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

1875-0362/10

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:

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

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

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

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 di-

rection of the covariance matrix $C = \left(M_p - we^T\right)\left(M_p - we^T\right)^T$, where $e = (1,1,...,1)^T$ 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.

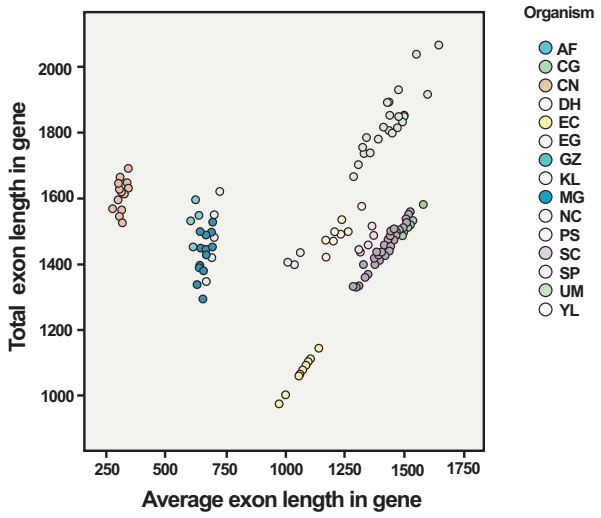

Fig. (1). Scatter-plot of the average exon length per gene a_{ex} (*x*-axis) vs. the total exon length l_{ex} (*y*-axis) for all 140 processed chromosomes of 15 fungi species.

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

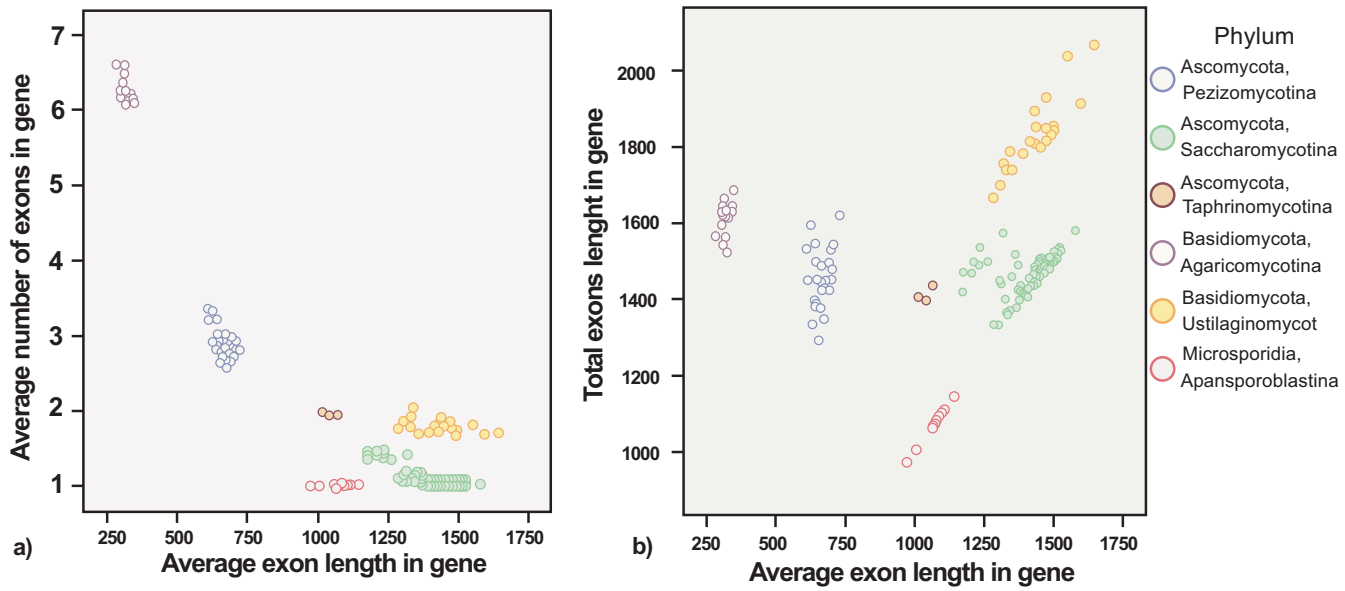

Fig. (2). Scatter-plot for all 140 processed chromosomes of six fungi phyla of the average exon length per gene a_{ex} (x-axis). **a**) vs. the average number of exons per gene n_{ex} (y-axis); **b**) vs. the average net exon length per gene l_{ex} (y-axis).

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

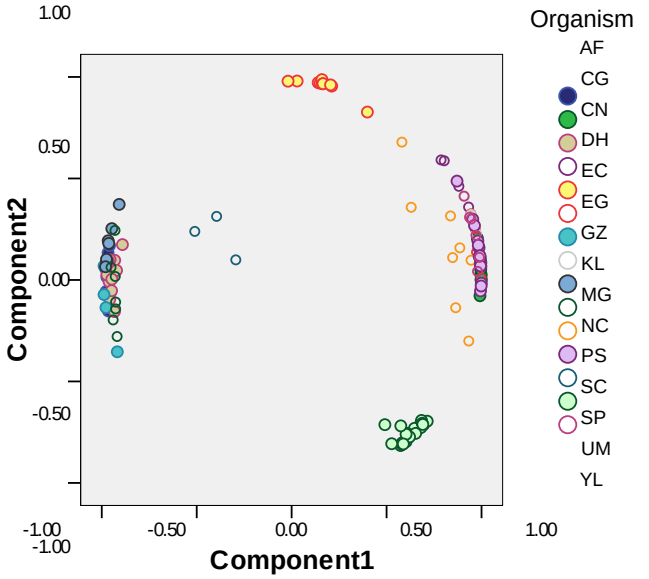

Fig. (3). Factor analysis of 140 processed chromosomes of 15 fungi species by seven parameters (n_{ex}, l_{ex}, a_{ex}, p_c, $l0_{ex}$, $n1_{ex}$, $l1_{ex}$) colored in reference to species.

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).

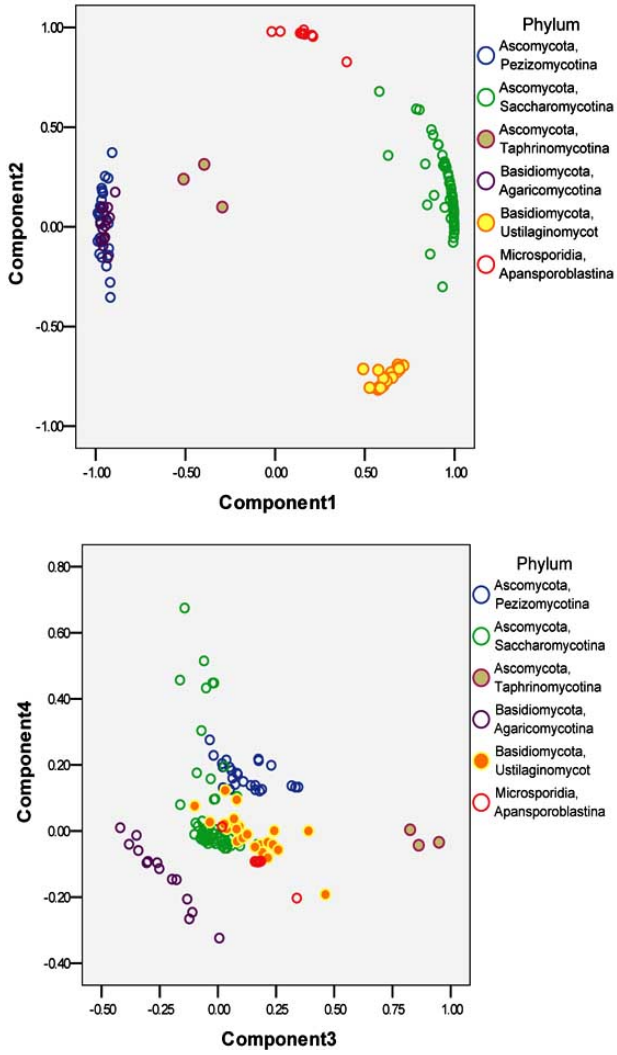

Fig. (4). Factor analysis of 140 processed chromosomes of 15 fungi species by seven parameters (n_{ex}, l_{ex}, a_{ex}, p_c, $l0_{ex}$, $n1_{ex}$, $l1_{ex}$) colored in reference to phylum.

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.

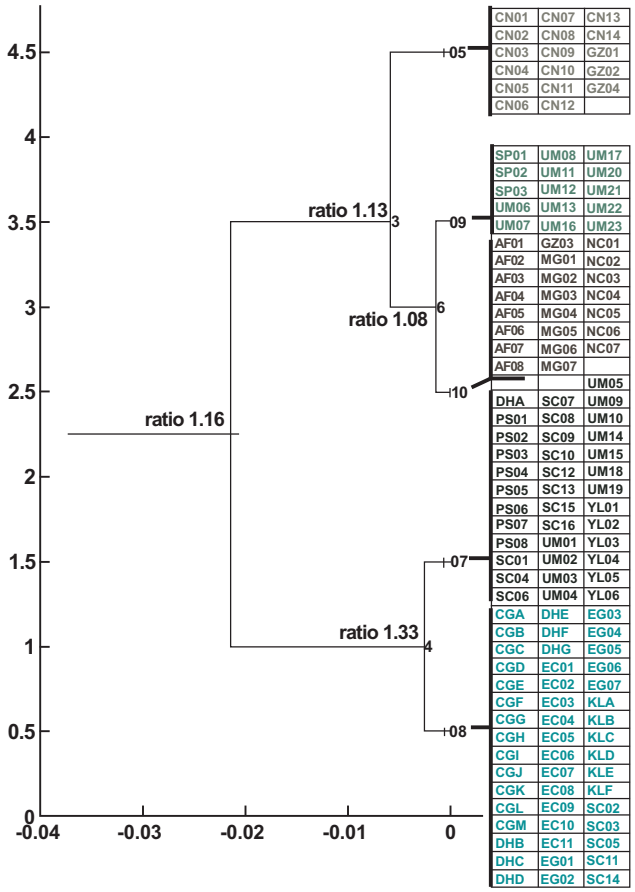

Fig. (5). A dendrogram of clusters obtained by the PDDP method based on scaled distances among the vectors (n_{ex}, l_{ex}, a_{ex}, p_c, $l0_{ex}$, $n1_{ex}$, $l1_{ex}$) presenting the chromosomes.

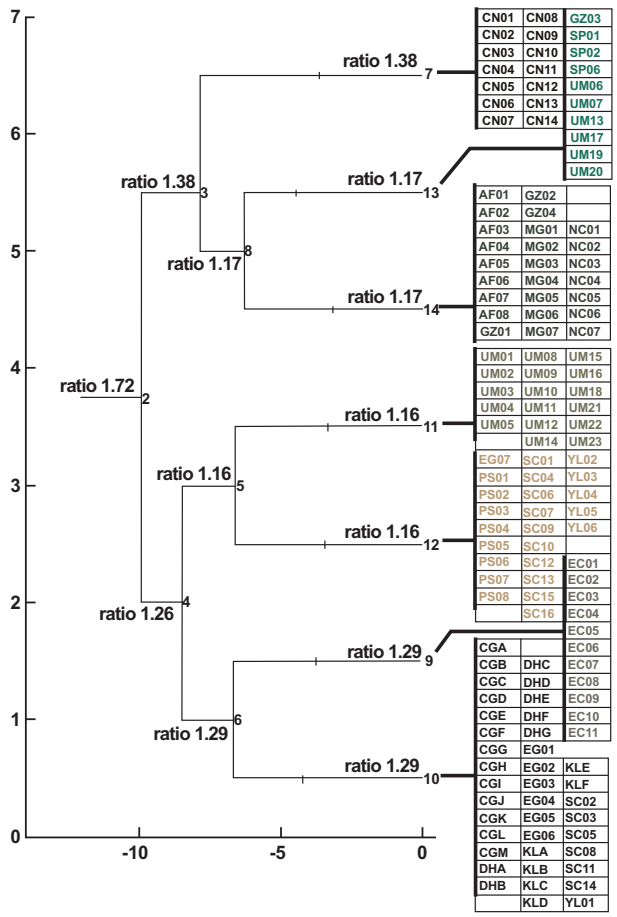

Fig. (6). Dendrogram of clusters obtained by the PDDP method based on standardized distances among the vectors (n_{ex}, l_{ex}, a_{ex}, p_c, $l0_{ex}$, $n1_{ex}$, $l1_{ex}$) presenting the chromosomes.

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.

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 ex-

ons 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**

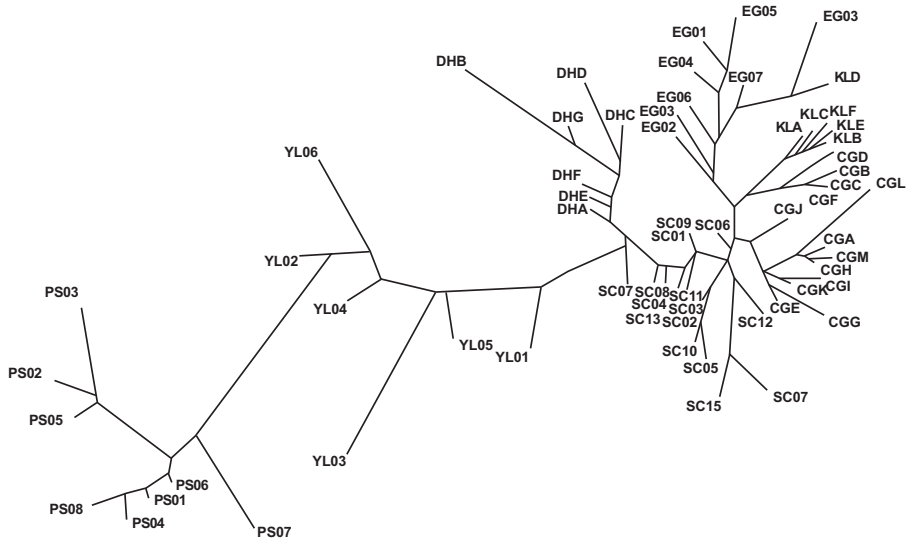

Fig. (7). Dendrogram of the 63-processed chromosomes of seven yeast species based on scaled distances among parameters n_{ex}, l_{ex}, a_{ex}, *an*, *al*, *nl*, *a*, *b*, and R^2 (*an*, *al*, and *nl* are correlation coefficients; *a*, *b*, and R^2 are parameters of the linear regression $n_{ex}=a+b\cdot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