + HS), and (v) CE + CN. This means that clustering based on the linear regression parameters *a* and *b* follows the major differences between species from different kingdoms, and some reasonably observable differences between species from the same kingdom. There are some exceptions, and we would like to eliminate them by using the R^2 parameter - percent of the explained variation - of the regression analysis. It has negligible value for protists, medium values for plants and fungi, and relatively high values for animals.

Fig. 6 presents scatter plots for *a vs* R^2 (left) and *b vs* R^2 (right). It shows slightly improved resolution between the species: the CE and the CN chromosomes now belong to separate clusters, while the AD and the OS are almost (but not quite) separate. Hopefully, combining all the parameters together would give a better resolution than looking at any two parameters at a time.

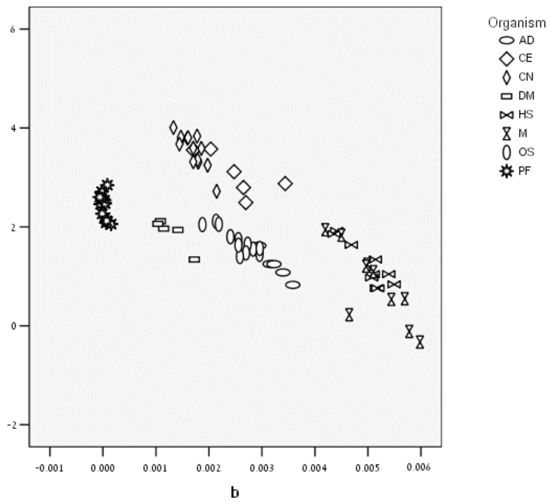

Fig. 5 Scatter-plot of parameters *a* (y-axis) and *b* (x-axis) of the linear regression $N_{ex}=a+b\cdot L_{ex}$

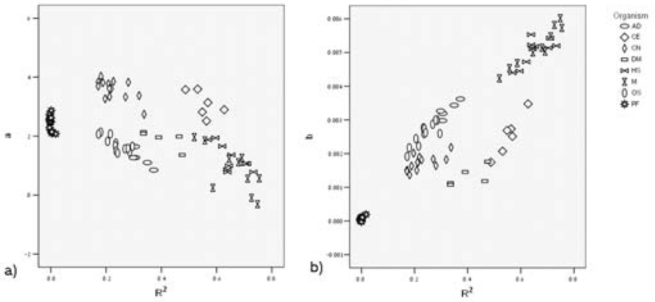

Fig. 6 Scatter-plots of parameters *a*, *b*, and R^2 of linear regression $N_{ex}=a+b\cdot L_{ex}$ for all the processed chromosomes. Left plot: *a* (y-axis) vs. R^2; right plot: *b* (y-axis) vs. R^2

B. Dendrogram of Chromosomes of All the Genomes

Our next goal is to visualize the chromosome classification using all of the parameters: n_{ex}, a_{ex}, *a*, *b*, and R^2 we have calculated (see Supplementary Material for the complete table of their values). We standardize each of the parameters to the interval [−1 ÷ +1], and then calculate the Euclidean distances in six-dimensional parameter space between all pairs of chromosomes *i* and *j* according to

$$d_{ij}^2 = \left(\mathrm{n'}_{\mathrm{i,ex}} - \mathrm{n'}_{\mathrm{j,ex}}\right)^2 + \left(\mathrm{l'}_{\mathrm{i,ex}} - \mathrm{n'}_{\mathrm{j,ex}}\right)^2 + \left(\mathrm{l'}_{\mathrm{i,ex}} - \mathrm{n'}_{\mathrm{j,ex}}\right)^2 + \left(\mathrm{a'}_{\mathrm{i}} - \mathrm{a'}_{\mathrm{j}}\right)^2 + \left(\mathrm{b'}_{\mathrm{i}} - \mathrm{b'}_{\mathrm{j}}\right)^2 + \left(\left(\mathrm{R}^2\right)_{\mathrm{i}} - \left(\mathrm{R}^2\right)_{\mathrm{j}}\right)^2,$$

where $n'_{i,ex}$, $l'_{i,ex}$, $a'_{i,ex}$, a'_{i}, b'_{i}, and $R^{2\prime}_i$ are the standardized parameters of the chromosome *i*. Having calculated the distance matrix d_{ij}, we used the *Neighbor Joining Algorithm* to obtain the dendrogram of our chromosomes. The chromosomes of one species were grouped together but separately from other species. There is only one exception: the chromosomes of the two mammal species *M. musculus* and *H. sapiens* form a single mixed branch (Fig. 7).

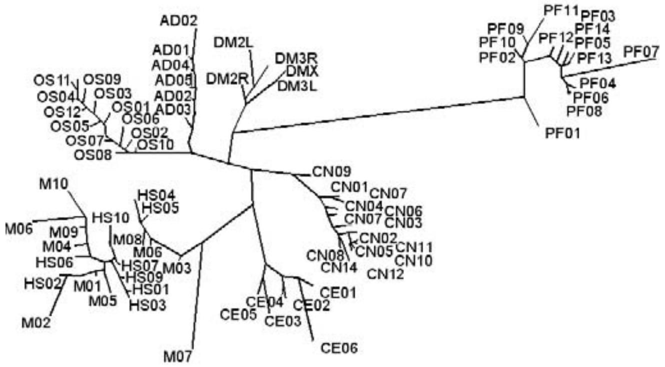

Fig. 7 Dendrogram of the 76 processed chromosomes of eight species based on weighted distances among parameters n_{ex}, l_{ex}, a_{ex}, *a*, *b*, and R^2 (*a, b,* and R^2 are parameters of the linear regression $N_{ex}=a+b\cdot L_{ex}$)

IV. Conclusion

Our results show both general and genome-specific features of the exon-intron organization of eukaryotic genes. The most general feature found in all genomes is the positive correlation between the number of introns in a gene and the corresponding protein's length (and equivalently, the net length of all the exons of the gene). In addition, in all the genomes we have studied, the average exon length in a gene decreases with the number of those exons. By while these laws of exon-intron statistics are quite general, the correlation parameters are genome-specific. For the first time, for our best knowledge, it was shown that they are specific to genomes rather than to individual chromosomes. Indeed, in the parameter space of average chromosome properties and linear regression parameters (between exon numbers and protein lengths), all chromosomes from the same genome form obvious clusters.

Clearly, the exon-intron structures of eukaryotic genes have many important parameters that we did not consider in this work; we have left them for the future research. The main goals of this article are to draw attention to the statistical properties of exon size distributions, and to visualize both the general laws of exon-intron organizations of genes and the genome-specific features.

Acknowledgment

A. K. thanks S. Hosid for the technical support.

References

[1] H. Naora and N.J. Deacon, “Relationship between the total size of exon and introns in the protein-coding genes of higher eukaryotes,” *Proc. Natl. Acad. Sci.. USA*, vol. 79: pp. 6196–6200, 1982.

[2] J.D. Hawkins, “A survey on intron and exon lengths,” *Nucleic Acids Res*., vol. 16: pp. 9893–9908, 1988.

[3] M. Deutsch and M. Long, “Intron-exon structures of eukaryotic model organisms,” *Nucleic Acids Res*., vol. 27: p. 3219–3228, 1999.

[4] E.V. Kriventseva and M.S. Gelfand, “Statistical analysis of the exon-intron structure of higher and lower eukaryote genes,” *J. Biomol. Struct. Dyn.*, vol. 17, pp. 281–288, 1999.

[5] A.A. Mironov and M.S. Gelfand, “Prediction and computer analysis of the exon-intron structure of human genes,” *Mol. Biol*., vol. 38, pp. 70–77, 2004.

[6] A.T. Ivashchenko and S.A. Atambayeva, “Variation in lengths of introns and exons in genes of the Arabidopsis thaliana nuclear genome,” *Russian Journal of Genetics*, vol. 40, pp. 1179–1181, 2004.

[7] S.W. Roy and D. Penny, “Intron length distributions and gene prediction,” *Nucleic Acids Res*., vol. 35, pp. 4737–4742, 2007.

[8] S.A. Atambayeva, V.A. Khailenko, and A.T. Ivashchenko, “Intron and exon length variation in arabidopsis, rice, nematode, and human,” *Mol. Biol*., vol. 42, pp. 312–320, 2008.

[9] A.T Ivashchenko,. V.A. Khailenko, and S.A. Atambayeva, “Variation of the lengths of exons and introns in Human Genome genes,” *Russian Journal of Genetics*, vol. 45, pp.16–22, 2009.

[10] F.S. Collins *et al*., “Finishing the euchromatic sequence of the human genome. International Human Genome Sequencing Consortium,” *Nature*, vol. 431, pp. 931–945, 2004.

[11] E.M. Schwarz *et al*., “WormBase: better software, richer content,” *Nucleic Acids Res*., vol. 34 (Database), pp. D475–D478, 2006.

[12] R.A. Drysdale and M.A. Crosby, “FlyBase: genes and gene models,” *Nucleic Acids Res*., vol. 33 (Database), pp. D390–D395, 2005.

[13] B.J. Haas *et al*., “Complete reannotation of the Arabidopsis genome: methods, tools, protocols and the final release,” *BMC Biol*., vol. 3, p. 7, 2005.

[14] J.M.J. Logsdon, A. Stoltzfus, and W.F. Doolittle, “Molecular evolution: recent cases of spliceosomal intron gain?” *Curr. Biol*., vol. 8: pp. R560–R563, 1998.

[15] J.M Archibald,. C.J. O'Kelly, and W.F. Doolittle, “The chaperonin genes of jakobid and jakobid-like flagellates: implications for eukaryotic evolution,” *Mol. Biol. Evol*., vol. 19, pp. 422–431, 2002.

[16] A.T Ivashchenko, M.I. Tauasarova, and S.A. Atambayeva, “Exon–Intron Structure of Genes in Complete Fungal Genomes,” *Mol. Biol*., vol. 43, pp. 24–31, 2009.

[17] B.J. Loftus *et al*., “The genome of the basidiomycetous yeast and human pathogen Cryptococcus neoformans,” *Science*, vol. 307, pp. 1321–1324, 2005.

[18] D. Martinez *et al*., “Genome sequence of the lignocellulose degrading fungus Phanerochaete chrysosporium strain RP78,” *Nat. Biotechnol*., vol. 22, pp. 695–700, 2004.

[19] S. Gudlaugsdottir, D.R. Boswell, G.R. Wood, and J. Ma, “Exon size distribution and the origin of introns,” *Genetica*, vol. 131, pp. 299–306, 2007.

[20] Y. Ryabov and M. Gribskov, “Spontaneous symmetry breaking in genome evolution,” *Nucleic Acids Res*., vol. 36, pp. 2756–2763, 2008.

[21] G. Cho and R.F. Doolittle, “Intron distribution in ancient paralogs supports random insertion and not random loss,” *J. Mol. Evol*., vol. 44, pp. 573–584, 1997.

[22] S.W. Roy, “The origin of recent introns: transposons?” *Genome Biol*., vol. 5, p. 251, 2004.

[23] W. Gilbert, “The exon theory of genes,” in *Symp. Quant. Biol*., Cold Spring Harbor, vol.52, 1987, pp.901–905.

[24] T. Cavalier-Smith, “Selfish DNA and the origin of introns,” *Nature*, vol. 315, pp. 283–284, 1985.

[25] J.M. Logsdon and J.D. Palmer, “Origin of introns – early or late?” *Nature,* vol. 369, pp. 526–528, 1994.

[26] M.K. Sakharkar, V.T. Chow, and P. Kangueane, “Distributions of exons and introns in the human genome,” *In Silico Biol*., vol. 4, pp. 387–393, 2004.

SUPPLEMENTARY MATERIAL
AVERAGE CHROMOSOME CHARACTERISTICS AND REGRESSION PARAMETERS OBTAINED FOR 76 PROCESSED CHROMOSOMES

Chromo-some	l_{ex}	n_{ex}	a_{ex}	$N_{ex}=a+b\cdot L_{ex}$		
				a	$b\cdot 10^{-3}$	R^2
PF01	1972	2.59	1138	2.54	.03	.000
PF02	2079	2.29	1384	2.11	.09	.008
PF03	2308	2.65	1581	2.85	.08	.006
PF04	2375	2.41	1583	2.48	.03	.001
PF05	2284	2.36	1596	2.46	-.05	.005
PF06	2419	2.57	1578	2.47	.04	.002
PF07	2772	2.30	1835	2.19	.04	.003
PF08	2379	2.66	1532	2.71	-.02	.000
PF09	2096	2.53	1343	2.57	-.02	.000
PF10	2085	2.21	1357	2.08	.06	.007
PF11	2149	2.23	1472	2.27	-.01	.000
PF12	2303	2.42	1529	2.06	.16	.022
PF13	2267	2.47	1526	2.60	-.06	.005
PF14	2316	2.29	1543	2.12	.07	.008
AD01	1277	5.42	411	1.08	3.40	.354
AD02	1185	5.07	407	.83	3.58	.377
AD03	1246	5.18	427	1.25	3.15	.311
AD04	1252	5.30	395	1.62	2.94	.310
AD05	1252	5.28	418	1.25	3.23	.303
OS01	1246	5.08	418	1.55	2.83	.273
OS02	1241	5.11	425	1.45	2.96	.290
OS03	1225	5.19	401	1.56	2.96	.284
OS04	1245	4.92	428	1.74	2.56	.240
OS05	1187	4.89	431	1.65	2.73	.238
OS06	1241	4.82	454	1.47	2.70	.241
OS07	1213	4.72	451	1.80	2.41	.210
OS08	1225	4.73	438	2.11	2.13	.185
OS09	1218	4.72	424	2.05	2.19	.220
OS10	1256	4.65	468	1.40	2.59	.246
OS11	1305	4.50	498	2.05	1.88	.177
OS12	1239	4.80	432	1.63	2.56	.300
CN01	1594	6.17	303	3.31	1.80	.275
CN02	1613	6.12	323	3.75	1.47	.213
CN03	1545	6.25	308	3.79	1.59	.193
CN04	1642	6.28	306	3.35	1.79	.324
CN05	1617	6.59	314	3.59	1.86	.223
CN06	1664	6.48	312	3.81	1.61	.285
CN07	1627	6.09	340	3.32	1.70	.217
CN08	1628	6.21	343	3.81	1.47	.176
CN09	1564	6.07	317	2.72	2.15	.342
CN10	1644	6.19	341	4.01	1.33	.186
CN11	1686	6.11	345	3.67	1.44	.176
CN12	1523	6.25	321	3.25	1.98	.201
CN13	1567	6.62	277	3.84	1.78	.231
CN14	1626	6.37	306	3.59	1.70	.216
DM2L	1597	3.79	522	1.97	1.15	.469
DM2R	1565	4.15	482	1.94	1.42	.395
DM3L	1582	3.83	535	2.12	1.09	.340
DM3R	1547	4.01	494	1.34	1.72	.480
DMX	1684	3.80	547	2.06	1.04	.340
CE01	1383	6.52	218	3.12	2.47	.573
CE02	1225	5.79	222	2.49	2.70	.568
CE03	1377	6.37	216	3.57	2.03	.536
CE04	1229	6.05	214	2.80	2.65	.553
CE05	1185	5.57	221	3.56	1.71	.492
CE06	1301	7.35	185	2.88	3.44	.632
M01	1557	9.01	283	1.24	4.99	.696
M02	1548	9.60	264	.55	5.70	.759
M03	1373	7.70	293	1.94	4.20	.524
M04	1407	8.20	284	.54	5.44	.717
M05	1572	9.12	284	1.09	5.10	.684
M06	1238	7.08	295	-.33	5.98	.753
M07	1386	6.67	395	.22	4.65	.591
M08	1439	8.31	291	1.83	4.50	.562
M09	1500	8.65	310	1.21	4.97	.650
M10	1466	8.36	307	-.11	5.78	.731
HS01	1504	8.72	279	1.05	5.10	.708
HS02	1611	9.53	245	1.05	5.39	.715
HS03	1627	9.78	269	.84	5.50	.643
HS04	1538	8.71	293	1.92	4.41	.600
HS05	1605	8.86	313	1.86	4.36	.573
HS06	1503	8.52	272	.76	5.16	.738
HS07	1468	8.52	280	1.64	4.69	.625
HS08	1453	8.35	280	.98	5.07	.649
HS09	1499	8.54	303	.76	5.19	.644
HS10	1507	9.09	255	1.34	5.14	.658

Preface to section 3.2

Kaplunovsky, A., Zabrodsky, D., Volkovich, Z., Ivashchenko, A.T., and Bolshoy, A. (2010).

Statistics of exon lengths in fungi.
The Open Bioinformatics Journal **4**: 31-40.

This manuscript deals with statistical properties of exon-intron organizations of genes in fungi. The exon-intron structures of fungi genes are quite different from each other, and the evolution of such structures raises many questions. We tried to address some of these questions with an accent on methods of revealing evolutionary factors based on the analysis of gene exon-intron structures using statistical analysis. Taking whole genomes of fungi, we went through all the protein-coding genes in each chromosome separately and calculated the portion of intron-containing genes and average values of the net length of all the exons in a gene, the number of the exons, and the average length of an exon. We found striking similarities between all of these average properties of chromosomes of the same species and significant differences between properties of the chromosomes belonging to species of different divisions (Phyla) of the kingdom of Fungi. Comparing those chromosomal and genomic averages, we have developed a technique of clustering based on characteristics of the exon-intron structure. This technique of clustering separates different fungi species, grouping them according to Fungi taxonomy. The main conclusion of this article is that the statistical properties of exon-intron organizations of genes are the genome-specific features preserved by evolutionary processes.

Open Access

Statistics of Exon Lengths in Fungi

Alexander Kaplunovsky[1], David Zabrodsky[2], Zeev Volkovich[2], Anatoliy Ivashchenko[3] and Alexander Bolshoy*,[1]

[1]*Department of Evolutionary and Environmental Biology and Genome Diversity Center at the Institute of Evolution, University of Haifa, Israel*

[2]*Department of Software Engineering, ORT Braude College, Karmiel, Israel*

[3]*Department of Biotechnology, Biochemistry, Plant Physiology at the Al-Farabi Kazakh National University, Kazakhstan*

Abstract: The exon-intron structures of fungi genes are quite different from each other, and the evolution of such structures raises many questions. We tried to address some of these questions with an accent on methods of revealing evolutionary factors based on the analysis of gene exon-intron structures using statistical analysis. Taking whole genomes of fungi, we went through all the protein-coding genes in each chromosome separately and calculated the portion of intron-containing genes and average values of the net length of all the exons in a gene, the number of the exons, and the average length of an exon. We found striking similarities between all of these average properties of chromosomes of the same species and significant differences between properties of the chromosomes belonging to species of different divisions (Phyla) of the kingdom of Fungi. Comparing those chromosomal and genomic averages, we have developed a technique of clustering based on characteristics of the exon-intron structure. This technique of clustering separates different fungi species, grouping them according to Fungi taxonomy. The main conclusion of this article is that the statistical properties of exon-intron organizations of genes are the genome-specific features preserved by evolutionary processes.

Keywords: Comparative genomics, Exon-intron structure, Eukaryotic clustering.

INTRODUCTION

The exon–intron structure is an important feature of a gene. The exon and intron lengths, as well as intron density, vary within a broad range [1-5]. In spite of a large amount of accumulated information on how the diversity of the exon–intron structure of genes is produced remains unclear and investigating the underlying factors will give further insight into the evolution of exon-intron structure.

A putative link between the biological role of introns and the distribution of exon sizes in protein-coding genes was established soon after intron discovery [6]. Since then, many studies – including statistical analysis – of the exon-intron structures of higher and lower eukaryote genes were performed [2, 5, 7-10]. The problem of exon and intron lengths' variability has a long history [10, 11], and it remains unsolved. We observed a huge variation of intron lengths, both between different organisms and between different genes of the same organism.

Likewise, we do not understand the evolutionary forces shaping species-specific chromosomal distributions of the intron densities (average numbers of introns per gene). At first, the intron density was thought to be related to organismal complexity. The initial studies supported this hypothesis: Homo sapiens have 8.1 introns per gene on average [12], Caenorhabditis elegans – 4.7 [13], Drosophila melanogaster – 3.4 [14], and Arabidopsis thaliana – 4.4 [15]; by contrast, unicellular species were found to have less introns per gene [16]. However, further studies found significantly high intron densities in many single-celled species [17, 18], and intron densities in basidiomycetes and zygomycete fungi are among the highest known among eukaryotes (4-6 per gene) [19, 20]. Diversity in intron densities among fungal genomes makes them extremely attractive for exploring questions of exon-intron structure evolution. Indeed, fungi display a wide diversity of gene structures, ranging from far less than one intron per gene for yeasts, to approximately 1–2 introns per gene on average for many recently sequenced ascomycetes (including the organisms in this study), to roughly seven introns per gene on average for some basidiomycetes (e.g., Cryptococcus).

Following the genome sequencing of several lower eukaryotes, it has become possible to examine exon–intron statistics with sufficiently large samples of genes. The lower eukaryotic genomes appeared to differ in many aspects, including the portion of intron-containing genes [19, 21]. Lower eukaryotes are of particular interest for studying the biological role of introns, since some of their genomes have only a few intron-containing genes, while the portion of such genes in other genomes is extremely high. The exon–intron structure of lower fungal genes has been examined in several works [1, 2, 8, 19, 21-26], but our current knowledge of the structure is still far away from being complete.

*Address correspondence to this author at the Department of Evolutionary and Environmental Biology and Genome Diversity Center at the Institute of Evolution, University of Haifa, Israel; Tel: +972-4-8240382; Fax: +972-4-8240382; E-mail: bolshoy@research.haifa.ac.il

In our previous paper [27] we have shown both general and genome-specific features of the exon-intron organization of eukaryotic genes of different kingdoms. We have shown that the most general feature found in all genomes is the positive correlation between the number of introns in a gene and the corresponding protein's length (equivalently, the net length of all the exons of the gene). In addition, in all the genomes we have studied, the average exon length negatively correlates with the average number of exons. Recently, analyses of patterns of exon-intron architecture variation brought Zhu and co-authors to the same conclusions [28]. One of their main conclusions was a decrease of average exon length as the total exon numbers in a gene increased. By while these laws of exon-intron statistics appeared to be quite general, nevertheless, many of the correlation parameters are genome-specific. In this study we continue the efforts of the previous one [27] to define genome-specific features of the exon-intron organization of fungal genomes.

There is mixture of different chromosomal characters of exon-intron organization. Among them we chose to limit ourselves to consideration of pure exonic properties and, additionally, proportions of intron-containing genes among all protein coding genes. In *A. fumigates*, for example, this proportion is ~80%. Does this mean that this property is consistent for every chromosome of *A. fumigates* and is the variation of this parameter negligible? For NC and GZ the values of this proportion are very close to 80% as well – does it mean that all other exonic properties should be similar as well? To answer this question we calculate and compare such exonic properties as exon densities, average exon lengths, and average net exon lengths. It was shown that in all genomes with a high proportion of intron-containing genes there is positive correlation between exon density and average protein length. As this was found for the genomes with a high proportion of intronless genes, the rule should be modified.

DATA AND METHODS

Fungi Species Data

Nucleotide sequences of 140 chromosomes of 15 fungi species presented in Table **1** have been obtained from GenBank ftp://ftp.ncbi.nih.gov/genomes/Fungi

A standard gene annotation looks like the following annotation of a randomly chosen gene NCU08052.1 of *Neurospora crassa*

gene	<25457..>26451.
mRNA	join(<25457..25690,25755..26055,26117..>26451),
CDS	join(25457..25690,25755..26055,26117..26451).

The annotation means the first exon of this gene starts somewhere upstream of the position 25457, and the last exon of the gene ends somewhere downstream of the position 26451. (In genomic annotations only, coding parts of exons are predicted sufficiently well, so everywhere in this study, when referring to "exons", we mean "coding parts of exons". In other words, only those introns within coding sequences and exons without UTR (untranslated regions) were used for analysis. The data related to coding parts of exons are taken from CDS (coding sequence) lines. For example, the CDS of NCU08052.1 consists of the three "exons" [25457:25690], [25755:26055], [26117:26451] with lengths of 234bp, 301bp, and 335bp. The length of the gene is larger than 995 bp, the number of exons is equal to 3, the net length of the exons (the protein size in bp) is equal to 870, and the average exon length is equal to 290.

Exon-Intron Structure Statistical Parameters

Each gene was assigned three values: the net length L_{ex} of all its exons, the number N_{ex} of those exons, and an average exon length A_{ex}: $A_{ex} = \frac{L_{ex}}{N_{ex}}$.

For each chromosome of each genome several absolute and averaged chromosomal characters were calculated. The proportion of intron-containing genes (p_c) is a relevant attribute; the average net length l_{ex} of all the exons in a gene per chromosome, the average number n_{ex} of the exons per gene per chromosome, and the average exon length a_{ex} are the characteristics of exons. a_{ex} is the mean of the A_{ex} values of individual genes per chromosome, $a_{ex} = \frac{1}{n}\sum_{1}^{n} A_{ex}$, where n denotes a number of genes in the chromosome here. Note that the a_{ex} is different from the average length $\bar{a}_{ex}$ of all the exons in the chromosome, regardless of which gene(s) they belong to. (The $\bar{a}_{ex}$, is calculated as a total length of all exons in a chromosome divided by a total number of all exons in a chromosome, see ref. [4]. The a_{ex} usually have significantly larger values than the $\bar{a}_{ex}$ because an average length of *i*-th exon exponentially decreases with an index *i*, see ref. [29].

We also calculated species-averaged exon parameters: N_g (total number of genes per genome), AN_{ex} (average number of exons in a gene per genome), AL_{ex} (average net length of all exons in a genome), AA_{ex} (average exon length in a gene per genome), $AN1_{ex}$ (average number of exons in a intron-containing gene per genome), $AL0_{ex}$ = average (over a genome) length of an intronless gene, $L1_{ex}$ (average net length of all exons in intron-containing genes), and P_g (proportion of intron-containing genes in genome in percents).

Distances Between Pairs of Fungal Chromosomes

One of our goals was to cluster the chromosomes using exon-intron structure parameters. We used distance-based methods of clustering; therefore, we had to define a method for a distance measuring. The distance between a pair of chromosomes was calculated as the distance between vectors constructed from several standardized parameters defined above. The complete vector x_r of chromosomal parameters related to chromosome r consists of (n_{ex}, l_{ex}, a_{ex}, p_c, $l0_{ex}$, $n1_{ex}$, $l1_{ex}$).

After having extracted parameters, our next task was to find an appropriate dissimilarity measure d such that $d(x_r, x_s)$ is small if and only if x_r and x_s are close. The simplest dissimilarity measure is the Euclidean distance:

$$d^2(x_r, x_s) = \sum_{k=1}^{K} (x_{r,k} - x_{s,k})^2$$

However the Euclidean distance is not suitable for further clustering, since it is isotropic, while the abovementioned exonic characters do not have similar behaviors. That is why

Table 1. List of Processed Species and their Chromosomes

N	Abbreviation	Name of the organism	Phylum / Class	Number of chromosomes
1	AF	*Aspergillus fumigatus*	Ascomycota Pezizomycotina	8
2	CG	*Candida glabrata CBS138*	Ascomycota Saccharomycotina	13
3	CN	*Cryptococcus neoformans*	Basidiomycota Agaricomycotina	14
4	DH	*Debaryomyces hansenii CBS767*	Ascomycota Saccharomycotina	7
5	EC	*Encephalitozoon cuniculi GB-M1*	Microsporidia Apansporoblastina	11
6	EG	*Eremothecium (Ashbya) gossypii*	Ascomycota Saccharomycotina	7
7	GZ	*Gibberella zeae*	Ascomycota Pezizomycotina	4
8	KL	*Kluyveromyces lactis*	Ascomycota Saccharomycotina	6
9	MG	*Magnaporthe grisea*	Ascomycota Pezizomycotina	7
10	NC	*Neurospora crassa*	Ascomycota Pezizomycotina	7
11	PS	*Pichia stipitis*	Ascomycota Saccharomycotina	8
12	SC	*Saccharomyces cerevisiae*	Ascomycota Saccharomycotina	16
13	SP	*Schizosaccharomyces pombe 972h*	Ascomycota Taphrinomycotina	3
14	UM	*Ustilago maydis*	Basidiomycota Ustilaginomycotina	23
15	YL	*Yarrowia lipolytica CLIB122*	Ascomycota Saccharomycotina	6
		Total		140

it was relevant to use a standardized Euclidean distance defined by:

$$d^2(x_r, x_s) = \sum_{k=1}^{K} \frac{(x_{r,k} - x_{s,k})^2}{\operatorname{var} x_k},$$

where var x_k is the empirical variance of x_k, i.e.,

$$\operatorname{var} x_k = \sum_{n=1}^{N} (x_{n,k} - m_k)^2; m_k = \frac{1}{N}\sum_{n=1}^{N} x_{n,k}$$

The reason for introducing this distance is of a statistical matter. The $x_{r,k}$ are considered as N realizations of a random variable x_r, such that the x_r are independent. Then var x_k is only the squared empirical standard deviation of x_k.

We also used a scaled Euclidean distance based on the scaling of all values $x_{r,k}$ of the objected parameter of x_r according to the given interval $[l_1, l_2]$:

$$x_{r,k} = l_1 + \frac{\left(x_{r,k} - x_{r,\min}\right)\left(l_2 - l_1\right)}{\left(x_{r,\max} - x_{r,\min}\right)},$$

Clustering of Fungal Chromosomes

Two methods of clustering were used: a well-known Neighbor Joining algorithm [30] and a Principal Directions Divisive Partitioning (PDDP) algorithm [31]. NJ constructs a tree that does not assume an evolutionary clock, so that it is, in effect, an unrooted tree. We used the program *Neighbor* of *Phylip* Package (the University of Washington) http://evolution.genetics.washington.edu/phylip/doc/neighbor.html, which is an implementation of NJ. Matrices of standardized and scaled distances between all pairs of 63 yeast chromosomes were exported to the program *Neighbor*. The output file was drawn by the program *TreeView* of Prof. Rod Page http://taxonomy.zoology.gla.ac.uk/rod/treeview.html.

The Principal Directions Divisive Partitioning (PDDP) algorithm, introduced by D. Boley [31], is a top-down hierarchical clustering method producing a binary tree in which each node is a data structure containing data items. Inherently, the algorithm has been designed to operate with a text mining task, based on the term–document matrix representation; although in reality this approach can be employed to different objects admitting similar matrix representation. Specifically, the algorithm manages instances given by an $n \times m$ matrix $M_m = [d_1, ..., d_m]$, whose columns and rows represent the "documents" and "terms", accordingly. In this study the "documents" are the fungal chromosomes, and the "terms" are the exon-intron statistical parameters described above.

At the start, all set M_m fits in the root of the tree. The algorithm continues by splitting all document vectors into two disjointed subsets resting upon principal data directions. Consecutively, both of the two partitions are recursively divided into two sub-partitions. As a result, a nested partitions' assembly is organized as a binary tree (the "PDDP tree") such that every partition is either a leaf node or is separated into two children in the PDDP tree.

Let us suppose, we have a partition represented by means $n \times p$ matrix M_p, $p \leq m$. The splitting of this partition is provided by the projection on the main leading eigenvector direction of the covariance matrix $C = \left(M_p - we^T\right)\left(M_p - we^T\right)^T$, where $e = \left(1,1,...,1\right)^T$ and w is the sample mean of the chromosomes $\left[d_1,...,d_p\right]$. In the simplest version of the algorithm used in this paper the chromosomes $\left[d_1,...,d_p\right]$ are put into the clusters exactly with respect to their projections sign. All the documents with non-positive projections form the left child and the remaining documents are fed into the right one. The cluster chosen for splitting in the PDDP process is the one having the largest variance calculated as the square Frobenius norm.

$$\left\|M_p - we^T\right\|_F^2 = \sum_{i,j}\left(M_p - we^T\right)_{i,j}^2$$

Note that this criterion usually leads to clusters with more or less similar sizes.

Analyses of the Structural-Functional Organization of the System

One-way ANOVA statistical method was used to test for differences in the exon-intron structure between several groups of fungi species. We also used Factor analysis (FA) as an integral statistical method, giving the opportunity to define and to evaluate the structural-functional organization of the system. We chose Principal components analysis (PCA) as one of the techniques of FA. The method produces a set of eigenvectors calculated from the matrix of correlations between parameters where each of them represents a causal connection of elements. It is important to note that by using the technique of PCA, all factors become orthogonal and are caused by different properties of the system.

RESULTS AND DISCUSSION

All of the abovementioned chromosomal characteristics (n_{ex}, l_{ex}, a_{ex}, p_c, $l0_{ex}$, $n1_{ex}$, $l1_{ex}$) were calculated for all 140 chromosomes. The intragenomic variation was found to be pretty small everywhere, exactly as it was expected. As an illustration, the values of these characteristics for a randomly selected organism, *A. fumigatus*, are given in Supplementary (Table **S1**).

Every column in Table **S1** contains of practically indistinguishable parameters. For example, there is the same proportion of intron-containing genes in all eight chromosomes of *A. fumigatus* P_c = 78.5±0.5%.

Table **S2** (Supplementary) shows that the sets L_{ex} and N_{ex} do not demonstrate significant differences among various chromosomes of *A. fumigatus*. We can see that F-statistics comparing variances between and within groups of chromosomes is not significant; therefore, all chromosomes have only indistinguishable distributions of L_{ex} and N_{ex}.

Analogical results were obtained for the chromosomal parameters of all other organisms as well. For all chromosomal characters of all genomes the differences between two chromosomes of an identical genome appeared not to be statistically significant. Would the differences between two chromosomes of two different species depend on the evolutionary distance between these two organisms? Would it be possible to identify an organism by a combination of chromosomal characters? As it appeared (Figs. **1-2**) a pair of characters does not provide full partition of all species.

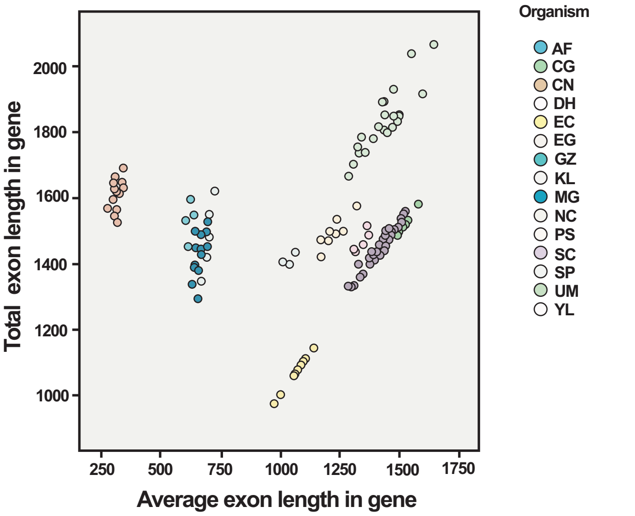

Fig. (1). Scatter-plot of the average exon length per gene a_{ex} (*x*-axis) vs. the total exon length l_{ex} (*y*-axis) for all 140 processed chromosomes of 15 fungi species.

Species-Averaged Statistical Parameters

In Table **2**, in addition to parameters averaged over all genes, there are data related to intron-containing ($L1_{ex}$) and intronless genes ($AL0_{ex}$) separately. For the set of intronless genes, the parameters AL_{ex} and AA_{ex} are identical and equal to an average gene length $AL0_{ex}$. In the section Methods there are descriptions and formulas for calculations of these parameters. Some putative empirical rules may be observed in Table **2**. For example, regarding average gene lengths of intron-containing and intronless genes, it seems that if there is only a small amount of intron-containing genes in a genome, these genes are shorter in average than other intronless genes of the same genome. This property is especially strongly expressed in EC, CG, and KL, and also exists for EG, DH, SP, and UM. Another observation may be done regarding a lack of correlation between amounts of genes in a genome and other genomic statistical parameters.

Chromosome-Averaged Statistical Parameters

Let us consider the average parameters l_{ex}, n_{ex} and a_{ex}. Scatter-plot of a_{ex} *vs.* l_{ex} is shown in Fig. (**1**). Every organism in the plot is presented by a specific combination of a color and the filling in of a circle. As we mentioned above, Fig. (**1**) shows that the averages l_{ex} and a_{ex} turned out to be pretty similar for different chromosomes of the same species but rather distant for different species. Moreover, five separate groups of points may be observed in Fig. (**1**). The two parameters l_{ex} and a_{ex} cluster separately all 14 chromosomes of *C. neoformans* (CN) in one group, 8 chromosomes of *E. cuniculi* (EC) in another group, and all 23 chromosomes of *U. maydis* (UM) in the third group. All other points are distributed between two additional groups.

Analyzing the contents of the groups presented in Fig. (**1**), one can suppose that the partitions follow fungal taxonomy. Fig. (**2b**) is obtained from Fig. (**1**) by coloring all

Table 2. Exon Parameters by Species

Organism	N_g	AN_{ex}	AL_{ex}	AA_{ex}	$AN1_{ex}$	$L1_{ex}$	P_g	$AL0_{ex}$	$AL0_{ex}$ / AL_{ex}	$AL0_{ex}$ / $L1_{ex}$
AF	9002	2.935	1476	671	3.462	1522	78.58	1304	0.883	0.856
CG	5174	1.016	1513	1507	2.024	671	1.59	1527	1.009	2.275
CN	6318	6.262	1608	317	6.428	1624	96.95	1112	0.691	0.684
DH	6231	1.057	1387	1357	2.057	1092	5.38	1403	1.011	1.284
EC	1995	1.008	1079	1078	2.071	435	0.50	1084	1.005	2.492
EG	2952	1.049	1460	1441	2.032	882	4.38	1501	1.028	1.165
GZ	6745	3.238	1520	624	3.682	1564	83.44	1299	0.854	0.830
KL	5257	1.024	1422	1413	2.016	733	2.40	1439	1.012	1.963
MG	9675	2.875	1411	852	3.490	1485	75.31	1185	0.839	0.798
NC	6343	2.699	1459	690	3.123	1481	80.01	1370	0.939	0.925
PS	5299	1.417	1493	1220	2.566	1746	25.86	1402	0.939	0.803
SC	5859	1.055	1489	1450	2.029	1466	5.31	1491	1.001	1.017
SP	4990	1.952	1417	1042	3.089	1310	45.56	1507	1.063	1.150
UM	5539	1.751	1831	1443	2.979	1642	37.93	1947	1.063	1.186
YL	6425	1.160	1460	1339	2.135	1646	14.10	1430	0.979	0.869
Total	87804	2.229	1484	1023	3.774	1526	44.31	1450	0.977	0.950

points in six colors related to six fungi classes (see Table **1**): *Ascomycota Pezizomycotina*, *Ascomycota Saccharomycotina*, *Ascomycota Taphrinomycotina*, *Basidiomycota Agaricomycotina*, *Basidiomycota Ustilaginomycotina*, and *Microsporidia Apansporoblastina*.

Fig. (**2a**) presents a scatter plot of the a_{ex} *vs.* n_{ex}, and clearly shows four separate groups of chromosomes: CN chromosomes belonging to *Basidiomycota Agaricomycotina* form the most left group, *Ascomycota Pezizomycotina* chromosomes make the second left group, three chromosomes of *S. pombe* (*Ascomycota Taphrinomycotina*) are located together but separately from other points on the plot, and the points belonging to other fungi classes (*Basidiomycota Ustilaginomycotina*, *Microsporidia Apansporoblastina*, and *Ascomycota Saccharomycotina*) appear more or less together. The CN chromosomes have the greatest exon density (n_{ex}) and the shortest exons (l_{ex}) among all the fungi chromosomes

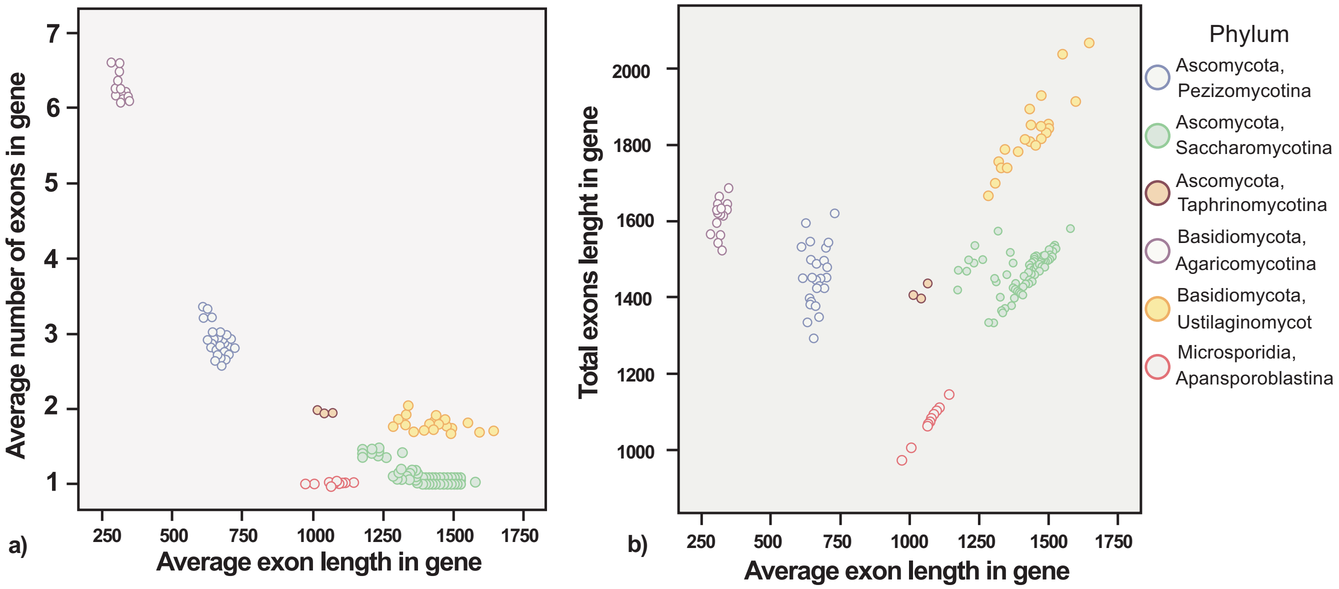

Fig. (2). Scatter-plot for all 140 processed chromosomes of six fungi phyla of the average exon length per gene a_{ex} (x-axis). **a**) vs. the average number of exons per gene n_{ex} (y-axis); **b**) vs. the average net exon length per gene l_{ex} (y-axis).

we have studied. Scatter-plots of a_{ex} vs. n_{ex} (Fig. **2a**) and a_{ex} vs. l_{ex} (Fig. **2b**) show that already three parameters a_{ex}, n_{ex} and l_{ex} are sufficient for successful classification of 140 chromosomes to six fungal classes.

At this point, we use factor analysis of the system of 140 chromosomes that led us to the synthesis of the following successive logical structure:

1. Dividing the system into sets of "elementary" components – all of the abovementioned chromosomal characteristics (n_{ex}, l_{ex}, a_{ex}, p_c, $l0_{ex}$, $n1_{ex}$, $l1_{ex}$)
2. Analysis of the relationships of these components in species
3. Revealing system-forming relations
4. Description of the structure of the system (model) and its properties

As we can see from Table **3**, four main components are responsible for the whole system organization, and two of them can describe 93.9% of the whole variability of the system.

Table 3. Total Variance Explained

Component	% of Variance	Cumulative %
1	73.841	73.841
2	20.042	93.884
3	3.887	97.771
4	2.229	100.000

The detailed Table **S3** placed in Supplementary data, shows relationships of these principal components in species as a component structure of 140 chromosomes on the basis of their exon-intron structure. Results of Table **S3** (Supplementary) are shown also in Figs. (**3**, **4**). We can see that the first component strongly divides all species into yeasts (*Saccharomycotina*) vs. *Pezizomycotina* and *Taphrinomycotyna*, and the second component demonstrates the difference between *Microsporidia* and *Basidiomycota*. Unfortunately, we can also see that the chromosomes of the species of the phylum Basidiomycota are split by the first component between two groups: they appear in the first group together with *Agaricomycotina* (CN) and in the second group together with *Ustilaginomycotina* (UM).

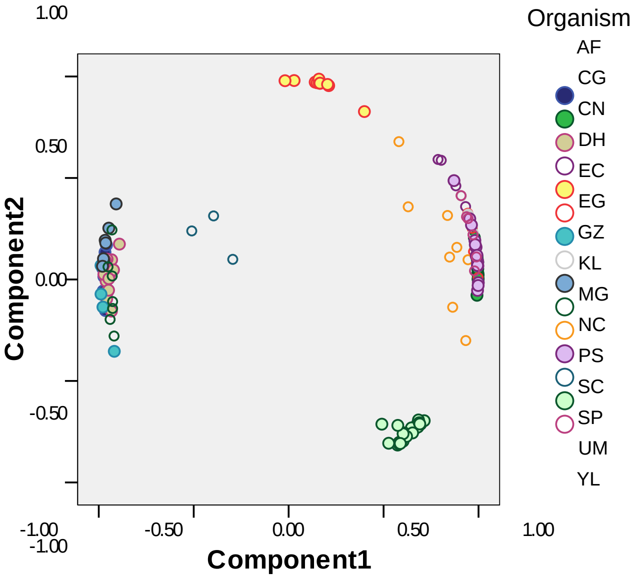

Fig. (3). Factor analysis of 140 processed chromosomes of 15 fungi species by seven parameters (n_{ex}, l_{ex}, a_{ex}, p_c, $l0_{ex}$, $n1_{ex}$, $l1_{ex}$) colored in reference to species.

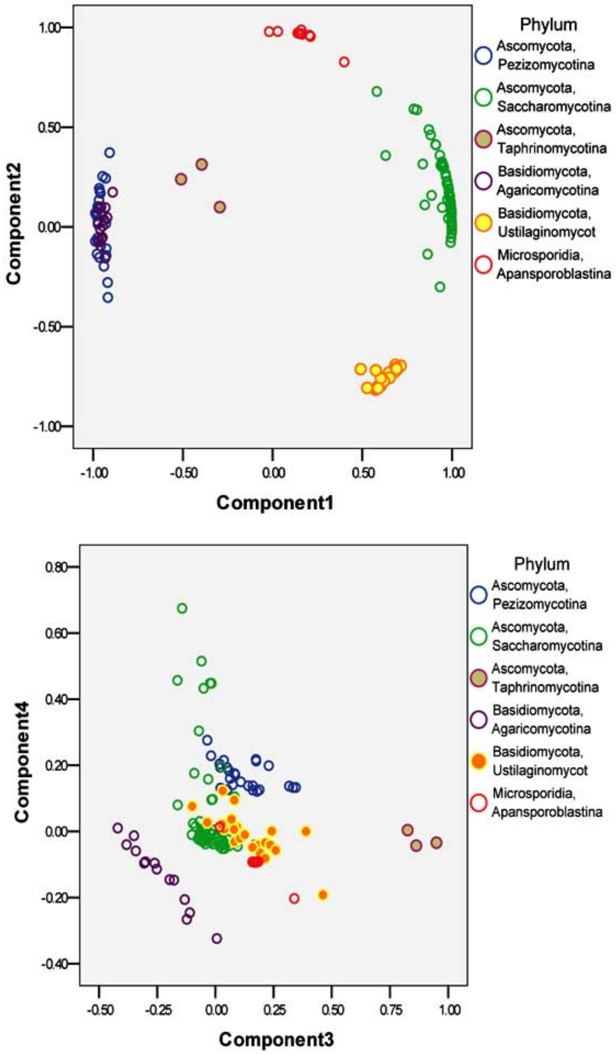

Fig. (4). Factor analysis of 140 processed chromosomes of 15 fungi species by seven parameters (n_{ex}, l_{ex}, a_{ex}, p_c, $l0_{ex}$, $n1_{ex}$, $l1_{ex}$) colored in reference to phylum.

The PDDP method based on scaled distance measures produced a tree presented in Fig. (**5**). There are 5 terminal nodes at the tree: 05, 07, 08, 09, and 10. Some species may be characterized by a homogeneous distribution of the chromosomes: all chromosomes of CN are in cluster 05, SP chromosomes are in 09, all 8 chromosomes of AF are in cluster 10, and so on. However, there are species with "non-uniform" distribution: for example, the third chromosome of GZ is located in cluster 10 while the other 3 chromosomes are in cluster 05.

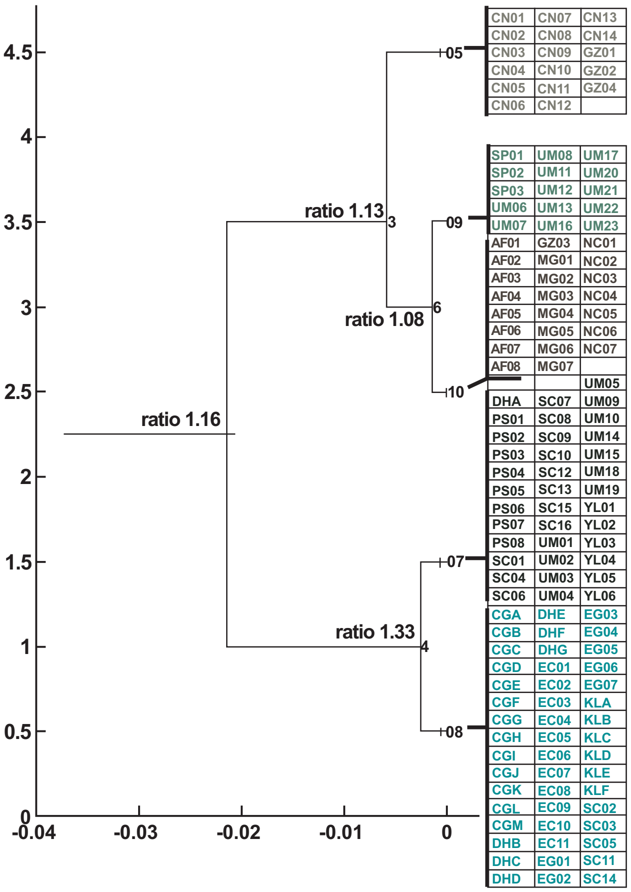

Fig. (5). A dendrogram of clusters obtained by the PDDP method based on scaled distances among the vectors (n_{ex}, l_{ex}, a_{ex}, p_c, $l0_{ex}$, $n1_{ex}$, $l1_{ex}$) presenting the chromosomes.

The standardized distances (Fig. **6**) led to better results. There are more terminal nodes, and the clusters corresponding to the leaves of the tree are more homogeneous than in the previous dendrogram (Fig. **5**); nevertheless, the third chromosome of GZ is located differently from other chromosomes of the same organism similarly to the previous tree. Moreover, the first chromosome of YL and the chromosome 07 of EG appear separately in "wrong" clusters.

Clustering results presented in Figs. (**5**, **6**) appear to be sufficiently similar. It may be considered as evidence of the consistency of recovered cluster structures.

Table **4** presents the measure of strength of association between two final partitions. The Cramer's contingency coefficient built on a contingency table is 0.865. Therefore, it can be concluded that there is a strong association among the partitions.

Clustering results presented in Figs. (**5**, **6**) are based on different distance measures: scaled distances and standardized distances. The denominators are different for these measures. One of the discussed problems is the choice of the denominator for the distance parameters. What is the proper scaling parameter needed to make the data dimensionless? Because *a priori* we do not know contribution of which parameters will take the highest effect, we can only try different kinds of multiplying factors and compare results of classification. As we have found, the normalization to unite standard deviation gave us the best result but, of course, it is not the only way to dimensionless data representation.

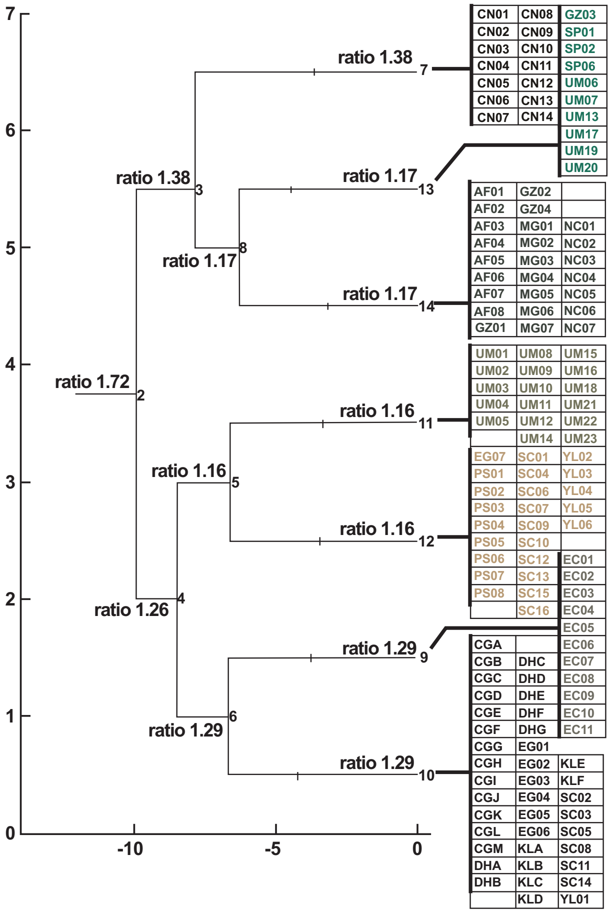

Fig. (6). Dendrogram of clusters obtained by the PDDP method based on standardized distances among the vectors (n_{ex}, l_{ex}, a_{ex}, p_c, $l0_{ex}$, $n1_{ex}$, $l1_{ex}$) presenting the chromosomes.

We know that in bioinformatics there are many other methods in use. For example, in the very popular correspondence analysis (positive) data are normalized to unite mean. For bistochastization or binormalization more sophisticated methods were used, see, for example, a highly cited paper [32], a review [33] or a very seminal mathematical paper [34]. The reason for all alternative approaches to data normalization is, usually, very simple. Any normalization, either to unit variance of variables or to unit interval or to any other factor may cause many mistakes and may give enormously high weight to unimportant features, and only the final result may judge whether our choice was justified. As we mentioned above, clustering results are sufficiently similar, which may be considered as justification of our choices.

Dendrogram of Yeast Chromosomes

All applied clustering techniques based on distances among vectors (n_{ex}, l_{ex}, a_{ex}, p_c, $l0_{ex}$, $n1_{ex}$, $l1_{ex}$) sometimes did not succeed in distinguishing between chromosomes of different species, especially between yeasts. Therefore, we decided to use linear regression between the net length of exons of a gene l_{ex} and the number of exons n_{ex} in genes on all processed chromosomes. Now, the vectors presenting the chromosomes additionally to averaged chromosomal parameters n_{ex}, l_{ex}, and a_{ex} contained correlation coefficients *an, al,* and *nl*, linear regression parameters *a, b,* and a parameter of explained variation R^2 of a regression $n_{ex}=a+b \cdot l_{ex}$, as in our previous paper [27]. We applied the program *Neighbor* using scaled distances. The dendrogram presented in Fig. (**7**) was drawn by the program *TreeView*. There are two main features of the dendrogram: a) practically all chromosomes of the same yeast species are distributed compactly along the tree, and 2) the chromosomes belonging to the same species form a separate cluster.

CONCLUSIONS

We applied statistical analysis of the exon-intron structure in order to reveal general and genome-specific features of fungi genes. Taking the complete genomes of fungi, we went through all of the protein-coding genes in each chromosome separately and calculated the portion of intron-containing genes and average values of the net length of all the exons in a gene, the number of the exons, and the average length of an exon. The purpose of this research has been to determine the most appropriate approach to classify fungal chromosomes, according to these simple exon-intron statistics. We tested a few clustering techniques measuring distances among the chromosomes in different ways.

Firstly, we found that intragenomic variation is substantially smaller than intergenomic variance everywhere. In other words, we found that the laws of exon-intron statistics are specific to genomes rather than to individual chromosomes.

Secondly, we commented on the consistent similarity of the partitions, which resulted from rather different clustering methods. Clustering results obtained with scaled and normalized Euclidean distances appear to be sufficiently similar. The Principal Components (PC) clustering, the Principal Directions Divisive Partitioning (PDDP) method, and the Neighbor joining (NJ) algorithm produced very similar clustering results.

Thirdly, we propose techniques of clustering that are able to distinguish between chromosomes of different species with satisfactory success. The addition of regression parame-

Table 4. Contingency

Cluster index in Fig. (5)		**05**	**09**	**10**	**07**	**08**
Cluster index in Fig. (6)	**# of items**	**17**	**15**	**23**	**37**	**48**
07	14	**14**	0	0	0	0
13	10	0	**8**	1	1	0
14	25	3	0	**22**	0	0
11	17	0	7	0	**10**	0
12	24	0	0	0	**23**	1
09	11	0	0	0	0	**11**
10	39	0	0	0	3	**36**

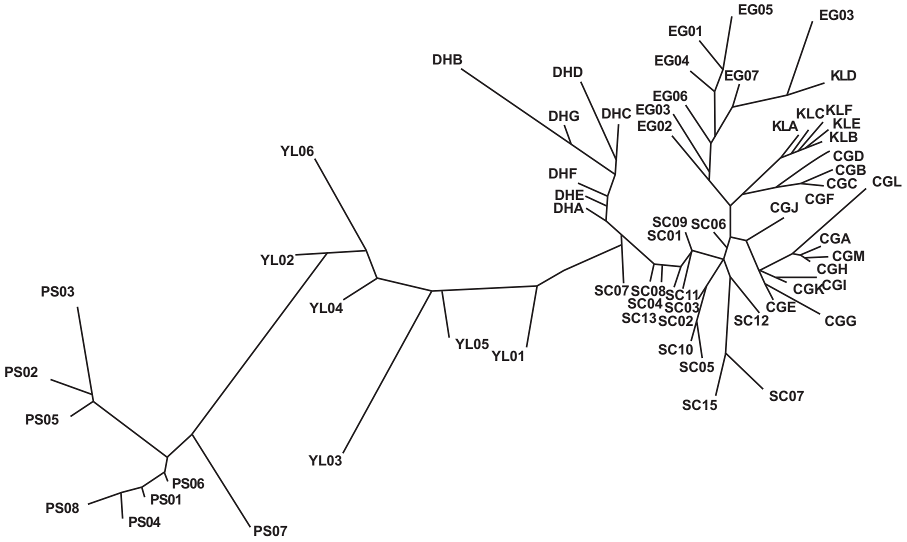

Fig. (7). Dendrogram of the 63-processed chromosomes of seven yeast species based on scaled distances among parameters n_{ex}, l_{ex}, a_{ex}, *an, al, nl, a, b,* and R^2 (*an, al,* and *nl* are correlation coefficients; *a, b,* and R^2 are parameters of the linear regression $n_{ex}=a+b \cdot l_{ex}$) obtained by NJ clustering technique.

ters to averaged chromosomal parameters n_{ex}, l_{ex}, and a_{ex} improved the resolution of clustering. We added to parameters n_{ex}, l_{ex}, and a_{ex} parameters of linear regression $n_{ex}=a+b\cdot l_{ex}$ and got a phylogenetic tree of the yeasts.

Clearly, the exon-intron structures of eukaryotic genes have many important parameters that we did not consider in this work; we intend to pursue these in future research. In particular, the ratio between the exon and intron lengths appears to be an important feature of a gene. In some genomes the intron length is comparable with the exon length: in unicellular eukaryotes [1, 2], plants [2, 35], and particular animals [2-4]. In general, introns are longer than exons in mammalian genes [11].

ACKNOWLEDGEMENTS

We thank Prof. Daniel Boley for his kind assistance in using his Retrieve Experimental Software for Principal Direction Divisive Partitioning method. We thank Ms. Robin Permut for the editing.

ABBREVIATIONS

N_{ex} = number of exons in a gene

L_{ex} = net length of all exons in a gene

A_{ex} = average exon length in a gene

n_{ex} = average (over a chromosome) number of exons in a gene

l_{ex} = average (over a chromosome) net length of all exons in a gene

a_{ex} = average (over a chromosome) of the average exon length in a gene

p_c = is a proportion of intron-containing genes in a chromosome

$l0_{ex}$ = average (over a chromosome) length of an intronless gene

$a0_{ex}$ = $l0_{ex}$

$l1_{ex}$ = average (over a chromosome) net length of all exons in an intron-containing gene

$a1_{ex}$ = average (over a chromosome) of the average exon length of an intron-containing gene

$n1_{ex}$ = average number of exons in a intron-containing gene per chromosome

N_g = total number of genes per genome

AN_{ex} = average (over a genome) number of exons in a gene

AL_{ex} = average (over a genome) net length of all exons in a gene

AA_{ex} = average (over a genome) of the average exon length in a gene

P_g = is a proportion of intron-containing genes in a genome

$AL0_{ex}$ = average (over a genome) length of an intronless gene

$AA0_{ex}$ = $AL0_{ex}$

$AL1_{ex}$ = average (over a genome) net length of all exons in an intron-containing gene

$AA1_{ex}$ = average (over a genome) of the average exon length of an intron-containing gene

$AN1_{ex}$ = average number of exons in an intron-containing gene per genome

SUPPLEMENTARY MATERIAL

Supplementary material is available on the publishers Web site along with the published article.

REFERENCES

[1] D. M. Kupfer, S. D. Drabenstot, K. L. Buchanan, H. Lai, H. Zhu, D. W. Dyer, B. A. Roe, and J. W. Murphy, "Introns and splicing elements of five diverse fungi," *Eukaryot. Cell,* vol. 3, pp. 1088-1100, 2004.

[2] M. Deutsch, and M. Long, "Intron-exon structures of eukaryotic model organisms," *Nucleic Acids Res.*, vol. 27, pp. 3219-3228., 1999.

[3] J. F. Wendel, R. C. Cronn, I. Alvarez, B. Liu, R. L. Small, and D. S. Senchina, "Intron size and genome size in plants," *Mol. Biol. Evol.*, vol. 19, pp. 2346-2352, 2002.

[4] M. K. Sakharkar, V. T. Chow, and P. Kangueane, "Distributions of exons and introns in the human genome," *In Silico Biol.*, vol. 4, pp. 387-393, 2004.

[5] S. W. Roy and D. Penny, "Intron length distributions and gene prediction," *Nucleic Acids Res.*, vol. 35, pp. 4737-4742, 2007.

[6] H. Naora and N. J. Deacon, "Relationship between the total size of exon and introns in the protein-coding genes of higher eukaryotes," *Proc. Natl. Acad. Sci. USA.* , vol. 79, pp. 6196-6200, 1982.

[7] J. D. Hawkins, "A survey on intron and exon lengths," *Nucleic Acids Res.*, vol. 16, pp. 9893-9908, 1988.

[8] E. V. Kriventseva and M. S. Gelfand, "Statistical analysis of the exon-intron structure of higher and lower eukaryote genes," *J. Biomol. Struct. Dyn.*, vol. 17, pp. 281-288, 1999.

[9] A. T. Ivashchenko and S. A. Atambayeva, "Variation in lengths of introns and exons in genes of the Arabidopsis thaliana nuclear genome," *Russ. J. Genet.*, vol. 40, pp. 1179-1181, 2004.

[10] S. A. Atambayeva, V. A. Khailenko, and A. T. Ivashchenko, "Intron and exon length variation in arabidopsis, rice, nematode, and human," *Mol. Biol.*, vol. 42, pp. 312-320, 2008.

[11] A. T. Ivashchenko, V. A. Khailenko, and S. A. Atambayeva, "Variation of the lengths of exons and introns in Human Genome genes," *Russ. J. Genet.*, vol. 45, pp. 16-22, 2009.

[12] F. S. Collins, E. S. Lander, J. Rogers, and R. H. Waterston, "International human genome sequencing consortium: finishing the euchromatic sequence of the human genome," *Nature,* vol. 431, pp. 931-945, 2004.

[13] E. M. Schwarz, I. Antoshechkin, C. Bastiani, T. Bieri, D. Blasiar, P. Canaran, J. Chan, N. Chen, W. J. Chen, P. Davis, T. J. Fiedler, L. Girard, T. W. Harris, E. E. Kenny, R. Kishore, D. Lawson, R. Lee, H-M, Muller, C. Nakamura, P. Ozersky, A. Petcherski, A. Rogers, W. Spooner, M. A. Tuli, K. Van Auken, D. Wang, R. Durbin, J. Spieth, L. D. Stein, and P. W. Sternberg, "WormBase: better software, richer content," *Nucleic Acids Res.*, vol. (34 Database), pp. D475-478, 2006.

[14] R. A. Drysdale, M. A. Crosby, and C. FlyBase, "FlyBase: genes and gene models," *Nucleic Acids Res.*, vol. 33, pp. D390-D395, 2005.

[15] B. J. Haas, J. R. Wortman, C. M. Ronning, L. I. Hannick, R. K. J. Smith, R. Maiti, A. P. Chan, C. Yu, M. Farzad, D. Wu, O. White, and C.D. Town, "Complete reannotation of the Arabidopsis genome: methods, tools, protocols and the final release," *BMC Biol.*, vol. 3, p. 7, 2005.

[16] J. M. J. Logsdon, A. Stoltzfus, and W. F. Doolittle, "Molecular evolution: recent cases of spliceosomal intron gain?," *Curr. Biol.*, vol. 8, pp. R560-R563, 1998.

[17] J. M. Archibald, C. J. O'Kelly, and W. F. Doolittle, "The chaperonin genes of jakobid and jakobid-like flagellates: implications for eukaryotic evolution," *Mol. Biol. Evol.*, vol. 19, pp. 422-431, 2002.

[18] A. T. Ivashchenko, M. I. Tauasarova, and S. A. Atambayeva, "Exon–intron structure of genes in complete fungal genomes," *Mol. Biol.,* vol. 43, pp. 24-31, 2009.

[19] B. J. Loftus, E. Fung, P. Roncaglia, D. Rowley, P. Amedeo, D. Bruno, J. Vamathevan, M. Miranda, I. J. Anderson, J. A. Fraser, J. E. Allen, I. E. Bosdet, M. R. Brent, R. Chiu, T. L. Doering, M. J. Donlin, C. A. D'Souza, D. S. Fox, V. Grinberg, J. Fu, M. Fukushima, B. J. Haas, J. C. Huang, G. Janbon, S. J. Jones, H. L. Koo, M.I. Krzywinski, J. K. Kwon-Chung, K. B. Lengeler, R. Maiti, M. A. Marra, R. E. Marra, C. A. Mathewson, T. G. Mitchell, M. Pertea, F. R. Riggs, S. L. Salzberg, J. E. Schein, A. Shvartsbeyn, H. Shin, M. Shumway, C. A. Specht, B. B. Suh, A. Tenney, T. R. Utterback, B. L. Wickes, J. R. Wortman, N. H. Wye, J. W. Kronstad, J. K. Lodge, J. Heitman, R. W. Davis, C. M. Fraser, and R. W. Hyman, "The genome of the basidiomycetous yeast and human pathogen *Cryptococcus neoformans*," *Science*, vol. 307, pp. 1321-1324, 2005.

[20] D. Martinez, R. M. Berka, B. Henrissat, M. Saloheimo, M. Arvas, S. E. Baker, J. Chapman, O. Chertkov, P. M. Coutinho, D. Cullen, E. G. Danchin, I. V. Grigoriev, P. Harris, M. Jackson, C. P. Kubicek, C. S. Han, I. Ho, L. F. Larrondo, A. L. de Leon, J. K. Magnuson, S. Merino, M. Misra, B. Nelson, N. Putnam, B. Robbertse, A. A. Salamov, M. Schmoll, A. Terry, N. Thayer, A. Westerholm-Parvinen, C. L. Schoch, J. Yao, R. Barabote, M. A. Nelson, C. Detter, D. Bruce, C. R. Kuske, G. Xie, P. Richardson, D. S. Rokhsar, S. M. Lucas, E. M. Rubin, N. Dunn-Coleman, M. Ward, and T. S. Brettin, "Genome sequencing and analysis of the biomass-degrading fungus Trichoderma reesei (syn. Hypocrea jecorina)," *Nat. Biotechnol.,* vol. 26, pp. 553-560, 2008.

[21] M. D. Katinka, S. Duprat, E. Cornillot, G. Méténier, F. Thomarat, G. Prensier, V. Barbe, E. Peyretaillade, P. Brottier, P. Wincker, F. Delbac, H. El Alaoui, P. Peyret, W. Saurin, M. Gouy, J. Weissenbach, and C. P. Vivarès, "Genome sequence and gene compaction of the eukaryote parasite *Encephalitozoon cuniculi*," *Nature,* vol. 414, pp. 450-453, 2001.

[22] M. Spingola, L. Grate, and D. A. Haussler, "Genomewide bioinformatic and molecular analysis of intron in S. cerevisiae," *RNA*, vol. 5, pp. 221-234, 1999.

[23] T. J. Sharpton, D. E. Neafsey, J. E. Galagan, and J. W. Taylor, "Mechanisms of intron gain and loss in Cryptococcus," *Genome Biol.,* vol. 9, pp. R24.1-R24.10, 2008.

[24] C. Nielsen, B. Friedman, B. Birren, C. Burge, and J. E. Galagan, "Patterns of intron gain and loss in fungi," *PLoS Biol.,* vol. 2, p. e422, 2004.

[25] J. E. Stajich and F. S. Dietrich, "Evidence of mRNA mediated intron loss in the human-pathogenic fungus Cryptococcus neoformans," *Eukaryot. Cell*, vol. 5, pp. 789-793, 2006.

[26] J. E. Stajich, F. S. Dietrich, and S. W. Roy, "Comparative genomic analysis of fungal genomes reveals intron-rich ancestors," *Genome Biol.,* vol. 8, p. R223, 2007.

[27] A. Kaplunovsky, V. A. Khailenko, A. Bolshoy, S. A. Atambayeva, and A. T. Ivashchenko, "Statistics of exon lengths in animals, Plants, fungi, and protists," *Int. J. Biol. Life Sci.*, vol. 1, pp. 139-144, 2009.

[28] L. C. Zhu, Y. Zhang, W. Zhang, S. H. Yang, J. Q. Chen, and D. C. Tian, "Patterns of exon-intron architecture variation of genes in eukaryotic genomes," *BMC Genomics,* vol. 10, p. 12, 2009.

[29] S. Gudlaugsdottir, D. R. Boswell, G. R. Wood, and J. Ma, "Exon size distribution and the origin of introns," *Genetica,* vol. 131, pp. 299-306, 2007.

[30] N. Saitou, and M. Nei, "The neighbor-joining method: a new method for reconstructing phylogenetic trees," *Mol. Biol. Evol.*, vol. 4, pp. 406-425, 1987.

[31] D. Boley, "Principal directions divisive partitioning," *Data Min. Knowl. Disc.,* vol. 2, pp. 325-344, 1988.

[32] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander, "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring," *Science,* vol. 286, pp. 531-537, 1999.

[33] S. C. Madeira and A. L. Oliveira, "Biclustering Algorithms for Biological Data Analysis: A Survey," *IEEE Transactions on Computational Biology and Bioinformatics,* vol. 1, pp. 24-45, 2004.

[34] J. Franklin and J. Lorenz, "On the scaling of multidimensional matrices," *Linear Algebra and its Application,* vol. 114/115, pp. 717-735, 1989.

[35] X. Y. Ren, O. Vorst, M. W. E. J. Fiers, W. J. Stiekema, and J. P. Nap, "In plants, highly expressed genes are the least compact," *Trends Genet.,* vol. 22, pp. 528-532, 2006.

Received: June 01, 2010 Revised: October 12, 2010 Accepted: October 13, 2010

SUPPLEMENTARY TABLES

Table S1. Exonic Chromosomal Parameters of *Aspergillus fumigatus*

	n_{ex}	l_{ex}	a_{ex}	$n1_{ex}$	$l1_{ex}$	p_c	$l0_{ex}$	$l0_{ex}$ / l_{ex}	$l0_{ex}$ / $l1_{ex}$
AF1	2.990±0.09	1488±55	664±32	3.491±0.10	1520±63	79.87	1360±101	0.913	0.895
AF2	2.935±0.09	1529±60	698±39	3.450±0.09	1562±66	78.99	1405±131	0.918	0.899
AF3	2.894±0.10	1448±56	675±34	3.476±0.11	1507±67	76.48	1254±96	0.866	0.832
AF4	2.953±0.11	1452±59	647±33	3.462±0.11	1504±70	79.29	1254±99	0.864	0.864
AF5	2.938±0.11	1494±68	691±50	3.481±0.12	1526±68	78.09	1381±189	0.924	0.904
AF6	3.007±0.10	1498±63	646±35	3.452±0.11	1562±71	81.81	1212±126	0.809	0.776
AF7	2.784±0.14	1426±75	669±45	3.388±0.15	1507±92	74.72	1187±107	0.832	0.788
AF8	2.839±0.15	1378±80	656±51	3.424±0.17	1416±93	75.86	1257±150	0.912	0.888
AF	2.935±0.04	1476±23	671±14	3.462±0.04	1522±25	78.58	1304±47	0.883	0.856

Table S2. Results of ANOVA Test to Parameters of Chromosomes of *Aspergillus fumigatus*

Number_of_exons	**Sum of Squares**	**df**	**Mean Square**	**F**	**Sig**
Between Groups	33.524	7	4.789	1.374	0.211
Within Groups	33516.100	9615	3.486		
Total	33549.624	9622			
Total_exon_length	**Sum of Squares**	**df**	**Mean Square**	**F**	**Sig**
Between Groups	15022658.093	7	2146094.013	1.730	0.097
Within Groups	11925704534.704	9615	1240322.885		
Total	11940727192.798	9622			

Table S3. Component Matrix. Extraction Method: Principal Component Analysis with 4 extracted components.

	Component			
	1	**2**	**3**	**4**
AF01	-.977	-.057	.161	.125
AF02	-.965	-.153	.160	.138
AF03	-.978	.076	.082	.175
AF04	-.980	.056	.085	.172
AF05	-.973	-.077	.179	.121
AF06	-.973	.014	-.019	.229
AF07	-.967	.134	.036	.215
AF08	-.959	.169	.189	.126
CGA	.978	.206	-.009	-.038
CGB	.997	.031	-.070	-.024
CGC	.995	.066	-.075	-.022
CGD	.995	.064	-.072	-.024
CGE	.998	.022	-.055	-.026
CGF	.994	.097	-.044	-.029
CGG	.996	.084	-.016	-.031
CGH	.996	.052	-.073	-.021
CGI	.992	.113	-.057	-.030
CGJ	.998	-.001	-.056	-.018

(Table S3). Contd…..

	Component			
	1	2	3	4
CGK	.992	.108	-.049	-.030
CGL	.992	-.078	-.101	-.010
CGM	.998	.004	-.061	-.022
CN01	-.932	.097	-.350	-.013
CN02	-.967	-.067	-.132	-.206
CN03	-.956	.100	-.261	-.096
CN04	-.937	.033	-.342	-.059
CN05	-.954	-.062	-.123	-.266
CN06	-.922	.047	-.382	-.040
CN07	-.961	-.009	-.252	-.114
CN08	-.958	-.096	-.110	-.246
CN09	-.973	.025	-.178	-.147
CN10	-.947	.004	-.306	-.096
CN11	-.948	-.053	-.301	-.092
CN12	-.965	.092	-.197	-.146
CN13	-.891	.174	-.420	.010
CN14	-.933	-.156	.006	-.324
DHA	.952	.305	.030	-.006
DHC	.973	.224	-.043	-.038
DHD	.804	.587	.085	-.035
DHE	.932	.359	.040	-.036
DHF	.950	.308	.050	-.023
DHG	.881	.461	.095	-.045
DNB	.787	.591	.169	-.040
EC01	.029	.980	.177	-.092
EC02	.211	.955	.185	-.091
EC03	.140	.972	.165	-.093
EC04	.150	.969	.174	-.093
EC05	.161	.969	.162	-.093
EC06	.160	.987	.018	.013
EC07	.165	.966	.179	-.093
EC08	-.019	.979	.182	-.091
EC09	.165	.967	.171	-.094
EC10	.399	.827	.339	-.203
EC11	.205	.961	.159	-.092
EG01	.967	.252	.029	-.038
EG02	.997	.071	-.017	-.024
EG03	.987	.149	.043	-.040
EG04	.944	.326	.034	-.035
EG05	.974	.137	-.162	.080
EG06	.992	.124	.011	-.021
EG07	.998	.009	-.054	-.012
GZ01	-.978	-.136	.074	.141
GZ02	-.988	.069	.051	.126
GZ03	-.918	-.354	.170	-.044
GZ04	-.988	-.072	.022	.131
KLA	.945	.322	.019	-.046
KLB	.949	.311	.020	-.046
KLC	.973	.229	-.010	-.038

(Table S3). Contd…..

	Component			
	1	**2**	**3**	**4**
KLD	.910	.411	.029	-.052
KLE	.945	.319	.039	-.052
KLF	.988	.147	-.048	-.029
MG01	-.974	.092	.018	.204
MG02	-.966	.193	.063	.160
MG03	-.975	.102	.021	.194
MG04	-.963	.180	.050	.193
MG05	-.908	.372	.059	.183
MG06	-.980	.065	.109	.150
MG07	-.947	.253	.141	.138
NC01	-.927	-.143	.317	.137
NC02	-.930	.244	.175	.213
NC03	-.927	-.109	.333	.133
NC04	-.940	-.197	.174	.218
NC05	-.929	.017	.345	.133
NC06	-.919	-.279	-.035	.276
NC07	-.951	.062	.229	.199
PS01	.886	.158	-.051	.433
PS02	.630	.358	-.143	.675
PS03	.837	.315	-.017	.448
PS04	.848	.110	-.060	.515
PS05	.581	.679	-.023	.448
PS06	.945	.097	-.071	.304
PS07	.933	-.301	.027	.195
PS08	.864	-.137	-.163	.457
SC01	.995	.074	-.064	-.007
SC02	.996	.087	.010	-.006
SC03	.871	.487	.057	-.044
SC04	.995	.082	-.052	.008
SC05	.954	.299	.028	-.019
SC06	.981	.192	-.012	.005
SC07	.988	.116	-.096	.024
SC08	.963	.268	.026	.003
SC09	.997	.067	-.042	.008
SC10	.995	-.054	-.086	.015
SC11	.999	-.016	-.027	-.020
SC12	.998	-.030	-.059	.006
SC13	.995	.069	-.068	.030
SC14	.987	.159	-.017	-.009
SC15	.983	.171	-.069	-.005
SC16	.992	.119	-.035	.011
SP01	-.294	.099	.950	-.034
SP02	-.510	.239	.826	.004
SP03	-.395	.313	.863	-.043
UM01	.695	-.713	.092	.014
UM02	.678	-.711	.181	-.042
UM03	.663	-.738	.081	.095
UM04	.681	-.727	.082	-.031
UM05	.647	-.731	.214	-.034

(Table S3). Contd…..

	Component			
	1	**2**	**3**	**4**
UM06	.603	-.797	.038	.007
UM07	.574	-.816	.074	.020
UM08	.591	-.799	.106	-.021
UM09	.686	-.699	.191	-.065
UM10	.714	-.696	.080	.006
UM11	.609	-.756	.236	-.041
UM12	.582	-.810	.069	.037
UM13	.582	-.803	.031	.123
UM14	.653	-.756	-.035	.027
UM15	.684	-.692	.214	-.081
UM16	.621	-.773	.126	-.010
UM17	.527	-.807	.259	-.057
UM18	.688	-.706	.159	-.048
UM19	.691	-.712	-.100	.076
UM20	.603	-.760	.242	.001
UM21	.575	-.719	.389	.000
UM22	.588	-.808	.027	.021
UM23	.491	-.713	.462	-.192
YL01	.908	.413	.059	.026
YL02	.979	.042	-.091	.176
YL03	.937	.309	-.031	.158
YL04	.989	.111	-.017	.096
YL05	.971	.220	-.012	.097
YL06	.945	.300	.082	.105

Preface to section 3.3

Kaplunovsky, A., Ivashchenko, A.T., and Bolshoy, A. (2011).

Statistical analysis of exon lengths in various eukaryotes.

Open Access Bioinformatics **3:** 1-15.

The goal of this research is to determine the most appropriate approach to classify eukaryotic chromosomes, according to simple exon-intron statistics. The exon-intron structures of eukaryotes genes are quite different from each other, and the evolution of such structures raises many problematical questions. As a preliminary attempt to address some of these questions we performed statistical analysis of gene exon-intron structures. Taking whole genomes of eukaryotes, we went through all the protein-coding genes in each chromosome separately and calculated the portion of intron-containing genes and average values of the net length of all the exons in a gene, the number of the exons, and the average length of an exon. Comparing those chromosomal and genomic averages, we have developed a technique of clustering based on characteristics of the exon-intron structure. This technique of clustering separates different species, grouping them according to eukaryotes taxonomy. Our conclusion is that the best approach is based on distances among four principal components obtained by Factor analysis and following by application of such clustering algorithms as Neighbor Joining, *k*-means and Partitioning Around Medoids.

Open Access Bioinformatics

Dovepress

open access to scientific and medical research

Open Access Full Text Article

ORIGINAL RESEARCH

Statistical analysis of exon lengths in various eukaryotes

Alexander Kaplunovsky[1]
Anatoliy Ivashchenko[2]
Alexander Bolshoy[1]

[1]Department of Evolutionary and Environmental Biology, Genome Diversity Center, Institute of Evolution, University of Haifa, Israel; [2]Department of Biotechnology, Biochemistry, Plant Physiology, Al-Farabi Kazakh National University, Kazakhstan

Purpose: The principal goals of this research were to investigate correlations between certain properties of exons in a gene (ie, between exon density and the corresponding protein length) and to compare genomic trees obtained with different approaches of clustering based on exonic parameters. The aim was a better understanding of exon–intron structures and their origin and development. The exon–intron structures of eukaryote genes are quite different from each other, and the evolution of such structures raises many problematic questions. As a preliminary attempt to address some of these questions, we performed a statistical analysis of gene exon–intron structures.

Methods: Taking whole genomes of eukaryotes, we went through all the protein-coding genes in each chromosome separately and calculated the portion of intron-containing genes and average values of the net length of all the exons in a gene, the number of the exons, and the average length of an exon. Comparing those chromosomal and genomic averages, we developed a technique of clustering based on characteristics of the exon–intron structure. This technique of clustering separates different species, grouping them according to eukaryote taxonomy.

Conclusion: Our conclusion is that the best approach is based on distances among four principal components obtained by factor analysis and followed by application of clustering algorithms, such as neighbor-joining, *k*-means, and partitioning around medoids.

Keywords: comparative genomics, exon–intron structure, eukaryotic clustering, principal component analysis

Introduction

It is no secret that people are fond of classifying things. Genomics is no exception. A lot of methods exist in comparative genomics that can be used for the purpose of genome classification.[1] The objective of cluster analysis is to divide objects into clusters in a way that similarity among the items belonging to the same group is higher than similarity among items belonging to distinct groups. In this study, we intend to show that genome clustering based on exon–intron structural characteristics is essentially accurate and reliable and expands on the results of previous studies.[2,3] This kind of clustering neither supports a widely accepted taxonomy nor argues against it. Uncovering and further analyzing the exon–intron structural properties that unify or distinguish genomes in the clustering procedure improve our understanding of the nature and evolutionary history of splicing.

One of the greatest enigmas of eukaryotic genome evolution is the widespread existence of introns. Introns have been detected in the genes of viruses,

Correspondence: Alexander Bolshoy
Department of Evolutionary
and Environmental Biology, University
of Haifa, Haifa 39105, Israel
Tel +972 4824 0382
Fax +972 4824 0382
Email bolshoy@research.haifa.ac.il

submit your manuscript | www.dovepress.com
Dovepress
DOI: 10.2147/OAB.S14448

chloroplasts, and mitochondria of both lower and higher eukaryotes. This study focuses on the most important type of introns, ie, the spliceosomal introns of nuclear-encoded protein genes. Here we survey some of the properties of the exon–intron structure of these genes in almost all completely sequenced eukaryotic genomes. Net and averaged exonic lengths are among the attributes considered in this study.

The exon and intron lengths vary across a broad range.[4–8] Statistical analyses of exon and intron lengths have been performed several times on different sets of eukaryotes.[2,3,5,8–15]

Previously, we have shown some genome-specific features of the exon–intron organization of eukaryotic genes using a limited set of genomes from different kingdoms.[2] We have shown that the most general feature found in all genomes is a positive correlation between the number of introns in a gene and the corresponding protein length (ie, the net length of all the exons of the gene). In addition, we have shown that the average exon length correlates negatively with the average number of exons. Recently, analyses of patterns of exon–intron architecture variation brought Zhu et al to the same conclusion.[16] One of their main observations was a decrease in average exon length as the total exon numbers in a gene increased. Although the laws of exon–intron statistics appeared to be quite general, many of the correlation parameters were genome-specific.

Intron density, which is the average number of introns per gene, is an evolutionary riddle. At first, it was thought that one could simply predict intron density from organism complexity. Initial studies supported this hypothesis, ie, *Homo sapiens* has 8.1 introns per gene on average,[17] *Caenorhabditis elegans* has 4.7,[18] *Drosophila melanogaster* has 3.4,[19] and *Arabidopsis thaliana* has 4.4.[20] In contrast, unicellular species were found to have fewer introns per gene.[21] However, further studies found significantly higher intron densities in many unicellular species,[15,22] and intron densities in Basidiomycetes and Zygomycete fungi appeared to be among the highest known for eukaryotes (4–6 per gene).[23,24] Diversity in intron densities among fungal genomes makes them extremely attractive for exploring possible answers to questions concerning exon–intron structure evolution. Indeed, fungi display a wide diversity of gene structures, ranging from less than one intron per gene for yeasts to approximately 1–2 introns per gene, on average, for many recently sequenced lower fungi (including the organisms in this study) and to roughly 5.5 introns per gene on average for some Basidiomycetes (eg, *Cryptococcus*).

Following the genome sequencing of several lower eukaryotes, it has become possible to examine exon–intron statistics with sufficiently large samples of genes. The purpose of our recent publication[3] was to determine the most appropriate approach to classify fungal chromosomes according to simple exon–intron statistics. We tested a few clustering techniques measuring distances among the chromosomes in different ways. As a result of our analysis, we commented on the consistent similarity of the partitions, resulting from different clustering methods. Clustering results[3] obtained with scaled and normalized Euclidean distances appeared to be sufficiently similar. The principal components-based clustering method, the principal directions divisive partitioning method, and the neighbor-joining algorithm produced very similar clustering results. Therefore, we propose techniques of clustering that are able to distinguish between chromosomes of different species with satisfactory results. The addition of regression parameters to averaged chromosomal parameters improves the resolution of clustering.

There is a mixture of different chromosomal characteristics in exon–intron organization. In this study, similar to our previous publications, we considered only pure exonic properties and, additionally, proportions of intron-containing genes among all protein-coding genes. We calculated and compared exonic properties, including exon densities, average exon lengths, and average net exon lengths. In this study we investigated the correlation between the number of exons in a gene (exon density) and the corresponding protein length; compared intragenomic variation with intergenomic variance of exon densities, average exon lengths, and average net exon length; compared genomic trees obtained using different approaches of clustering based on exonic parameters; and paved a road for further evolutionary in silico research of exon–intron structure and its origins and development.

Methods

Data set

The nucleotide sequences of 322 chromosomes of 32 species presented in Table 1 were obtained from the database of the Eukaryotic Genome Sequencing Projects (http://www.ncbi.nlm.nih.gov/genomes/leuks.cgi). Gene annotations were used to calculate genic statistical properties. A standard gene annotation looks like the following annotation of a randomly chosen gene, NCU08052.1 of *Neurospora crassa*:

Gene <25457..>26451.
mRNA join (<25457..25690,25755..26055,26117..
>26451).

Coding sequence join (25457..25690,25755..26055, 26117..26451).

The annotation means the first exon of this gene starts somewhere upstream of position 25457, and the last exon of the gene ends somewhere downstream of position 26451. For the purposes of this study, the term "exons" refers to "coding parts of exons". In other words, only those introns within coding sequences and exons without untranslated regions were used for analysis. The data related to coding parts of exons are taken from coding sequence lines. For example, the coding sequence of NCU08052.1 consists of three "exons" (25457:25690, 25755:26055, and 26117:26451) with lengths of 234 bp, 301 bp, and 335 bp, respectively. The length of the gene is greater than 995 bp, the number of exons is equal to 3, the net length of the exons (the protein size in bp) is equal to 870, and the average exon length is equal to 290.

Exon–intron structure and statistical parameters

Each gene was assigned three gene-related exonic values, ie, the net length, $L_{ex,}$ of all its exons, the number, N_{ex}, of those exons, and an average exon length, A_{ex}:

$$A_{ex} = \frac{L_{ex}}{N_{ex}}$$

For each chromosome of each genome, several absolute and averaged chromosomal characters were calculated. In addition to the three averaged characteristics of exons, the average net length, l_{ex}, of all the exons in a gene per chromosome, the average number, n_{ex}, of the exons in a gene per chromosome, the average exon length, a_{ex}, per chromosome, and the proportion of intron-containing genes, p_c, as a relevant attribute were calculated. It should be

Table 1 List of processed species and their chromosomes

Kingdom/ supergroup	Phylum	Class	Organism	Abbreviation	Chromosomes (n)
Animalia	Arthropoda	Insecta	*Drosophila melanogaster*	DM	6
	Chordata	Mammalia	*Canis familiaris*	CF	19
			Homo sapiens	HS	10
			Mus musculus	MM	10
	Nemata	Caenorhabditis	*Caenorhabditis elegans*	CE	6
Fungi	Ascomycota	Ascomycetes	*Neurospora crassa*	NC	7
		Eurotiomycetes	*Aspergillus fumigatus*	AF	8
		Saccharomycotina	*Candida glabrata*	CG	13
			Debaryomyces hansenii	DH	7
			Eremothecium gossypii	EG	7
			Kluyveromyces lactis	KL	6
			Pichia stipitis	PS	8
			Saccharomyces cerevisiaei	SC	16
			Yarrowia lipolytica	YL	6
		Sordariomycetes	*Gibberella zeae*	GZ	4
			Magnaporthe grisea	MG	7
		Taphrinomycotina	*Schizosaccharomyces pombe*	SP	3
	Basidiomycota	Agaricomycotina	*Cryptococcus neoformans*	CN	14
		Ustilaginomycotina	*Ustilago maydis*	UM	23
	Microsporidia	Apansporoblastina	*Encephalitozoon cuniculi*	EC	11
Plantae	Magnoliophyta	Liliopsida	*Oryza sativa*	OS	12
		Magnoliopsida	*Arabidopsis thaliana*	AD	5
Plantae/Viridiplantae	Chlorophyta	Prasinophyceae	*Micromonas* sp. RCC299	MS	17
			Ostreococcus_lucimarinus	OL	21
Protista/ Chromalveolata	Ciliophora	Ciliatea	*Paramecium tetraurelia*	PT	1
	Apicomplexa	Aconoidasida	*Plasmodium falciparum*	PF	14
			Plasmodium knowlesi	PK	14
			Theileria annulata	TA	3
Protista/Chromista	Cryptophyta	Cryptophyceae	*Guillardia theta*	GT	3
			Hemiselmis anderenii	HA	3
Protista/Protozoa	Euglenozoa	Kinetoplastea	*Leishmania braziliensis*	LB	35
Protista/Rhizaria	Cercozoa	Chlorarachniophycea	*Bigelowiella natans*	BN	3
Total					**322**

mentioned that a_{ex} is the mean of the A_{ex} values of individual genes per chromosome:

$$a_{ex} = \frac{1}{n}\sum_{1}^{n} A_{ex}$$

where n denotes a number of genes in the chromosome here. The measure a_{ex} defined in this is different from the average length, $\bar{a}_{ex}$, of all the exons in the chromosome, regardless of which gene(s) they belong to. The $\bar{a}_{ex}$ is calculated as the total length of all exons in a chromosome divided by the total number of all exons in a chromosome.[7] The a_{ex} usually have significantly larger values than the $\bar{a}_{ex}$ because an average length of i-th exon exponentially decreases with an index, i.[25]

We also calculated species-averaged exon parameters, ie, N_g (total number of genes per genome), AN_{ex} (average number of exons in a gene per genome), AL_{ex} (average net length of all exons in a gene per genome), AA_{ex} (average exon length in a gene per genome), $AN1_{ex}$ (average number of exons in an intron-containing gene per genome), $AL0_{ex}$ (average length of an intronless gene per genome), $AL1_{ex}$ (average net length of all exons in an intron-containing gene per genome), and P_g (proportion of intron-containing genes in a genome in percent).

Distances between pairs of genomes

One of our goals was to cluster genomes using exon–intron structure parameters. We used distance-based methods of clustering, so had to define a method for distance measurement. The distance between a pair of genomes was calculated as the distance between vectors constructed from several standardized parameters defined above. The vector $\overline{x}_r$ of genomic parameters related to genome r consists of (AN_{ex}, AL_{ex}, AA_{ex}, $AN1_{ex}$, $AL1_{ex}$, $AL0_{ex}$), and is equal to

$$\overline{x}_r = \left\{ \frac{j_{ex,r} - \mu_j}{\sigma_j} \right\},\ j \in \left\{ AN_{ex}, AL_{ex}, AA_{ex}, AN1_{ex}, AL1_{ex}, AL0_{ex} \right\},$$

where μ_j is the mean value of a genomic parameter j and σ_j is its standard deviation.

Having extracted these parameters, our next task was to find an appropriate dissimilarity measure, d, such that $d(x_r, x_s)$ is small if x_r and x_s are close. The simplest dissimilarity measure is a normalized (standardized) Euclidean distance:

$$d(x_r, x_s) = \sqrt{\sum_{k=1}^{K} \left(\overline{x}_{r,k} - \overline{x}_{s,k} \right)^2}$$

Clustering of genomes

A few popular algorithms were used to cluster all 32 genomes. First of all, the well known neighbor-joining algorithm[26] was used. Using neighbor-joining, a tree that does not assume an evolutionary clock was constructed, and therefore, in effect, an unrooted tree results. We used the Neighbor of Phylip program package from the University of Washington (http://evolution.genetics.washington.edu/phylip/doc/neighbor.html), which is an implementation of neighbor-joining. Matrices of standardized distances between all pairs of chromosomes were exported to the Neighbor program. The output file was drawn by the TreeView program of Professor Rod Page (http://taxonomy.zoology.gla.ac.uk/rod/treeview.html). We also used well known k-medoid clustering by partitioning around medoids and k-means algorithms. The k-medoids and k-means algorithms are described elsewhere.[1]

Analyses of structural–functional organization of the system

One-way analysis of variance (ANOVA) was used to test for differences in exon–intron structure between several groups of species. We also used factor analysis as an integral statistical method, affording an opportunity to define and evaluate the structural–functional organization of the system. We chose principal components analysis as one of the techniques for factor analysis. This method produces a set of eigenvectors calculated from the matrix of correlations between parameters where each set represents a causal connection of elements. It is important to note that by using the technique of principal components analysis, all factors become orthogonal and are caused by different properties of the system.

Results

The aforementioned chromosomal characteristics (n_{ex}, l_{ex}, a_{ex}, p_c, $l0_{ex}$, $n1_{ex}$, $l1_{ex}$) were calculated for all 322 chromosomes. As an illustration, the values of these characteristics for a randomly selected unicellular organism, *Plasmodium knowlesi*, are given in supplementary Table S1. Every column in Table S1 contains indistinguishable parameters. The intragenomic variation was found to be rather small for other unicellular organisms as well, as shown with fungi.[3]

The results of the one-way ANOVA test for differences in the first three parameters, n_{ex}, l_{ex}, and a_{ex}, of chromosomes of all genomes are presented in Table 2. In general, we found intragenomic variation in l_{ex} and a_{ex} to be quite small for almost all unicellular organisms, and this was significant in n_{ex}, l_{ex}, and a_{ex} for Plantae and Animalia (especially for

Table 2 Results of one-way ANOVA test for differences in parameters between chromosomes for several species

Organism	Kingdom/supergroup	n_{ex}	l_{ex}	a_{ex}
AD	Plantae	0.006**	0.000***	0.007**
AF	Fungi	0.211	0.097	0.431
BN	Protista/Rhizaria	0.591	0.790	0.193
CE	Animalia	0.000***	0.000***	0.000***
CF	Animalia	0.000***	0.000***	0.000***
CG	Fungi	–	0.979	0.976
CN	Fungi	0.591	0.764	0.077
DH	Fungi	–	0.190	0.058
DM	Animalia	0.000***	0.000***	0.000***
EC	Fungi	–	0.203	0.226
EG	Fungi	–	0.377	0.423
GT	Protista/Chromista	–	0.128	0.112
GZ	Fungi	0.000***	0.040**	0.000***
HA	Protista/Chromista	–	0.599	0.599
HS	Animalia	0.002**	0.123	0.000***
KL	Fungi	–	0.427	0.389
LB	Protista/Protozoa	–	0.003**	0.002**
MG	Fungi	0.045**	0.014*	0.565
MM	Animalia	0.000***	0.000***	0.000***
MS	Plantae/Viridiplantae	0.000***	0.342	0.002**
NC	Fungi	0.037*	0.009**	0.947
OL	Plantae/Viridiplantae	0.000***	0.305	0.863
OS	Plantae	0.000***	0.075	0.000***
PF	Protista/Chromalveolata	0.083	0.168	0.053
PK	Protista/Chromalveolata	0.471	0.770	0.548
PS	Fungi	0.253	0.392	0.203
PT	Protista/Chromalveolata	–	–	–
SC	Fungi	0.692	0.993	0.985
SP	Fungi	0.651	0.570	0.321
TA	Protista/Chromalveolata	0.004**	0.771	0.680
UM	Fungi	0.309	0.539	0.366
YL	Fungi	–	0.319	0.523

Notes: *significance $0.01 < P < 0.05$; **significance $0.001 < P < 0.01$; ***significance $P < 0.001$; –parameters that did not pass the Levene test of homogeneity.

D. melanogaster chromosomes, with an outstanding and short chromosome 4). Table 2 shows that the sets a_{ex}, l_{ex}, and n_{ex} in various chromosomes demonstrate significant differences. We can see that F-statistics comparing variances between and within groups of chromosomes are significant. The ANOVA method was used only for parameters that passed the Levene test of homogeneity. As can be seen, most species with a low percentage of intron-containing genes in chromosome p_c did not pass this test for n_{ex}.

Problems investigated in this study included correlations between different species-averaged parameters of exon–intron structure, clustering chromosomes of a few organisms belonging to the same kingdom (Protista, Plantae, and Animalia) by combinations of chromosome-averaged exonic characteristics, and clustering of all 32 organisms by combinations of species-averaged characteristics of exons.

Correlations among species-averaged statistical parameters

In Table 3, in addition to parameters averaged over all genes, there are data related to a set of "intron-containing" genes ($AL1_{ex}$) and to a set of "intronless" genes ($AL0_{ex}$). In the Methods section, there are descriptions and formulae for calculations of these parameters. Some putative empiric rules may be deduced from Table 3. For example, regarding average protein lengths of intron-containing and intronless genes (net length of all exons), it seems that if there is only a small amount of intron-containing genes in a genome, such proteins are shorter on average than other proteins coded by intronless genes of the same genome. This property is especially strongly expressed for some species of fungi (*Encephalitozoon cuniculi*, *Candida glabrata*, and *Kluyveromyces lactis* and also exists for *Eremothecium gossypii*, *Debaryomyces hansenii*, *Schizosaccharomyces pombe*, and *Ustilago maydis*), and for three Protista species (*Leishmania braziliensis*, *Hemiselmis anderenii*, and *Guillardia theta*). Figure 1 shows a scatter-plot of P_g versus a fraction of $AL0_{ex}/AL1_{ex}$ and is obtained from Table 3. *H. anderenii* does not appear in Figure 1 because it has no intron-containing genes. There are three main groups of points in the plot, ie, a group of genomes with a low concentration of intron-containing genes ($P_g < 10\%$), a group of genomes with a high concentration of intron-containing genes ($P_g > 70\%$), and an intermediate group. The first group is mainly characterized by a striking prevalence of longer genes among intronless genes compared with intron-containing ones. We could deduce a rule that, in genomes with a low presence of intron-containing genes, such genes are coding shorter proteins. However, there is an exception to this empiric rule, ie, *L. braziliensis*, which has a fraction $AL0_{ex}/AL1_{ex}$ similar to genomes with high P_g. An empiric rule for the second group may be formulated that there is a (linear) positive correlation between a proportion of intron-containing genes in a genome and a fraction $AL0_{ex}/AL1_{ex}$ while values of a fraction are lower than 1. Unfortunately, we have an exception to this rule as well, ie, *Bigelowiella natans*, which has a surprisingly high value of the ratio $AL0_{ex}/AL1_{ex}$. Regarding the central group, we may say only that it has an intriguing configuration that requires further investigation.

Chromosome-averaged statistical parameters

Let us consider the average parameters l_{ex}, n_{ex}, and a_{ex}. A scatter-plot of a_{ex} versus l_{ex} is shown in Figure 2B for Protista and illustrates the statement made previously that

Table 3 Species dependent exonic parameters

Organism	Kingdom/supergroup	AN_{ex}	AL_{ex}	AA_{ex}	$AN1_{ex}$	$AL1_{ex}$	P_g	$AL0_{ex}$
AD	Plantae	5.255	1243	412	6.404	1322	78.65	951
AF	Fungi	2.918	1465	668	3.453	1513	78.14	1289
BN	Protista/Rhizaria	4.054	960	359	4.580	907	85.37	1142
CE	Animalia	6.283	1284	213	6.428	1306	97.34	457
CF	Animalia	10.697	1696	232	11.450	1762	92.74	843
CG	Fungi	1.016	1509	1504	2.025	662	1.56	1522
CN	Fungi	6.271	1611	319	6.445	1627	96.87	1123
DH	Fungi	1.057	1387	1357	2.075	1070	5.41	1402
DM	Animalia	3.914	1824	503	4.686	2007	81.12	906
EC	Fungi	1.008	1071	1069	2.143	438	0.71	1075
EG	Fungi	1.048	1472	1452	2.035	874	4.58	1485
GT	Protista/Chromista	1.033	939	930	2.000	583	3.29	952
GZ	Fungi	3.261	1531	623	3.590	1553	82.04	1359
HA	Protista/Chromista	1.000	1019	1019	–	–	0.00	1019
HS	Animalia	8.868	1533	280	10.167	1656	85.76	790
KL	Fungi	1.025	1418	1409	2.017	760	2.47	1435
LB	Protista/Protozoa	1.012	1905	1882	2.040	3854	1.16	1882
MG	Fungi	2.844	1394	654	3.480	1468	74.34	1179
MM	Animalia	8.248	1457	302	9.658	1575	83.53	848
MS	Plantae/Viridiplantae	1.516	1488	1166	2.447	1636	34.58	1407
NC	Fungi	2.703	1476	694	3.136	1505	79.73	1366
OL	Plantae/Viridiplantae	1.279	1253	1100	2.344	1388	20.06	1222
OS	Plantae	4.846	1237	440	6.054	1348	75.96	890
PF	Protista/Chromalveolata	2.440	2238	1490	3.603	2131	55.30	2377
PK	Protista /Chromalveolata	2.591	2189	1486	4.094	2021	51.43	2373
PS	Fungi	1.408	1495	1227	2.551	1732	26.28	1409
PT	Protista/Chromalveolata	3.337	1583	583	3.803	1674	83.37	1128
SC	Fungi	1.055	1482	1444	2.035	1434	5.31	1485
SP	Fungi	1.951	1413	1040	3.098	1305	45.36	1501
TA	Protista/Chromalveolata	3.775	1581	785	4.964	1525	69.96	1716
UM	Fungi	1.782	1839	1439	3.025	1649	38.60	1961
YL	Fungi	1.158	1458	1339	2.131	1637	13.92	1428
Total		3.121	1579	1057	4.204	1606	42.97	1417

the averages of l_{ex} and a_{ex} were fairly similar for different chromosomes of the same species but, as a rule, rather distant for different species. Moreover, six separate groups of points may be observed in Figure 2B.

We colored all points using four colors relating to four Protista supergroups, ie, Chromalveolata (*Plasmodium falciparum*, *P. knowlesi*, *Paramecium tetraurelia*, and *Theileria annulata*), Chromista (*Guillardia theta, H. anderenii*), Protozoa (*L. braziliensis*), and Rhizaria (*B. natans*, see Table 1). Analyzing the contents of the groups presented in Figure 2, one can suppose that the divisions follow their taxonomy. Indeed, scatter-plots of a_{ex} vs n_{ex} (Figure 2A) and a_{ex} vs l_{ex} (Figure 2b) clearly show six separate groups of chromosomes; *B. natans* chromosomes belonging to Rhizaria form the left-most group, *G. theta* and *H. anderenii* chromosomes belonging to Chromista are located together, and Protozoa (*L. braziliensis*) form the third cluster. Chromosomes belonging to Chromalveolata form three clusters, according to their phylum and class, ie, Apicomplexa Plasmodium (*P. falciparum* and *P. knowlesi*), Apicomplexa Theileria (*T. annulata*), and a single chromosome of Paramecium (*P. tetraurelia*). These scatter-plots show that the three parameters a_{ex}, n_{ex}, and l_{ex} are sufficient for successful classification of 76 chromosomes to eight unicellular organisms.

The same conclusion regarding classification mirroring the phyla taxonomy can be made following an analysis of the matching chromosomal parameters for Animalia. Scatter-plots of a_{ex} versus l_{ex} and a_{ex} versus n_{ex} for Animalia are shown in Figure 3. Points related to averages l_{ex} and a_{ex} were related to different chromosomes of the same species and were located quite close to one another, whereas points related to chromosomes of different species are placed distant from one another. Striking exceptions are the points associated with chromosome 4 of *D. melanogaster* and chromosome 7 of *Mus musculus*. These points form clusters of a single member clearly disjointed from other

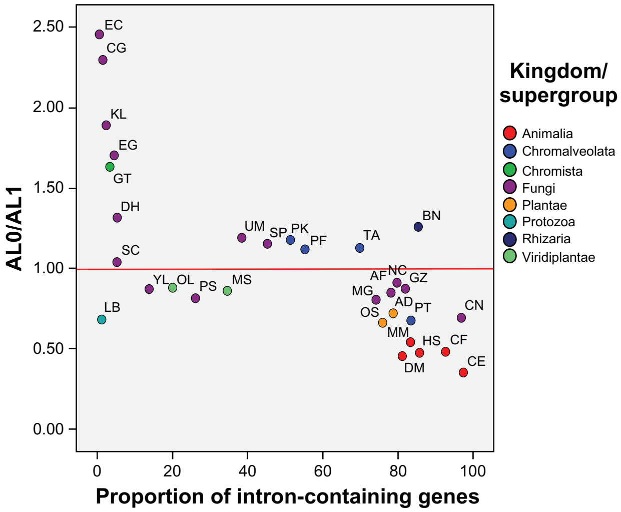

Figure 1 Scatter-plot of P_g showing the proportion of intron-containing genes in a genome on the x-axis versus the ratio between ALO_{ex} (an average length of an intronless gene) and $AL1_{ex}$ (an average net length of all exons in an intron-containing gene) on the y-axis for all 32 genomes. The red line marks the level of equality of ALO_{ex} and $AL1_{ex}$.

groups. All other points form three separate groups, which may be observed in Figure 3. The two parameters l_{ex} and a_{ex} separately cluster five chromosomes of *D. melanogaster* in one group, six chromosomes of *C. elegans* in another group, and all 39 chromosomes of *Canis familiaris*, *H. sapiens*, and *M. musculus* in the third group.

Let us repeat our observations deduced from Figure 3 relating to the phyla. We colored all points in three colors related to three animal phyla (see Table 1), ie, Arthropoda, Chordata, and Nemata. Figure 3a presents a scatter-plot of a_{ex} versus n_{ex} and clearly shows three separate groups of chromosomes and two outliers. *C. familiaris*, *H. sapiens*, and *M. musculus* chromosomes belonging to Chordata Mammalia form the left-most group; *C. elegans* chromosomes belonging to Nemata Caenorhabditis appear in the second left group; and the points belonging to *D. melanogaster* (Arthropoda Insecta) appear in the right group. Two chromosomes, ie, DM4 (the shortest chromosome of *D. melanogaster*) and MM07, form two separate groups, each one with a single member. The *C. elegans* chromosomes have the greatest exon density (n_{ex}) and the shortest exons (l_{ex}) among all the animal chromosomes studied.

Clustering of genomes by species-averaged statistical parameters

After the relatively satisfying success of partial clustering based on only three chromosomal characteristics, our next objective was to cluster all 32 genomes. We took seven species-averaged exon parameters mentioned previously, ie, AN_{ex} (average number of exons in a gene per genome), AL_{ex} (average net length of all exons in a gene per genome), AA_{ex} (average exon length in a gene per genome), $AN1_{ex}$ (average number of exons in an intron-containing gene per genome), $AL0_{ex}$ = average (over a genome) length of an intronless gene, $AL1_{ex}$ (average net length of all exons in an intron-containing gene per genome), and P_g (proportion of intron-containing genes in a genome expressed as a percentage). The expectation was that clustering would generally follow the kingdom/supergroup/phylum classification. However, the

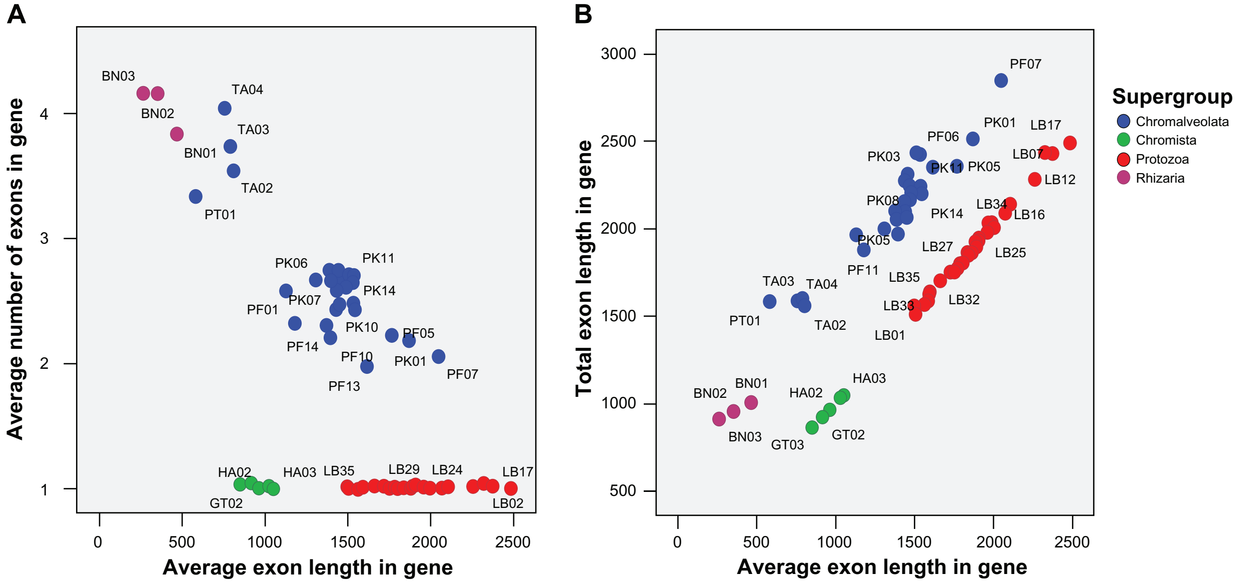

Figure 2 Scatter-plot for 76 processed chromosomes of eight Protista species, colored by four supergroups. Plot presents the average exon length per gene a_{ex} (x-axis) **A**) versus the average number of exons per gene n_{ex} (y-axis) and **B**) versus the average net exon length per gene l_{ex} (y-axis).

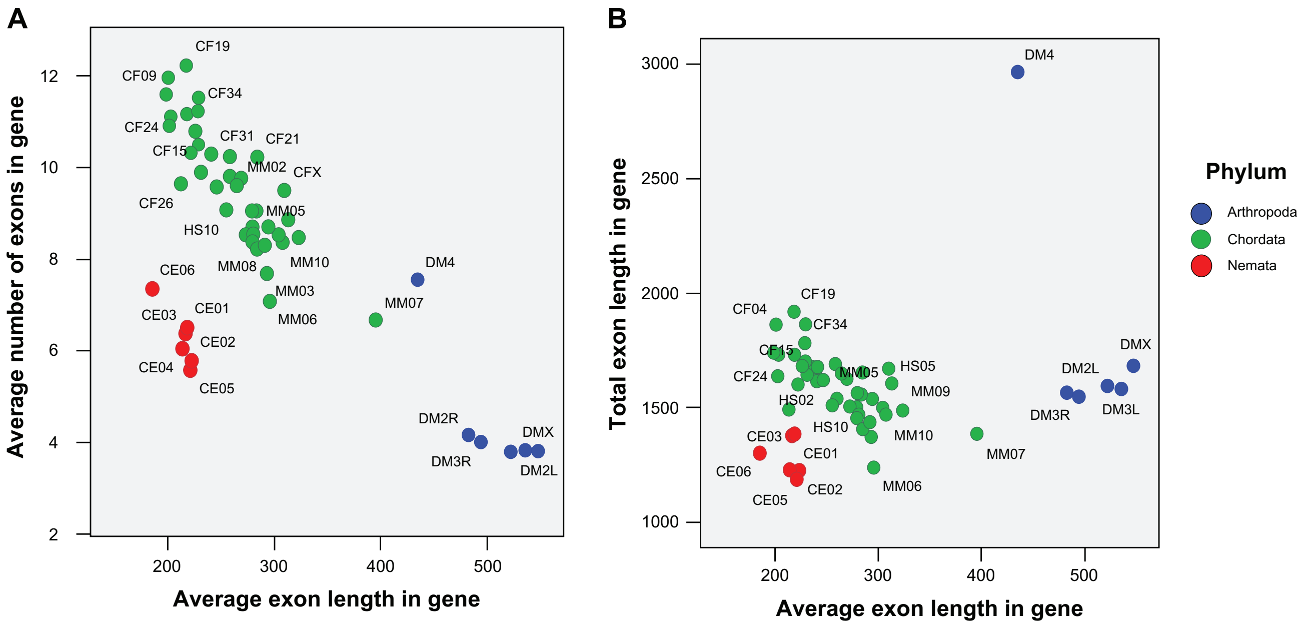

Figure 3 Scatter-plot for 59 processed chromosomes of five Animalia species, colored by three phyla. Plot presents the average exon length per gene a_{ex} (x-axis) **A**) versus the average number of exons per gene n_{ex} (y-axis) and **B**) versus the average net exon length per gene l_{ex} (y-axis).

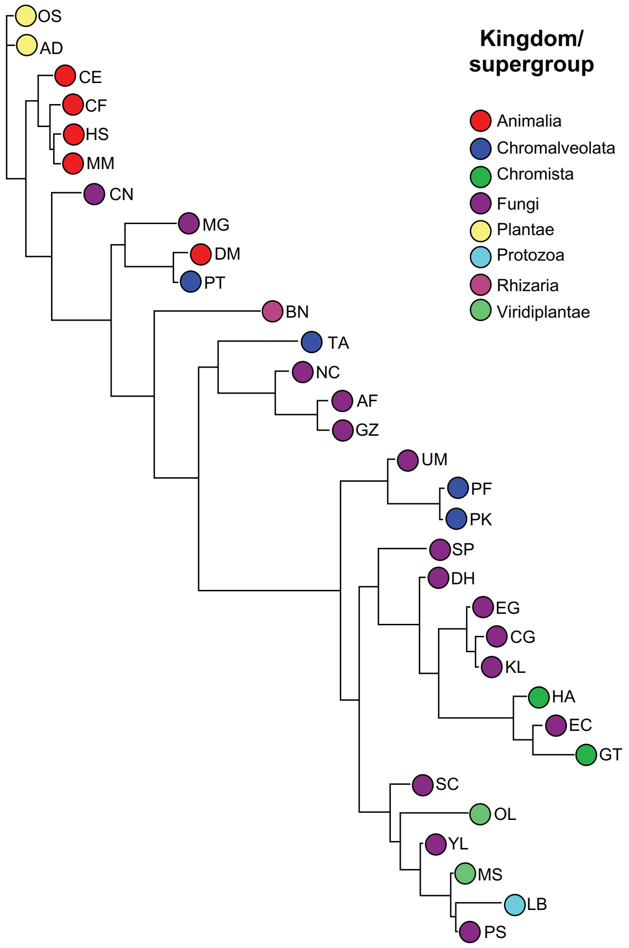

Figure 5 Dendrogram of 32 processed genomes obtained by neighbor-joining clustering technique and based on distances among four principal components obtained by factor analysis of AN_{ex}, AL_{ex}, AA_{ex}, $AN1_{ex}$, $AL1_{ex}$, and $AL0_{ex}$.

results were poor (data not shown). Assuming that a peculiar relationship between a parameter P_g and other parameters (see Figure 1) may negatively influence clustering, we excluded this parameter from further consideration.

At this point, we tried to cluster genomes of 32 different organisms using six parameters, namely AN_{ex}, AL_{ex}, AA_{ex}, $AN1_{ex}$, $AL1_{ex}$, and $AL0_{ex}$. As a first stage, we applied neighbor-joining clustering using standardized distances among the vectors (AN_{ex}, AL_{ex}, AA_{ex}, $AN1_{ex}$, $AL1_{ex}$, $AL0_{ex}$) and applying the Neighbor program. The dendrogram presented in Figure 4 was drawn by the TreeView program. As one can see, some organisms of the same kingdom/supergroup are distributed compactly along the tree. Nevertheless, not all species belonging to the same class form a monophyletic cluster. Mice (*M. musculus*), dogs (*C. familiaris*), and humans (*H. sapiens*) are located together, but flies (*D. melanogaster*), which form a cluster together with Protista/Chromalveolata, *T. annulata*, appear too far away from other Animalia. Viridiplantae species are placed distantly, and Protista are distributed along the tree in a strange manner. Such a classification, although better than the classification produced by seven parameters, cannot be considered adequate.

These discrepancies could be explained at least partially by the cross-dependencies of all the parameters considered. Therefore, the way to improve clustering is to replace these parameters by independent (orthogonal) parameters that could be obtained, eg, from results of a factor analysis of their correlation matrix as principal components.

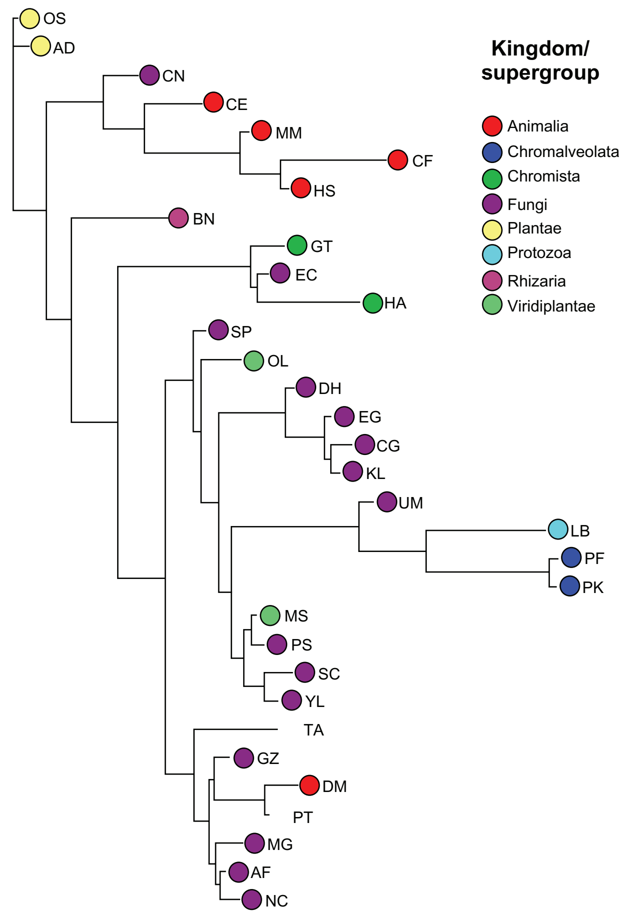

Figure 4 Dendrogram of 32 processed genomes obtained by the neighbor-joining clustering technique and based on standardized distances among parameters AN_{ex}, AL_{ex}, AA_{ex}, $AN1_{ex}$, $AL1_{ex}$, and $AL0_{ex}$.

The factor analysis led us to the synthesis of the following successive logical structure:

- Dividing the system into sets of "elementary" components, ie, all of the aforementioned genomic characteristics (AN_{ex}, AL_{ex}, AA_{ex}, $AN1_{ex,}$ $AL1_{ex}$, $AL0_{ex}$)
- Analysis of the relationships of these components in species
- Revealing system-forming relationships
- Description of the structure of the system (model) and its properties.

As shown in Table S2, four principal components are responsible for 99.4% of the organization of the whole system, and the first two describe 86.2% of the whole variability of the system. Four principal components (Table S3) have been used in genome clustering based on neighbor-joining, *k*-means, and partitioning around medoids. Results of neighbor-joining clustering are presented in Figure 5.

There are certain improvements comparing the clustering presented in Figure 4. Viridiplantae species are placed closely, and Protista are distributed along the tree less strangely than in Figure 4. However, *D. melanogaster* couples with the Protista *T. annulata* again.

Results of *k*-means (Table S4) clustering are very similar (practically identical) to the neighbor-joining results shown earlier. These *k*-means results are shown in Table S4. Results of partitioning around medoids clustering are presented in Table S5. These results are similar to neighbor-joining results as well. However, there are some additional improvements in partitioning genomes among different clusters. In general, the results show a high consistency of partitioning, in spite of differences in clustering techniques. Careful examination of Table S5 reveals hierarchic partitioning of organisms. Interestingly, partitioning around medoids clustering is not a hierarchic algorithm and should not necessarily produce any hierarchy. In our case of application of partitioning around medoids clustering to four principal components obtained by factor analysis, a strictly hierarchic structure is produced. In fact, the *k*-medoids clustering was performed for different values of *k* between 2 and 20, and it was observed that the clustering for a given value of *k* is always a strict subclustering of the clustering for *k–1*. This may be interpreted as existence of an intrinsic hierarchic structure of principal components analysis data. This may, in turn, serve as additional evidence of variance in the evolutionary nature of exon–intron structure.

Discussion

The origin of introns remains a mystery, and certain questions in molecular evolution are being investigated by in silico analysis of intron–exon structures in various organisms. To facilitate such studies, while taking advantage of the burgeoning amount of sequence data now available, we undertook a statistical analysis of the exon–intron structure for nearly all completely sequenced eukaryotic genomes in order to reveal general and genome-specific features of eukaryotic genes. We went through all of the protein-coding genes in each chromosome separately and calculated the portion of intron-containing genes and average values of the net length of all the exons in a gene, the number of exons, and the average length of an exon. Furthermore, we tried to determine the most appropriate approach to classifying eukaryotic chromosomes, according to these simple exon–intron statistics.

One of the main conclusions of the studies by Kaplunovsky et al[2] and Atambayeva et al[13] was that a positive correlation exists between the number of introns in a gene and the length of the corresponding protein (and equivalently the net length of all the exons of the gene). Here, like Kaplunovsky et al,[3] we confirmed the observation of Ivashchenko et al[15] that, for all fungal genomes with a proportion of intron-containing genes higher than 30%, gene size and total exon length depend on the intron number in a linear manner. The correlation problem is irrelevant for organisms with an extremely low proportion of intron-containing genes, such as yeasts, Protista/Chromista, and Protista/Protozoa.

In a previous publication,[2] we reported that intragenomic variation is substantially smaller than intergenomic variance in almost all fungal genomes. In other words, we found that the laws of exon–intron statistics are specific to genomes rather than to individual chromosomes. In this respect, the similarity in exon–intron structures for dogs (*C. familiaris*), mice (*M. musculus*), and humans (*H. sapiens*) is so striking that intragenomic and intergenomic variances of the sets a_{ex}, l_{ex}, and n_{ex} in various chromosomes are practically undetectable (see Table 2). A similar statement can be made regarding two plants in this study, ie, *Arabidopsis* and rice, and thus we confirmed the observations made by Atambayeva et al.[13]

Noteworthy is the similarity in the exon–intron structures of an insect, *D*. melanogaster, and a protist, *T. annulata* (see Table 3). Neither environmental habitat factors nor the evolutionary history of organisms provide any clue to solving the mystery of the proximity of these two genomes on the genome tree based on exon–intron characteristics. Perhaps the appearance of other eukaryotes in the data set of completely sequenced genomes will provide the answer.

The main advances of this study over previous research[2,3,5,8–15] lie in the larger amount of genomes considered and the concentrated efforts made to determine the most appropriate approach for clustering based on exonic characteristics. We checked a few procedures of clustering based on exon–intron structure features averaged over intron-containing or intronless genes. As a result, we conclude that the most successful procedure should be based on distances between four principal components obtained by factor analysis and followed by application of clustering algorithms. The consistency of recovered cluster structures may be considered evidence of hidden evolutionary resemblance.

We concentrated our efforts on comparison of exonic parameters, while planning to work on intron lengths later. Clearly, the exon–intron structures of eukaryotic genes have many important parameters that we did not consider in this work, and we intend to pursue these in future research. In particular, the ratio of exon and intron lengths promises to be an important feature of a gene. In some genomes, the intron length is comparable with the exon length, ie, in unicellular eukaryotes,[4,5] plants,[5,27] and particular animals.[5–7] In general, introns are longer than exons in mammalian genes.[14] Correlations of intronic characteristics with such genomic properties as gene density would be a goal for further research as well.

Acknowledgments

We are grateful to S Hosid and Z Volkovich for technical assistance with this research.

Disclosure

The authors report no conflicts of interest in this work.

References

1. Bolshoy A, Volkovich Z, Kirzhner V, Barzily Z. Genome clustering: From Linguistic Models to Classification of Genetic Texts. *Studies in Computational Intelligence Series*. Kacprzyk J, editor. Berlin, Heidelberg: Springer-Verlag; 2010.
2. Kaplunovsky A, Khailenko VA, Bolshoy A, Atambayeva SA, Ivashchenko AT. Statistics of exon lengths in animals, plants, fungi, and protists. *Int J Biol Life Sci.* 2009;1:139–144.
3. Kaplunovsky A, Zabrodsky D, Volkovich Z, Ivashchenko AT, Bolshoy A. Statistics of exon lengths in fungi. *Open Bioinformatics J.* 2010;4:31–40.
4. Kupfer DM, Drabenstot SD, Buchanan KL, et al. Introns and splicing elements of five diverse fungi. *Eukaryot Cell.* 2004;3:1088–1100.
5. Deutsch M, Long M. Intron-exon structures of eukaryotic model organisms. *Nucleic Acids Res.* 1999;27:3219–3228.
6. Wendel JF, Cronn RC, Alvarez I, Liu B, Small RL, Senchina DS. Intron size and genome size in plants. *Mol Biol Evol.* 2002;19: 2346–2352.
7. Sakharkar MK, Chow VT, Kangueane P. Distributions of exons and introns in the human genome. *In Silico Biol.* 2004;4:387–393.
8. Roy SW, Penny D. Intron length distributions and gene prediction. *Nucleic Acids Res.* 2007;35:4737–4742.
9. Naora H, Deacon NJ. Relationship between the total size of exon and introns in the protein-coding genes of higher eukaryotes. *Proc Natl Acad Sci U S A.* 1982;79:6196–6200.
10. Hawkins JD. A survey on intron and exon lengths. *Nucleic Acids Res.* 1988;16:9893–9908.
11. Kriventseva EV, Gelfand MS. Statistical analysis of the exon–intron structure of higher and lower eukaryote genes. *J Biomol Struct Dyn.* 1999;17:281–288.
12. Ivashchenko AT, Atambayeva SA. Variation in lengths of introns and exons in genes of the Arabidopsis thaliana nuclear genome. *Russ J Genet.* 2004;40:1179–1181.
13. Atambayeva SA, Khailenko VA, Ivashchenko AT. Intron and exon length variation in arabidopsis, rice, nematode, and human. *Mol Biol.* 2008;42:312–320.
14. Ivashchenko AT, Khailenko VA, Atambayeva SA. Variation of the lengths of exons and introns in human genome genes. *Russ J Genet.* 2009;45:16–22.
15. Ivashchenko AT, Tauasarova MI, Atambayeva SA. Exon–intron structure of genes in complete fungal genomes. *Mol Biol.* 2009;43:24–31.
16. Zhu LC, Zhang Y, Zhang W, Yang SH, Chen JQ, Tian DC. Patterns of exon–intron architecture variation of genes in eukaryotic genomes. *BMC Genomics.* 2009;10:12.
17. Collins FS, Lander ES, Rogers J, Waterston RH. International Human Genome Sequencing Consortium: Finishing the euchromatic sequence of the human genome. *Nature.* 2004;431:931–945.
18. Schwarz EM, Antoshechkin I, Bastiani C, et al. WormBase: Better software, richer content. *Nucleic Acids Res*. 2006;34 Database Issue: D475–D478.

19. Drysdale RA, Crosby MA, FlyBase C. FlyBase: Genes and gene models. *Nucleic Acids Res.* 2005;33:D390–D395.
20. Haas BJ, Wortman JR, Ronning CM, et al. Complete reannotation of the Arabidopsis genome: Methods, tools, protocols and the final release. *BMC Biol.* 2005;3:7.
21. Logsdon JMJ, Stoltzfus A, Doolittle WF. Molecular evolution: Recent cases of spliceosomal intron gain? *Curr Biol.* 1998;8:R560–R563.
22. Archibald JM, O'Kelly CJ, Doolittle WF. The chaperonin genes of jakobid and jakobid-like flagellates: Implications for eukaryotic evolution. *Mol Biol Evol.* 2002;19:422–431.
23. Loftus BJ, Fung E, Roncaglia P, et al. The genome of the basidiomycetous yeast and human pathogen Cryptococcus neoformans. *Science.* 2005;307:1321–1324.
24. Martinez D, Berka RM, Henrissat B, et al. Genome sequencing and analysis of the biomass-degrading fungus Trichoderma reesei (syn. Hypocrea jecorina). *Nat Biotechnol.* 2008;26:553–560.
25. Gudlaugsdottir S, Boswell DR, Wood GR, Ma J. Exon size distribution and the origin of introns. *Genetica.* 2007;131:299–306.
26. Saitou N, Nei M. The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol Biol Evol.* 1987;4:406–425.
27. Ren XY, Vorst O, Fiers MWEJ, Stiekema WJ, Nap JP. In plants, highly expressed genes are the least compact. *Trends Genet.* 2006;22:528–532.

Supplementary tables

Table S1 Exonic chromosomal parameters of *Plasmodium knowlesi*

Chromosome	n_{ex}	l_{ex}	a_{ex}	nl_{ex}	$L1_{ex}$	$L0_{ex}$	p_c
PK01	2.18 ± 0.32	2513 ± 373	1871 ± 311	3.79 ± 0.57	2326 ± 644	2651 ± 442	42.29
PK02	2.61 ± 0.39	2234 ± 384	1476 ± 340	3.88 ± 0.58	1991 ± 438	2541 ± 671	55.84
PK03	2.74 ± 0.36	2313 ± 346	1455 ± 252	4.16 ± 0.51	2216 ± 506	2432 ± 445	55.10
PK04	2.70 ± 0.38	2159 ± 291	1436 ± 263	4.08 ± 0.56	1893 ± 359	2487 ± 468	55.17
PK05	2.75 ± 0.38	2048 ± 304	1387 ± 277	4.30 ± 0.62	1773 ± 363	2356 ± 530	52.90
PK06	2.67 ± 0.38	1996 ± 284	1308 ± 205	4.33 ± 0.62	1972 ± 452	2020 ± 345	50.00
PK07	2.65 ± 0.28	2093 ± 218	1396 ± 189	4.07 ± 0.41	1974 ± 184	2232 ± 340	53.85
PK08	2.47 ± 0.23	2062 ± 216	1451 ± 193	3.71 ± 0.34	1746 ± 263	2436 ± 346	54.18
PK09	2.69 ± 0.24	2226 ± 203	1492 ± 175	4.15 ± 0.36	2016 ± 274	2469 ± 308	53.64
PK10	2.43 ± 0.24	2114 ± 305	1431 ± 205	3.84 ± 0.36	2109 ± 551	2119 ± 351	50.32
PK11	2.70 ± 0.26	2244 ± 330	1538 ± 203	4.44 ± 0.41	2025 ± 288	2459 ± 345	49.48
PK12	2.59 ± 0.20	2213 ± 187	1483 ± 145	4.17 ± 0.33	2119 ± 281	2308 ± 248	50.29
PK13	2.63 ± 0.29	2235 ± 225	1527 ± 186	4.30 ± 0.49	2071 ± 323	2395 ± 311	49.56
PK14	2.42 ± 0.19	2195 ± 166	1542 ± 155	4.01 ± 0.32	2062 ± 238	2315 ± 256	47.36
Total	2.59 ± 0.07	2185 ± 66	1487 ± 55	4.11 ± 0.11	2019 ± 93	2358 ± 95	51.43

Table S2 Total variance explained

Component	% of variance	Cumulative %
1	61.413	61.413
2	24.809	86.222
3	8.290	94.512
4	4.871	99.383

Table S3 Component matrix extraction method: principal component analysis with four extracted components

Component matrix (a)

Abbr	Kingdom/supergroup	Organism	Component			
			1	2	3	4
AD	Plantae	*Arabidopsis thaliana*	–0.928	–0.314	0.069	0.172
AF	Fungi	*Aspergillus fumigatus*	–0.566	0.785	–0.065	0.236
BN	Protista/Rhizaria	*Bigelowiella natans*	–0.601	–0.545	–0.210	0.544
CE	Animalia	*Caenorhabditis elegans*	–0.929	–0.296	0.189	–0.103
CF	Animalia	*Canis familiaris*	–0.956	–0.240	–0.025	–0.119
CG	Fungi	*Candida glabrata*	0.891	–0.314	–0.242	–0.221
CN	Fungi	*Cryptococcus neoformans*	–0.986	0.040	–0.123	–0.103
DH	Fungi	*Debaryomyces hansenii*	0.978	–0.204	0.046	–0.003
DM	Animalia	*Drosophila melanogaster*	–0.747	0.478	0.345	–0.298
EC	Fungi	*Encephalitozoon cuniculi*	0.582	–0.805	–0.082	0.085
EG	Fungi	*Eremothecium gossypii*	0.946	–0.250	–0.137	–0.152
GT	Protista/Chromista	*Guillardia theta*	0.417	–0.841	0.139	0.317
GZ	Fungi	*Gibberella zeae*	–0.555	0.769	–0.276	0.096
HA	Protista/Chromista	*Hemiselmis anderenii*	0.615	–0.704	–0.198	–0.064
HS	Animalia	*Homo sapiens*	–0.967	–0.229	0.050	–0.091
KL	Fungi	*Kluyveromyces lactis*	0.914	–0.346	–0.160	–0.139
LB	Protista/Protozoa	*Leishmania braziliensis*	0.677	0.628	–0.380	0.020
MG	Fungi	*Magnaporthe grisea*	–0.776	–0.492	0.237	0.314
MM	Animalia	*Mus musculus*	–0.965	–0.255	0.037	–0.038
MS	Plantae/Viridiplantae	*Micromonas* sp. RCC299	0.794	0.472	0.363	0.121
NC	Fungi	*Neurospora crassa*	–0.186	0.899	–0.225	0.287
OL	Plantae/Viridiplantae	*Ostreococcus lucimarinus*	0.737	–0.030	0.542	0.400
OS	Plantae	*Oryza sativa*	–0.927	–0.282	0.181	0.149
PF	Protista/Chromalveolata	*Plasmodium falciparum*	0.629	0.685	–0.327	–0.168
PK	Protista/Chromalveolata	*Plasmodium knowlesi*	0.646	0.632	–0.396	–0.152
PS	Fungi	*Pichia stipitis*	0.752	0.468	0.439	0.136
PT	Protista/Chromalveolata	*Paramecium tetraurelia*	–0.706	0.642	0.179	–0.239
SC	Fungi	*Saccharomyces cerevisiaei*	–0.969	0.119	0.215	0.026
SP	Fungi	*Schizosaccharomyces pombe*	0.890	0.063	–0.234	0.381
TA	Protista/Chromalveolata	*Theileria annulata*	–0.221	0.342	–0.815	0.391
UM	Fungi	*Ustilago maydis*	0.852	0.448	–0.253	–0.097
YL	Fungi	*Yarrowia lipolytica*	0.862	0.296	0.390	0.130

Table S4 Results obtained by *k*-means clustering technique and based on four principal components obtained by factor analysis of AN_{ex}, AL_{ex}, AA_{ex}, $AN1_{ex}$, $AL1_{ex}$, $AL0_{ex}$

	2	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
MG	1	2	2	2	6	7	7	8	5	8	1	6	7	2	11	8	13	12	7
AD	1	1	3	5	4	3	2	9	3	11	6	2	6	11	16	9	11	11	8
OS	1	1	3	5	4	3	2	9	3	11	6	2	6	11	16	9	11	11	8
HS	1	1	3	5	4	3	2	9	3	11	9	8	4	10	8	15	16	1	5
CF	1	1	3	5	4	3	2	9	3	11	9	8	4	10	8	15	16	1	5
MM	1	1	3	5	4	3	2	9	3	11	9	8	4	10	8	15	16	1	5
CE	1	1	3	5	4	3	2	9	3	11	8	8	5	10	8	15	16	1	14
CN	1	1	3	5	4	3	2	9	3	11	11	8	11	10	15	16	8	10	19
AF	1	2	2	2	5	4	3	2	8	5	5	10	1	3	14	11	1	3	4
GZ	1	2	2	2	5	4	3	2	8	5	5	10	1	3	14	11	1	3	4
DM	1	2	2	2	6	7	7	8	5	8	1	6	13	2	4	3	13	8	12
NC	1	2	2	2	5	4	3	2	8	5	5	10	1	3	14	11	1	3	4
PT	1	2	2	2	6	7	7	8	5	8	1	6	13	2	4	3	13	8	12
BN	1	1	3	5	3	3	2	7	7	7	4	13	8	9	9	13	5	9	9
TA	1	2	2	2	5	4	3	5	9	3	3	9	2	12	3	6	14	17	17
CG	2	3	1	3	2	6	8	1	2	2	10	12	10	5	6	1	6	16	15
EG	2	3	1	3	2	6	8	1	2	2	10	12	10	5	6	1	6	16	15
KL	2	3	1	3	2	6	8	1	2	2	10	12	10	5	6	1	6	16	15
DH	2	3	1	3	2	6	8	1	2	2	10	12	10	4	6	2	18	5	18
EC	2	3	1	3	2	5	4	6	10	9	2	5	3	14	2	12	7	14	10
GT	2	3	1	3	2	5	4	6	10	10	2	5	3	14	2	17	4	14	6
HA	2	3	1	3	2	5	4	6	10	9	2	5	3	14	2	12	7	14	10
LB	2	3	4	4	1	1	1	3	1	1	12	4	14	7	10	4	3	2	3
MS	2	3	4	4	1	1	1	3	1	1	12	4	14	15	10	4	15	18	20
PF	2	3	4	1	1	2	6	4	4	4	7	3	12	1	5	10	10	15	2
PK	2	3	4	1	1	2	6	4	4	4	7	3	12	1	7	10	10	15	2
UM	2	3	4	1	1	2	6	4	4	4	7	3	12	1	12	10	2	15	13
PS	2	3	4	4	1	1	1	3	1	1	12	4	14	15	10	4	15	19	20
SC	2	3	4	4	1	1	5	3	6	1	12	7	9	6	10	7	17	6	16
YL	2	3	4	4	1	1	1	3	6	1	12	7	14	15	10	7	15	4	20
SP	2	3	4	1	1	6	5	1	2	6	10	1	9	8	1	5	12	7	1
OL	2	3	4	4	1	1	5	3	6	1	12	11	9	13	13	14	9	13	11

Table S5 Results obtained by partitioning around medoids clustering technique and based on four principal components obtained by factor analysis of AN_{ex}, AL_{ex}, AA_{ex}, $AN1_{ex}$, $AL1_{ex}$, $AL0_{ex}$

	2	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
MG	2	3	3	3	3	3	3	3	3	3	12	12	12	12	12	12	12	12	12
AD	2	2	2	2	2	2	2	2	2	2	2	2	2	2	16	16	16	16	16
OS	2	2	2	2	2	2	2	2	2	2	2	2	2	2	16	16	16	16	16
HS	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
CF	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
MM	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
CE	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
CN	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	17	17	17	17
AF	2	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
GZ	2	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	20
NC	2	3	3	3	3	3	3	3	3	3	3	3	14	14	14	14	14	14	14
CG	1	1	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
EG	1	1	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
KL	1	1	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
DH	1	1	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	19	19
PF	1	1	1	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
PK	1	1	1	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
UM	1	1	1	5	5	5	5	5	5	5	5	5	5	5	5	5	18	18	18
EC	1	1	4	4	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6
GT	1	1	4	4	6	6	6	6	6	6	6	6	6	15	15	15	15	15	15
HA	1	1	4	4	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6
LB	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
MS	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
PS	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
SC	1	1	1	1	1	1	1	1	1	1	1	13	13	13	13	13	13	13	13
YL	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
DM	2	3	3	3	3	7	8	8	8	8	8	8	8	8	8	8	8	8	8
PT	2	3	3	3	3	7	8	8	8	8	8	8	8	8	8	8	8	8	8
TA	2	3	3	3	3	3	7	9	9	9	9	9	9	9	9	9	9	9	9
BN	2	2	2	2	2	2	2	7	7	7	7	7	7	7	7	7	7	7	7
SP	1	1	4	4	4	4	4	4	10	10	10	10	10	10	10	10	10	10	10
OL	1	1	1	1	1	1	1	1	1	11	11	11	11	11	11	11	11	11	11

4. Discussion

The exon-intron structures of different organisms are quite different from each other, and the evolution of such structures raises many questions. We tried to address some of these questions with an accent on methods of revealing evolutionary factors based on the analysis of gene exon-intron structures using statistical analysis. Is it possible to answer to these question, using methods of statistical analysis by calculating the portion of intron-containing genes and average values of the net length of all the exons in a gene, the number of the exons, and the average length of an exon? We found striking similarities between all of these average properties of chromosomes of the same species and significant differences between properties of the chromosomes belonging to species of different divisions or kingdoms. Comparing those chromosomal and genomic averages, we have developed a technique of clustering based on characteristics of the exon-intron structure. This technique of clustering separates different species, grouping them according to a taxonomy. The main conclusion of this work is that the statistical properties of exon-intron organizations of genes are the genome-specific features preserved by evolutionary processes.

In the previous publications (Kaplunovsky *et al.,* 2009-2011, Atambayeva *et al.,* 2008) was found that intragenomic variation is substantially smaller than intergenomic variance practically in all genomes. In other words, we found that the laws of exon-intron statistics are specific to genomes rather than to individual chromosomes. Here, was found that similarity in exon-intron structures of dogs (CF), mice (MM), and humans (HS) is so striking that intragenomic and intergenomic variances of the sets a_{ex}, $l_{ex,}$ and n_{ex} in various chromosomes are practically indistinguishable. Similar statement regarding two plants of study - Arabidopsis and rice - holds as well. Similarity in exon-intron structures of an insect *D. melanogaster* and a protist *T. annulata* is eye-catching. Neither environmental habitat factors nor evolutionary history of organisms provide any clue for the mystery of these two genomes proximity along the genome tree based on exon-intron characteristics. May be, appearance of other Eukaryotes in the dataset of completely sequenced genomes will bring an answer to it.

The purpose of this research has been to determine the most appropriate approach to classify eukaryotic chromosomes, according to these simple exon-intron statistics. One of the main conclusions of the studies of Kaplunovsky et al (2009) and Atambayeva et al (2008) was that exists positive correlation between the number of introns in a gene and the corresponding protein's length (and equivalently, the net length of all the exons of the gene). Here and in (Kaplunovsky *et al.,* 2010) we confirmed the statement of Ivashchenko et al (2009) that for all fungal genomes with proportion of intron-containing genes higher than 30%, gene size and total exon length linearly depend on the intron number. The correlation problem is irrelevant for the organisms with extremely low proportion of intron-containing genes, such as yeasts, Protista / Chromista and Protista / Protozoa.

We concentrated our efforts on comparison of exonic parameters, while planning to work on intron lengths later. Clearly, the exon-intron structures of eukaryotic genes have many important parameters that we did not consider in this work; we intend to pursue these in future research. In particular, the ratio between the exon and intron lengths promises to be an important feature of a gene. In some genomes the intron length is comparable with the exon length: in unicellular eukaryotes (Deutsch and Long, 1999; Kupfer, 2004), plants (Deutsch and Long, 1999; Ren *et al.,* 2006) and particular animals (Deutsch and Long, 1999; Wendel *et al.,* 2002; Sakharkar *et al.,* 2004). In general, introns are longer than exons in mammalian genes (Ivashchenko *et al*., 2009). Correlations of intronic characteristics with such genomic properties as gene density would be a goal for further research as well.

4.1. Comparison between Kingdoms

The exon-intron structures of different eukaryotic species are quite different from each other, and the evolution of such structures raises many questions. We try to address some of these questions using statistical analysis of whole genomes. We go through all the protein-coding genes in a genome and study correlations between the net length of all the exons in a gene, the number of the exons, and the average length of an exon. We also take average values of these features for each chromosome and study correlations between those averages on the chromosomal level. Our data show universal features of exon-intron structures common to animals, plants, and protists (specifically, *Arabidopsis thaliana, Caenorhabditis elegans, Drosophila melanogaster, Cryptococcus neoformans, Homo sapiens, Mus musculus, Oryza sativa,* and *Plasmodium falciparum).* We have verified linear correlation between the number of exons in a gene and the length of a protein coded by the gene, while the protein length increases in proportion to the number of exons. On the other hand, the average length of an exon always decreases with the number of exons. Finally, chromosome clustering based on average chromosome properties and parameters of linear regression between the number of exons in a gene and the net length of those exons demonstrates that these average chromosome properties are genome-specific features.

4.1.1. Average Numbers of Exons and Net Exon Lengths in Different Chromosomes

For each of 76 chromosomes of eight species, we have calculated the average parameters l_{ex} (net length of gene's exons), n_{ex} (number of exons in a gene) and a_{ex} (average exon length). These averages turned out to be pretty similar for different chromosomes of the same species but rather distant for different species. A scatter plot of the l_{ex} *vs* n_{ex}; shows clear clustering of the chromosomes by species. It also shows a wide separation between PF - a protist - and the other species (animals, fungi, and plants). The PF chromosomes have much longer average proteins (l_{ex}) and much lower exon density (n_{ex}) than all the other eukaryote chromosomes we have studied. Moreover, all species except PF have rather similar ranges of the l_{ex} parameter, but the n_{ex} fall into quite distinct regions on the plot for the DM (*D. melanogaster*) and CN, and more doubtful areas for plants (AD and OS) and mammals (*H. sapiens* and *M. musculus*).

A scatter-plot of the average exon length a_{ex} *vs* the average number of exons in a gene n_{ex} shows much better grouping of chromosomes belonging to the same species - all kingdoms are grouped separately. Still, the resolution is not sufficient and there is a slight overlapping between species from the same kingdom (M and HS, AD and OS). In addition, *C. elegans* chromosomes may be characterized by relatively short exons in average and rather big variation in intron density. To improve the resolution between the species, we are going to take a closer look at the relation between the average exon number and the average net exon length of a gene.

4.1.2. Relations between the Average Exon Number and the Average Net Length of Exons in a Gene

It was already shown (Atambaeva *et al.,* 2008) that the average exon length in *A. thaliana*, *O. sativa*, *C. elegans*, and *Homo sapiens* genes decreases with an increasing number of introns. In addition, positive linear correlation was observed between the sum of exon lengths and the number of exons. We can see the correlation between the net length of exons and the number of exons in 12156 genes on ten chromosomes of *H. sapiens* with significance $p< 0.001$.

Similar results have been obtained for all eight species. Each species is represented by a scatter-plot of L_{ex} *vs* N_{ex} with a linear regression. There are dramatic differences between average and maximal values of L_{ex} and N_{ex} for animals, plants, fungi, and protists, and especially between parameters a and b of the linear regression equation $y=a+bx$. In light of these differences, we decided to check if the regression parameters could be used in classification of genomes by their exon properties. We have calculated the linear regressions for all 76 processed chromosomes of all eight genomes. Our results show significant correlations between the protein lengths and the numbers of exons in all eight studied genomes. Their values testify to high reliability of the correlation.

Clustering based on the linear regression parameters a and b follows the major differences between species from different kingdoms, and some reasonably observable differences between species from the same kingdom. There are some exceptions, and we would like to eliminate them by using the R^2 parameter - percent of the explained variation - of the regression analysis. It has negligible value for protists, medium values for plants and fungi, and relatively high values for animals. Hopefully, combining all the parameters together would give a better resolution than looking at any two parameters at a time.

Our results show both general and genome-specific features of the exon-intron organization of eukaryotic genes. The most general feature found in all genomes is the positive correlation between the number of introns in a gene and the corresponding protein's length (or equivalently, the net length of all the exons of the gene). In addition, in all the genomes we have studied, the average exon length in a gene decreases with the number of those exons. But while these laws of exon-intron statistics are quite general, the correlation parameters are genome-specific. Moreover, they are specific to genomes rather than to individual chromosomes. Indeed, in the parameter space of average chromosome properties and linear regression parameters (between exon numbers and protein lengths), all chromosomes from the same genome form obvious clusters

4.2. Comparison within the Kingdom (Fungi)

All of the abovementioned chromosomal characteristics (n_{ex}, l_{ex}, a_{ex}, p_c, $l0_{ex}$, $n1_{ex}$, $l1_{ex}$) were calculated for all 140 chromosomes. The intragenomic variation was found to be pretty small everywhere, exactly as it was expected. We can see that there is the same proportion of intron-containing genes in all eight chromosomes, for example, for *A. fumigatus* P_c = 78.5±0.5%. Also, sets L_{ex} and N_{ex} in various chromosomes of *A. fumigatus* do not demonstrate significant differences. F-statistics comparing variances between and within groups of chromosomes is not significant; therefore, all chromosomes have only indistinguishable distributions of L_{ex} and N_{ex}. Analogical results were obtained for the chromosomal parameters of all other organisms as well. For all chromosomal characters of all genomes the differences between two chromosomes of an identical genome appeared not to be statistically significant. Would the differences between two chromosomes of two different species depend on the evolutionary distance between these two organisms? Would it be possible to identify an organism by a combination of chromosomal characters? As it appeared in our research, a pair of characters does not provide full partition of all species.

4.2.1. Species-averaged statistical parameters

In addition to parameters averaged over all genes, there are data related to intron-containing ($L1_{ex}$) and intronless genes ($AL0_{ex}$) separately. For the set of intronless genes, the parameters AL_{ex} and AA_{ex} are identical and equal to an average gene length $AL0_{ex}$. Some putative empirical rules may be observed. For example, regarding average gene lengths of intron-containing and intronless genes, it seems that if there is only a small amount of intron-containing genes in a genome, these genes are shorter in average than other intronless genes of the same genome. This property is especially strongly expressed in EC, CG, and KL, and also exists for EG, DH, SP, and UM. Another observation may be done regarding a lack of correlation between amounts of genes in a genome and other genomic statistical parameters.

4.2.2. Chromosome-averaged statistical parameters

Let us consider the average parameters l_{ex}, n_{ex} and a_{ex}. We can see that the averages l_{ex} and a_{ex} turned out to be pretty similar for different chromosomes of the same species but rather distant for different species. Moreover, five separate groups of points may be observed. The two parameters l_{ex} and a_{ex} cluster separately all 14 chromosomes of *C. neoformans* (CN) in one group, 8 chromosomes of *E. cuniculi* (EC) in another group, and all 23 chromosomes of *U. maydis* (UM) in the third group. All other points are distributed between two additional groups. Analyzing the contents of these groups, one can suppose that the partitions follow fungal taxonomy. *Ascomycota Pezizomycotina*, *Ascomycota Saccharomycotina*, *Ascomycota Taphrinomycotina, Basidiomycota Agaricomycotina, Basidiomycota Ustilaginomycotina*, and *Microsporidia Apansporoblastina.* We can also see that CN chromosomes have the greatest exon density (n_{ex}) and the shortest exons (l_{ex}) among all the fungi chromosomes we have studied. Scatter-plots of a_{ex} *vs.* n_{ex} and a_{ex} vs. l_{ex}

show that already three parameters a_{ex}, n_{ex} and l_{ex} are sufficient for successful classification of 140 chromosomes to six fungal classes.

At this point, we use factor analysis of the system of 140 chromosomes that led us to the synthesis of the following successive logical structure:

- Dividing the system into sets of "elementary" components - all of the abovementioned chromosomal characteristics (n_{ex}, l_{ex}, a_{ex}, p_c, $l0_{ex}$, $n1_{ex}$, $l1_{ex}$)
- Analysis of the relationships of these components in species
- Revealing system-forming relations
- Description of the structure of the system (model) and its properties

As we can see, four main components are responsible for the whole system organization, and two of them can describe 93.9% of the whole variability of the system. The first component strongly divides all species into yeasts (*Saccharomycotina*) vs. *Pezizomycotina* and *Taphrinomycotyna*, and the second component demonstrates the difference between *Microsporidia* and *Basidiomycota*. Unfortunately, we can also see that the chromosomes of the species of the phylum Basidiomycota are split by the first component between two groups: they appear in the first group together with *Agaricomycotina* (CN) and in the second group together with *Ustilaginomycotina* (UM).

4.3. Comparison between and within different Kingdoms

All of the abovementioned chromosomal characteristics and Species-averaged statistical parameters were calculated for all 322 chromosomes. Among problems that we investigated in this study are: a) correlations between different species-averaged parameters of exon-intron structure; b) clustering chromosomes of a few organisms belonging to the same Kingdom (Protista, Plantae, and Animalia) by combinations of chromosome-averaged exonic characteristics; c) clustering of all 32 organisms by combinations of species-averaged characteristics of exons.

4.3.1. Some correlations among species-averaged statistical parameters

Regarding average protein lengths of intron-containing and intronless genes (net length of all exons), it seems that if there is only a small amount of intron-containing genes in a genome, such proteins are shorter in average than other proteins coded by intronless genes of the same genome. This property is especially strongly expressed for some species of fungi (EC, CG, and KL, and also exists for EG, DH, SP, and UM), and also for three

Protista species: *Leishmania braziliensis* (LB), *Hemiselmis_anderenii* (HA), and *Guillardia theta* (GT). From a scatter-plot of P_g vs. a fraction of $AL0_{ex}/AL1_{ex,}$ we can see three main groups of points in the plot: a group of genomes with low concentration of intron-containing genes (P_g<10%), a group of genomes with high concentration of intron-containing genes (P_g >70%), and an intermediate group. The first group may be mainly characterized by a striking prevalence of longer genes among intronless genes compared to intron-containing ones. We could deduce a rule that, in genomes with a low presence of intron-containing genes, such genes are coding shorter proteins; however, there is an exclusion of this empirical rule – LB has a fraction $AL0_{ex}/AL1_{ex}$ similar to genomes with rather high P_g. An empirical rule for the second group may be formulated as "there is a (linear) positive correlation between a proportion of intron-containing genes in a genome and a fraction $AL0_{ex}/AL1_{ex}$, while values of a fraction are lower than one". Unfortunately, we have exclusion to this rule as well – *Bigelowiella natans* (BN) has a surprisingly high value of the ratio $AL0_{ex}/AL1_{ex}$. Regarding the central group, we may say only that it has the most intriguing configuration that requires further studies.

4.3.2. Chromosome-averaged statistical parameters

Let us consider the average parameters l_{ex}, $n_{ex,}$ and a_{ex}. Scatter-plot of a_{ex} *vs.* l_{ex} for Protista. This figure illustrates the statement claimed above that the averages l_{ex} and a_{ex} turned out to be pretty similar for different chromosomes of the same species but, as a rule, rather distant for different species. Moreover, six separate groups of points may be observed.

We colored all points in four colors relating to four Protista Supergroups: Chromalveolata (PF, PK, PT, TA), Chromista (GT, HA), Protozoa (LB), and Rhizaria (BN). Analyzing the contents of the groups, one can suppose that the divisions follow their taxonomy. Indeed, we can clearly see six separate groups of chromosomes: BN chromosomes belonging to Rhizaria form the most left group, GT and HA chromosomes belonging to Chromista are located together, and Protozoa (LB) form the third cluster. Chromosomes belonging to Chromalveolata form three clusters, according to their phylum and class: *Apicomplexa plasmodium* (PF and PK) form one cluster, *Apicomplexa theileria* (TA) forms another cluster, and a single chromosome of Paramecium (PT) – the third cluster. These scatter plots show that the three parameters a_{ex}, n_{ex} and l_{ex} are sufficient for successful classification of 76 chromosomes to eight unicellular organisms.

The same conclusion regarding classification mirroring the phyla taxonomy can be made following an analysis of the matching chromosomal parameters for Animalia. Points related to averages l_{ex} and a_{ex} related to different chromosomes of the same species and was located pretty close to one another, while points related to chromosomes of different species are placed rather distant from one another. Striking exceptions are the points associated with chromosome 4 of *D. melanogaster* and chromosome 7 of *M. musculus* – these points form clusters of a single member clearly disjointed from other groups. All other points form three separate groups which may be observed. The two parameters l_{ex} and a_{ex} separately cluster 5 chromosomes of *D. melanogaster* (DM) in one group, 6 chromosomes of *C. elegans* (CE) in another group, and all 39 chromosomes of *C. familiaris* (CF), *H. sapiens* (HS), and *M. musculus* (MM) in the third group.

Let us repeat our observations relating to the phyla. We colored all points in three colors related to three animal phyla: Arthropoda, Chordata, and Nemata. A scatter plot of the a_{ex} *vs.* n_{ex}, and clearly shows three separate groups of chromosomes and two outliers. CF, HS, and MM chromosomes belonging to Chordata Mammalia form the most left group, CE chromosomes belonging to Nemata Caenorhabditis make the second left group, and the points belonging to DM (Arthropoda Insecta) appear in the right group. Two chromosomes – DM4 (the shortest chromosome of DM) and MM07 form two separate groups; each one having a single member. The CE chromosomes have the greatest exon density (n_{ex}) and the shortest exons (l_{ex}) among all animal chromosomes studied.

4.3.3. Clustering of genomes by species-averaged statistical parameters

After a relatively satisfying success of partial clustering based on only three chromosomal characteristics, our next objective was to cluster all 32 genomes. We took seven species-averaged exon parameters mentioned above: AN_{ex} (average number of exons in a gene per genome), AL_{ex} (average net length of all exons in a gene per genome), AA_{ex} (average exon length in a gene per genome), $AN1_{ex}$ (average number of exons in an intron-containing gene per genome), $AL0_{ex}$ = average (over a genome) length of an intronless gene, $AL1_{ex}$ (average net length of all exons in an intron-containing gene per genome), and P_g (proportion of intron-containing genes in a genome in percent). The expectation was that clustering would generally follow Kingdoms / Supergroups / Phyla classification; however, the results appeared to be rather poor (data not shown). Assuming that peculiar relations between a parameter P_g and other parameters may negatively influence clustering, we excluded this parameter from further consideration.

At this point, we tried to cluster genomes of 32 different organisms using six parameters, namely, AN_{ex}, AL_{ex}, AA_{ex}, $AN1_{ex}$, $AL1_{ex}$, and $AL0_{ex}$. At a first stage, we applied NJ clustering using standardized distances among the vectors [AN_{ex}, AL_{ex}, AA_{ex}, $AN1_{ex}$, $AL1_{ex}$, $AL0_{ex}$] and applying the program *Neighbor*. As one can see, some organisms of the same Kingdom / Supergroup are distributed compactly along the tree. Nevertheless, not all species belonging to the same class form a monophyletic cluster. Mice (MM), dogs (CF), and humans (HS) are located together, but flies (DM), which form a cluster together with Protista / Chromalveolata *Theileria annulata* (TA), appear too far away from other Animalia. Viridiplantae species are placed distantly; Protista are distributed along the tree in a strange manner. Such classification, even better than the classification produced by 7 parameters, cannot be considered as "sufficiently good". These discrepancies could be explained at least partially by the cross dependencies of all considered parameters. Therefore, the natural way to improve clustering is to replace these parameters by independent (orthogonal) parameters that could be obtained, for example, from results of a factor analysis of their correlation matrix as principal components. Four principal components are responsible for 99.4% of the whole system organization, and the two first can describe 86.2% of the whole variability of the system.

Results of *k*-means are rather similar to NJ results as well; however, there are some additional improvements in partitioning genomes among different clusters. In general, the main results showed a high consistency of partitioning, in spite of differences in clustering techniques. Careful examination reveals hierarchical partition of organisms. Interestingly, PAM clustering is not a hierarchical algorithm and should not necessarily produce any hierarchy. In our case of application of PAM clustering to four principal components obtained by Factor Analysis, a strictly hierarchical structure is produced. It may be interpreted as existence of an intrinsic hierarchical structure of PCA data. This may, in turn, serve as an additional evidence of evolutionary nature of exon-intron structure variance.

5. Conclusion

The origin of introns remains a mystery, and certain questions in molecular evolution are being investigated through *in silico* analysis of intron–exon structures in various organisms. To facilitate such studies, while taking advantage of the exploding amount of sequence data now available, we applied statistical analysis of the exon-intron structure to practically all completely sequenced eukaryotic genomes in order to reveal general and genome-specific features of eukaryotic genes. We went through all of the protein-coding genes in each chromosome separately and calculated the portion of intron-containing genes and average values of the net length of all the exons in a gene, the number of the exons, and the average length of an exon. The purpose of this research has been to determine the most appropriate approach to classify eukaryotic chromosomes, according to these simple exon-intron statistics.

Our results show both general and genome-specific features of the exon-intron organization of eukaryotic genes. The most general feature found in all genomes is the positive correlation between the number of introns in a gene and the corresponding protein's length (or, equivalently, the net length of all the exons of the gene). In addition, in all the genomes we have studied, the average exon length in a gene decreases with the number of those exons. But while these laws of exon-intron statistics are quite general, the correlation parameters are genome-specific. Moreover, they are specific to genomes rather than to individual chromosomes. Indeed, in the parameter space of average chromosome properties and linear regression parameters (between exon numbers and protein lengths), all chromosomes from the same genome form obvious clusters.

Clearly, the exon-intron structures of eukaryotic genes have many important parameters that we did not consider in this work; we have left them for the future research. The main goals of this research are to draw attention to the statistical properties of exon size distributions, and to visualize both the general laws of exon-intron organizations of genes and the genome-specific features.

6. References

Ahmavaara, Y. and Markkanen, T. (1958). *The Unified Factor Model.* The Finnish Foundation for Alcohol Studies, Helsinki, Finland.

Archibald, J.M., O'Kelly, C.J., Doolittle, W.F. (2002). The chaperonin genes of jakobid and jakobid-like flagellates: implications for eukaryotic evolution. *Mol. Biol. Evol.* **19**: 422-431.

Atambaeva, S.A., Khailenko, V.A., and Ivashchenko, A.T. (2008). Intron and exon length variation in arabidopsis, rice, nematode, and human, *Mol. Biol.* **42**: 312-320.

Bartholomew, D.J. (1987). *Latent variable models and factor analysis*. Oxford Univ. Press, New York.

Cavalier-Smith, T. (1985). Selfish DNA and the origin of introns. *Nature* **315**: 283-284.

Cho, G. and Doolittle, R.F. (1997). Intron distribution in ancient paralogs supports random insertion and not random loss. *J. Mol. Evol.* **44** (6): 573-584.

Collins, F.S., Lander, E.S., Rogers, J., Waterston, R.H. (2004). International Human Genome Sequencing Consortium: Finishing the euchromatic sequence of the human genome. *Nature* **431**: 931-945.

Deutsch, M. and Long, M. (1999). Intron-exon structures of eukaryotic model organisms, *Nucl. Acids Res.* **27**: 3219-3228.

Donoghue, D.J. and Sharp, P.A. (1977). An improved bacteriophage lambda vector: Construction of model recombinants coding for kanamycin resistance. *Gene* **1** (3): 209-227.

Drysdale, R.A., Crosby, M.A and The FlyBase Consortium (2005). FlyBase: genes and gene models. *Nucleic Acids Res* **33**: D390-D395.

Engelgardt, V.A. (1970). Integratism – way from simple to complex in the life research. *Dokl. Biol. Sci*. **6**: 799-822 (1970).

Forni, M. and Lippi, M. (2000). *The Generalized Dynamic Factor Model: Representation Theory*. Mimeo, Universita di Modena.

Forsdyke, D.R. (1981). Are introns in-series error-detecting sequences? *J. Theor. Biol.* **93**: 861-866.

Forsdyke, D.R. (1995). A stem-loop kissing model for the initiation of recombination and the origin of introns. *Mol. Biol. Evol.* **12**: 949-958.

Gelinas, R.E. and Roberts, R.J. (1977). One predominant 5'-undecanucleotide in adenovirus 2 late messenger RNAs. *Cell* **11** (3):533-544.

Gilbert, W. (1987). The exon theory of genes. In: *Symp. Quant. Biol.*, Cold Spring Harbor, **52**: 901-905.

Gudlaugsdottir, S., Boswell, D.R. Wood, G.R. and Ma, J. (2007). Exon size distribution and the origin of introns. *Genetica* **131** (3): 299-306

Haas, B.J., Wortman, J.R., Ronning, C.M., *et al.* (2005). Complete reannotation of the Arabidopsis genome: methods, tools, protocols and the final release. *BMC Biology* **3**: 7.

Hawkins, J.D. (1988). A survey on intron and exon lengths. *Nucl. Acids Res.* **16**, 9893-9908.

Ivashchenko, A.T. and Atambaeva, Sh.A. (2004). Variation in lengths of introns and exons in genes of the *Arabidopsis thaliana* nuclear genome, Russian Journal of Genetics **40** (10): 1179-1181.

Ivashchenko, A.T., Khailenko V.A. and Atambayeva, S.A. (2009). Variation of the lengths of exons and introns in Human Genome genes, *Russian Journal of Genetics,* **45** (1): 16-22.

Ivashchenko, A.T., Tauasarova, M.I., and Atambayeva, S.A. (2009). Exon–intron structure of genes in complete fungal genomes. *Mol. Biol.* **43** (1): 24-31.

Järveläinen, V.P. (1971). Vähäsen faktorianalysistä, *Silva Fenn*. **5** (3): 281-290

Kaplunovsky, A., Khailenko, V.A., Bolshoy, A., Atambayeva, S.A., and Ivashchenko, A.T. (2009). Statistics of exon lengths in animals, plants, fungi, and protists. *International Journal of Biological and Life Sciences* **1** (3): 139-144.

Kaplunovsky, A., Zabrodsky, D., Volkovich, Z., Ivashchenko, A.T., and Bolshoy, A. (2010). Statistics of exon lengths in fungi. *The Open Bioinformatics Journal* **4**: 31-40.

Kaplunovsky, A., Ivashchenko, A.T., and Bolshoy, A. (2011). Statistical analysis of exon lengths in various eukaryotes. *Open Access Bioinformatics* **3:** 1-15.

Kliman, M. and Bernal, C.A. (2005). Unusual usage of AGG and TTG codons in humans and their viruses, *Gene* **352**: 92-99.

Kriventseva, E.V. and Gelfand, M.S. (1999). Statistical analysis of the exon-intron structure of higher and lower eukaryote genes, *J. Biomol. Struct. Dyn.* **17**: 281-288.

Kupfer, D.M., Drabenstot, S.D., Buchanan, K.L., *et al.* (2004). Introns and splicing elements of five diverse fungi. *Eukaryot Cell* **3**: 1088-1100.

Lewin, B. (2000). *Genes*. Oxford Press, Oxford, UK.

Liebovitch, L.S., Tao, Y., Todorov, A.T., and Levine, L. (1996). Is there an error-correcting code in the base sequence of DNA? *Biophys. J.* **71**: 1539-1544.

Loftus, B.J., Fung, E., Roncaglia, P., *et al.* (2005). The genome of the basidiomycetous yeast and human pathogen Cryptococcus neoformans. *Science* **307**: 1321-1324.

Logsdon, J.M. and Palmer, J.D. (1994). Origin of introns – early or late? *Nature* **369**: 526-528.

Logsdon, J.M.J., Stoltzfus, A., and Doolittle, W.F. (1998). Molecular evolution: recent cases of spliceosomal intron gain? *Curr. Biol.* **8**: R560-R563.

Martinez, D., Berka, R.M., Henrissat, B., *et al.* (2008). Genome sequencing and analysis of the biomass-degrading fungus Trichoderma reesei (syn. Hypocrea jecorina). *Nat. Biotechnol.* **26** (5): 553-560.

Naora, H. and Deacon, N.J. (1982). Relationship between the total size of exons and introns in protein-coding genes of higher eukaryotes. *Proc. Natl. Acad. Sci. USA* **79** (20): 6196-6200.

Odintsova, M.S. and Yurina, N.P. (2002). The Mitochondrial Genome of Protists, Russian Journal of Genetics **38** (6): 642-655.

Raible, F., Tessmar-Raible, K., Osoegawa, K**.** *et al.* (2005). Vertebrate-type intron-rich genes in the Murine Annelid *Platynereis dumerilii*. *Science* **310** (5752): 1325-1326.

Ren, X.Y., Vorst, O., Fiers, M.W.E.J., Stiekema ,W.J., Nap, J.P. (2006). In plants, highly expressed genes are the least compact. *Trends Genet.* **22** (10): 528-532.

Reuchlin, M. (2003). Contribution à l'histoire des méthodes statistiques employées en psychologie: les plans d'expérience et l'analyse de variances: Ronald Aymler Fisher (1890-1962). *Psychol. Histoire* **4**: 31-60.

Roy, S.W. (2004). The origin of recent introns: transposons? *Genome Biol*. **5:** 251.

Roy, S.W. and Penny, D. (2007). Intron length distributions and gene prediction. *Nucleic Acids Res.* **35:** 4737-4742.

Ryabov, Y. and Gribskov, M. (2008). Spontaneous symmetry breaking in genome evolution, *Nucleic Acids Res*. **36** (8): 2756-2763.

Saitou N. and Nei M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4** (4): 406-425.

Sakharkar, M.K., Chow, V.T., and Kangueane, P. (2004). Distributions of exons and introns in the human genome. *In Silico Biol.* **4** (4): 387-393.

Scherrer, K., Spohr, G., Granboulan, N., *et al.* (1970). Nuclear and cytoplasmic messenger-like RNA and their relation to the active messenger RNA in polyribosomes of HeLa cells. *Cold Spring Harb. Symp. Quant. Biol*. **35**: 539-554.

Schwarz, E.M., Antoshechkin, I., Bastiani, C., *et al.* (2006). WormBase: better software, richer content. *Nucleic Acids Res*. (**34** Database): D475-478.

Singer, M. and Berg, P. (1991). *Genes and Genomes: A Changing Perspective*, University Science Books, Mill Valley, Ca, USA.

Turmel, M., Otis C., and Lemieux C. (1999). The complete chloroplast DNA sequence of the green alga *Nephroselmis olivacea*: Insights into the architecture of ancestral chloroplast genomes, *Proc. Natl. Acad. Sci. USA* **96** (18): 10248-10253.

Umesono, K., Inokuchi, H., Shiki Y. *et al*. (1988). Strustural organization of *Marchantia polymorpha* chloroplast genome. *J. Mol. Biol.* **203**: 299-331.

Venter, J.C. Adams, M.D., Myers, E.W. *et al*. (2001). The sequence of the human genome, *Science* **291** (5507): 1304-1351.

Verhoog, H. (1993). *Zit er Toekomst in ons DNA*, Werkgroep Genenmanipulatie en Oordeelsvormung, Louis Bolk Instituut, Driebergen, Holland (in Dutch). German translation of part 2 (1994) in: *Genmanipulation an Pflanze, Tier und Mensch - Grundlagen zur Urteilsbildung*: 11-22. Verlag Freies Geistesleben, Stuttgart, Germany. English translation: http://www.ifgene.org/verhoog.htm

Wendel, J.F., Cronn, R.C., Alvarez, I., Liu, B., Small, R.L., and Senchina, D.S. (2002). Intron size and genome size in plants. *Mol. Biol. Evol.* **19** (12): 2346-2352.

Yap Yee Leng, Zhang Xue Wu, and Danchin, A. (2003). Relationship of SARS-CoV to other pathogenic RNA viruses explored by tetranucleotide usage profiling, *BMC Bioinformatics* **4**: 43.

Zhu, L.C., Zhang, Y., Zhang, W., Yang, S.H., Chen, J.Q., and Tian, D.C. (2009). Patterns of exon-intron architecture variation of genes in eukaryotic genomes. *BMC Genomics* **10**:12.

7. List of the author's selected publications in factor analysis

Bekhtereva, N.P., Bundzen, P.V., A.S.Kaplunovsky, and Matveev, Y.K. (1971). Functional organization of activity of cerebral neuronal assemblies in humans during short-term verbal memory. *Sechenov Physiol. Journ. USSR* **57** (12): 1605-1621.

Bundzen, P.V., Vasilevsky, N.N., Kaplunovsky, A.S., and Shabaev, V.V. (1971). Factor analysis in studies of functional organization of dynamic characteristics of the brain electrical activity. *Sechenov Physiol. Journ. USSR* **57** (7): 969-973.

Klimenko, V.M., Kaplunovsky, A.S., and Neroslavsky, I.A. (1972) Automatical classification of multiparametric experimental data, *Sechenov Physiol. Journ. USSR* **58** (4): 599-602.

Klimenko, V.M. and Kaplunovsky, A.S. (1972). Statistical study of impulse activity of neurons in the various hypothalamic areas, *Sechenov Physiol. Journ. USSR* **58** (10): 1484-1493.

Kaplunovsky, A.S. and Uriash, V.V. (1972). Factor analysis of the brain electrogram of cat in various sleep stadies. In: *Structure and Functions of Cerebral Brain,* Leningrad, 31-34.

Bekhtereva, N., Bundzen, P., Matveev, J., and Kaplunovski, A. (1973). Die Neuronenpopulationen des Gehirns beim Verbalen Kurzzeitgeduchtnis, *Moderne Medizin*, **4** (3): 179-186.

Bundzen, P.V., Gogolitsin, Yu.L., David, E., Kaplunovsky, A.S., and Perepelkin, P.D.(1973). Structural-systemic approach to the analysis of functional reorganization of neuronal pools. *Sechenov Physiol. Journ. USSR* **59** (12): 1803-1810.

Bundzen P.V., Gogolitsin, Y.L. and Kaplunovsky A.S. (1973). The computer analysis of multiunit activity of neuronal populations of the human brain, In: *Biological Diagnosis of Brain Disorders* (Proc. 5-th Intern. Conf. held at the New York Acad. Med., Oct.2-3, 1972.), Ed.: S. Bogoch, Spectrum Publ., Flushing, N.Y.: 275-280.

Bundzen P.V., Kaplunovsky A.S., and Gogolitsin, Y.L. (1973). Structural-functional approach to the analysis of bioelectrical processes in human brain. In: *Problems of Neurocybernetics*, Kiev, 34-70.

Bundzen, P.V., Kaplunovsky, A.S. and Perepelkin, P.D. (1973). Complex statistical analysis of multi-unit activity of human brain neuron populations during the mental activity. In: *Bionika-1973* (Mater. 4-th Vsesoyuzn. Conf. on Bionics), Moscow, vol.**3**:7-22.

Kaplunovsky A.S. and Bogoslovsky, M.M. (1973). The use of factor analysis for encephalographic characteristics of sleep, *Sechenov Physiol. Journ. USSR*, **59** (8): 1291-1292.

Bundzen, P.V. and Kaplunovsky, A.S. (1974). The Self-organization principles of the structural-systemic brain systems and memory mechanisms. In: *Problems of Physiology and Pathology of Higher Nervous Activity.* (Ed. by N.P. Bekhtereva). Medicina, Leningrad, 80-108.

Kaplunovsky A.S. (1974). To the problem of neuroholographic basis of the coding of verbal signals, In: *Neurophysiological Mechanisms of Human Mental Activity* (Proc. Intl. Symp.), Medicina, Leningrad: 212-214.

Kaplunovsky A.S. (1974). Principles of neuroholographic coding of information by the human memory. In: *Memory in Normal and Pathological Reactions of Organism*, Leningrad, 161-168.

Kaplunovsky A.S. and Perepelkin, P.D. (1974). Use of factor analysis methods for study multi-unit activity of human brain neuronal populations, In: *Use of Mathematical Methods in Medicine*, Moscow: 83-87.

Kaplunovsky A.S. and Perepelkin, P.D. (1974). Principal aspects of neuronal code of verbal signals from the point of view of bioholography. In: *Biological and Medical Cybernetics*, Moscow-Leningrad., vol.**3**: 69-73

Bundzen, P.V., Gogolitsin, Y.L., Kaplunovsky, A.S., and Malyshev, V.N. (1975). Systemic approach to the analysis of human brain neuronal populations during mental activity. *Human Physiology* **1** (1): 45-60.

Bundzen, P.V, Kaplunovsky, A.S., Klimenko,V.M., and Korneva, E.A. (1975). Methodological aspects and principles of the factor analysis use in neurophysiology, In: *Methodological Problems of Theoretical Medicine* (Ed.: N.P. Bekhtereva). Medicina, Leningrad: 25-39.

Bekhtereva, N.P., Bundzen, P.V.,.Kaplunovsky, A.S., and Perepelkin, P.D. (1976). On the neurophysiological coding of mental phenomena in man. In: *Memory of Mechanisms of Normal and Pathological Reactions*. (Ed. by N.P. Bechtereva). Medicina, Leningrad: 9-27.

Kaplunovsky A.S. (2005) Factor analysis in environmental studies, *HAIT Journal of Science and Engineering* B **2** (1-2): 54-94.

Kaplunovsky A.S. (2007) Why using factor analysis? Submitted to *HAIT Journal of Science and Engineering*. http://www.magniel.com/fa/kaplunovsky.html

שיטות לגילוי גורמים אבולוציונים

מבוססים על ניתוח מבנה אקסון-אינטרון של גנים

אלכסנדר קפלונובסקי

תקציר

האורך של אינטרון נע בין עשרה נוקליאוטידים ועד עשרות אלפי נוקליאוטידים. הראנו שקיימת קורלציה בין שונות האורכים של אינטרונים ואקסונים בגנומים עם חלק מתפקידייהם וייתכן וקורלציה זו נגרמה מגורמים אבולוציוניים.

מטרת מחקר זה היא לקבוע את הגישה הנאותה ביותר על מנת לסווג כרומוזומים איקריוטיים לפי סטטיסטיקה של אינטרון-אקסון. מבנה האקסונים/אינטרונים בגנים איקריוטיים הנו מגוון. וההתפתחות של מבנים אלו מעלה מספר שאלות מאתגרות. בניסיון הראשוני להתייחס לשאלות אלו ביצענו ניתוח של מבנים של אקסון-אינטרון. סרקנו את כל הגנים אשר מקודדים חלבון בכל כרומוזום בנפרד וחישבנו את היחס של הגנים אשר מכילים אינטרונים לכמות כוללת של הגנים, את האורך הממוצע של כלל האקסונים בגן ביחד, מספר האקסונים הממוצע והאורך הממוצע של אקסון. השוונו את הממוצעים הכרומוזומליים אל מול הממוצעים לגן ובעזרת כך פיתחנו מתודולוגיה של יצירת אשכולות לפי מאפייניי מבנה אקסון / אינטרון.

טכניקת הקבצה זו, יצרה הבדלה בין מינים שונים, מקובצים לפי מאפיינים איקריוטיים. מסקנתנו הייתה שהגישה הנכונה היא להתבסס על המרחקים בין ארבעת המרכיבים העיקריים (**Principal Components**) אשר הושגו בניתוח גורמים (**Factor Analysis**), כאשר על תוצאות אלו הפעלנו אלגוריתם שיוך אשכולות שכנים (**Neighbor Joining**).

שיטות לגילוי גורמים אבולוציוניים

מבוססים על ניתוח מבנה אקסון-אינטרון של גנים

מאת: אלכסנדר קפלונובסקי

בהדרכת: פרופ' אלכסנדר בולשוי

פרופ' אדוארד יעקובוב

חיבור לשם קבלת התואר "דוקטור לפילוסופיה"

אוניברסיטת חיפה

הפקולטה למדעי הטבע

החוג לביולוגיה אבולוציונית וסביבתית

פברואר, 2012

שיטות לגילוי גורמים אבולוציוניים

מבוססים על ניתוח מבנה אקסון-אינטרון של גנים

אלכסנדר קפלונובסקי

חיבור לשם קבלת התואר "דוקטור לפילוסופיה"

אוניברסיטת חיפה

הפקולטה למדעי הטבע

החוג לביולוגיה אבולוציונית וסביבתית

פברואר, 2012