Journal of Biological Physics **30:** 305–312, 2004.

Autosomal Similarity Revealed by Eukaryotic Genomic Comparison

ZHEN QI[1], YAN CUI[2], WEIWU FANG[3], LUNJIANG LING[1,*] and RUNSHENG CHEN[1,*]

[1]*Institute of Biophysics, Chinese Academy of Sciences, Beijing 100101, PR China;* [2]*Center of Genomics and Bioinformatics, University of Tennessee, Memphis, TN 38163, USA;* [3]*Institute of Mathematics, Chinese Academy of Sciences, Beijing 100080, PR China;*

[*]*Author for correspondence, e-mail: ling@sun5.ibp.ac.cn, crs@sun5.ibp.ac.cn*

Abstract. To describe eukaryotic autosomes quantitatively and determine differences between them in terms of amino acid sequences of genes, functional classification of proteins, and complete DNA sequences, we applied two theoretical methods, the Proteome-vector method and the function of degree of disagreement (FDOD) method, that are based on function and sequence similarity respectively, to autosomes from nine eukaryotes. No matter what aspect of the autosome is considered, the autosomal differences within each organism were less than that between species. Our results show that eukaryotic autosomes resemble each other within a species while those from different organisms differ. We propose a hypothesis (named intra-species autosomal random shuffling) as an explanation for our results and suggest that lateral gene transfer (LGT) did not occur frequently during the evolution of eukarya.

Key words: autosomal similarity, Proteome-vector, FDOD, shuffling, genomic structure

1. Introduction

As a number of genomes have been completely sequenced, it is possible and necessary to compare whole genomes instead of small fragments to resolve biological problems. Whole genomic comparison takes advantage of comprehensive information and has been heralded as the next logical step toward solving genomic puzzles, such as determining coding regions [1, 2], assigning protein functions [3], discovering regulatory signals [4], comparing organization of vertebrate genomes [5, 6] and deducing the history of evolution [7–10]. Traditionally, chromosomes were characterized qualitatively by karyotypes and banding patterns. Later, small segments of chromosomes were exploited to elucidate relationships between species. Now, in the genomic era large data sets, e.g., complete DNA sequences, can be used to describe chromosomes quantitatively and to do comparison between them. We applied two theoretical methods, the Proteome-vector method [11] developed by us recently, and the FDOD method [12, 13] to characterize and compare eukaryotic autosomes.

In the course of speciation and evolution of eukarya, the contents and functional properties of protein-encoding genes (simply referred to as "genes" hereafter) in an autosome may be changed substantially. To compare this feature between eukaryotic autosomes, we utilize the Proteome-vector approach that exploits information of coding areas to create a multi-dimensional vector whose components depict the relative contents of genes with different functions in an autosome.

As DNA sequences are the result of species evolution, the closer the evolutionary relationship between species, the more similar their genomic sequences should be. Utilizing complete or parts of genomic sequences, the FDOD method can calculate species-specific complete information set (CIS) as a representation of the sequence property of an autosome. Thereafter, a measure of sequence discrepancy between autosomes can be constructed. Different from the sequence alignment approach, the FDOD method is based on information theory and can exploit complete sequences of autosomes.

Using these two methods, we compared autosomes from nine eukaryotes in terms of functional classification of genes, amino acid sequences, and complete DNA sequences, and then both intra- and inter-species autosomal differences were calculated.

The software we developed and the supplementary materials are available upon request.

2. Materials and Methods

A total of 103 autosomes from nine eukaryotes (including one plant, two fungi, one protozoa, one nematode, one arthropod, and three mammals) were used as materials (Table I). Functional classification of a gene was determined by aligning its sequence to the COG (clusters of orthologous groups of proteins) database [14], that is defined by comparing protein sequences encoded in 43 complete genomes. These data were downloaded from relevant anonymous ftp servers (ftp://130.14.22.5/ and ftp://ftp.ensembl.org/).

Firstly, functional classification of genes was compared between autosomes by the Proteome-vector approach. A 17-dimension vector for functional classification of a gene was defined by aligning its amino acid sequence to the COG database using the FASTA program [15]. The COG database contains 17 functional subclasses that cover almost all known functions needed by an organism to survive. For each subclass, the M sequences with most matches to the query gene were extracted. After these sequences were sorted according to their match degree, the top M sequences were used to construct a 17-dimension vector $V_i = \{V_{i,1}, V_{i,2}, \ldots, V_{i,17}\}$ to depict COG-class attribute of the query gene. We set parameter M at 17 to count in possible contribution of every subclass, meanwhile to avoid introducing additional matches from other subclasses if the query gene was already included in a subclass. Then, the average vector that counts in the contribution of all genes in an autosome was calculated. After normalization of average vectors, we calculated

Table I. Autosomes from nine eukaryotes

ID	Species	Autosomes	Accession number	Revision date
Atha	*Arabidopsis Thaliana*	1–5	NC_003070.2, NC_003071.1, NC_003074.2, NC_003075.1, NC_003076.2 (From GenBank)	JAN-2002
Cele	*Caenorhabditis Elegan*	1–5	NC_003279.2, NC_003280.2, NC_003281.2, NC_003282.2, NC_003283.2, NC_003284.2, (From GenBank)	DEC-2001
Dmel	*Drosophila Melanogaster*	2,3	AE002566, AE002575, AE002584, AE002593, AE002602, AE002620, AE002629, AE002638, AE002647, AE002681, AE002690, AE002699, AE002708, AE002725, AE002769, AE002778, AE002787, AE002796, AE002804 (From GenBank)	OCT-2000
Ecun	*Encephalitozoon Cuniculi*	1–11	NC_003242.2, NC_003229.1, NC_003230.1, NC_003231.1, NC_003232.1, NC_003233.1, NC_003234.1, NC_003235.1, NC_003238.2, NC_003236.1, NC_003237.1 (From GenBank)	MAR-2002
Scer	*Saccharomyces Cerevisiae*	1–16	NC_001133.1, NC_001134.2, NC_001135.2, NC_001136.2, NC_001137.2, NC_001138.2, NC_001139.2, NC_001140.2, NC_001141.1, NC_001142.2, NC_001143.2, NC_001144.2, NC_001145.1, NC_001146.1, NC_001147.2, NC_001148.1, (From GenBank)	JUN-2002
Spom	*Schizosaccharomyces Pombe*	1,2,3	NC_003424.1, NC_003423.1, NC_003421.1 (From GenBank)	MAR-2002
Rnor	*Rattus norvegicus*	1–20	Version 9.1.1 (From Ensemble)	NOV-2002
Mmus	*Mus musculus*	1–19	Version 9.3a.1 (From Ensemble)	DEC-2002
Hsap	*Homo sapiens*	1–22	Version 8.30.1 (From Ensemble)	SEP-2002

Euclidian distances between them, and a pair-wise similarity matrix of autosomes was constructed.

We then applied the FDOD method to compare autosomes in terms of amino acid sequences. Based on Shannon's definition of information, entropy and degree of disagreement, the FDOD method calculates a species-specific CIS, and a distance matrix with elements determined by the discrepancies of CIS was created.

Finally, complete DNA sequences of autosomes from nine eukaryotes were subjected to the FDOD software to calculate autosomal difference in sequence composition.

Considering annotations of proteins for genomes of human, mouse and rat are at present incomplete and imprecise [16], we did not include them into the comparison of the functional classification of genes. For the comparison of amino acid sequences, the rat was omitted.

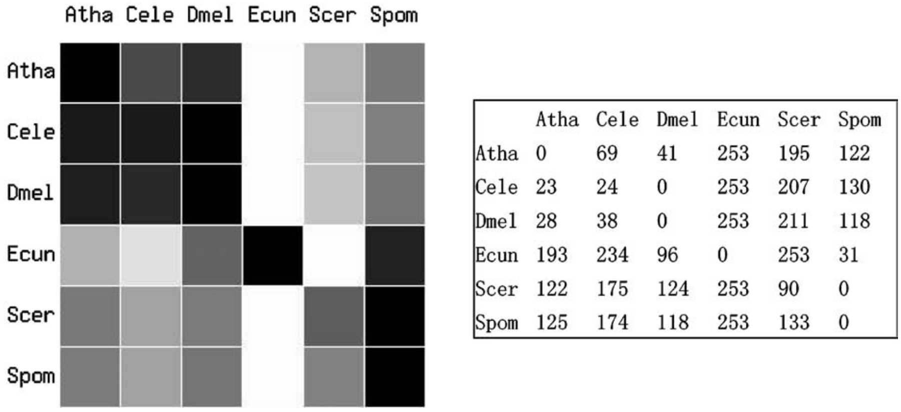
3. Results

The inter-species autosomal distances were calculated by averaging all differences between autosomes from two organisms. Similar calculations were performed for the intra-species autosomal distances except that all autosomes came from the same organism. Then, distance matrices were linearly discretized to integer values between 0 and 255 to draw gray level images. To make the minimum of each row stand out more clearly, we set its value to zero.

When functional classification of genes was compared, intra-species autosomal distances were less than inter-species distances, and for most rows of the distance matrix the minimum was located in the diagonal except that for roundworm and baker's yeast (Figure 1). Detailed analyses disclosed that several autosomes in these two organisms have very distinct gene content from the others that may be the cause of such inconsistency (data not shown).

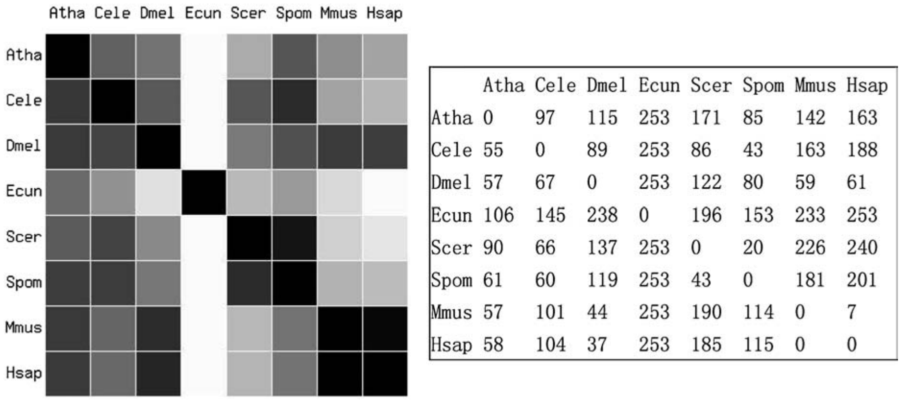
In the comparison of amino acid sequences, autosomal distances within each species were also less than between species (Figure 2). Though the distance between human and mouse was a little greater than the diagonal element of the related row before discretization, this element was set to zero due to the limitation of gray levels. Two yeasts (baker's yeast and fission yeast) showed very small difference and could be clustered together. Two mammals (mouse and human) also form one group for the same reason. These results are consistent with general knowledge of taxonomy.

Figure 3 shows the autosomal distance matrix when complete DNA sequences of autosomes were compared. Inter-species autosomal distances were greater than autosomal distances within each species, and except for baker's yeast, the minimal element of each row was located in the diagonal. The two yeasts (baker's yeast and fission yeast) have small differences in gray value and could be clustered together. Similarly, three mammals (mouse, rat and human) also form a group.



	Atha	Cele	Dmel	Ecun	Scer	Spom
Atha	0	69	41	253	195	122
Cele	23	24	0	253	207	130
Dmel	28	38	0	253	211	118
Ecun	193	234	96	0	253	31
Scer	122	175	124	253	90	0
Spom	125	174	118	253	133	0

Figure 1. Comparison of functional classification of genes by the Proteome-vector method. Autosomes were compared using 17-dimension proteome-vector in terms of functional classifications of proteins. Parameter M was set equal to 17. Euclidian distances were used to calculate differences between vectors.

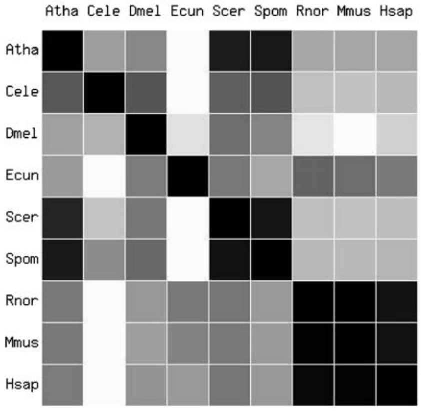


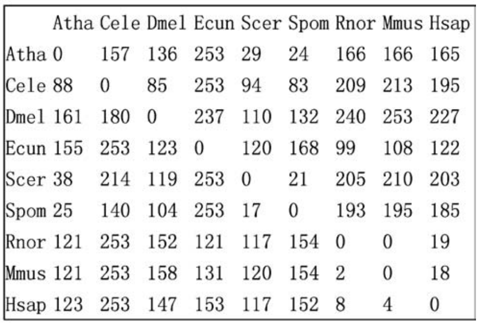
	Atha	Cele	Dmel	Ecun	Scer	Spom	Mmus	Hsap
Atha	0	97	115	253	171	85	142	163
Cele	55	0	89	253	86	43	163	188
Dmel	57	67	0	253	122	80	59	61
Ecun	106	145	238	0	196	153	233	253
Scer	90	66	137	253	0	20	226	240
Spom	61	60	119	253	43	0	181	201
Mmus	57	101	44	253	190	114	0	7
Hsap	58	104	37	253	185	115	0	0

Figure 2. Comparison of amino acid sequences using the FDOD method. The FDOD method was applied to compare autosomes in terms of amino acid sequences. Parameter K was set to two. Euclidian distances were used to calculate differences of amino acid sequences between autosomes. Results for different values of parameter K were similar.

4. Discussion

We applied two theoretical methods to quantitatively characterize and compare eukaryotic autosomes. The Proteome-vector approach provides unique definition for functional classification of genes and makes it easy to measure autosomal relationships in terms of coding area. The FDOD method is able to exploit whole genome information and avoid bias that may be associated with particular segments. Moreover, it circumambulates sequence alignment and related problems [17, 18].





	Atha	Cele	Dmel	Ecun	Scer	Spom	Rnor	Mmus	Hsap
Atha	0	157	136	253	29	24	166	166	165
Cele	88	0	85	253	94	83	209	213	195
Dmel	161	180	0	237	110	132	240	253	227
Ecun	155	253	123	0	120	168	99	108	122
Scer	38	214	119	253	0	21	205	210	203
Spom	25	140	104	253	17	0	193	195	185
Rnor	121	253	152	121	117	154	0	0	19
Mmus	121	253	158	131	120	154	2	0	18
Hsap	123	253	147	153	117	152	8	4	0

Figure 3. Comparison of complete DNA sequences using the FDOD method. The FDOD method was applied to compare complete DNA sequences between autosomes. Parameter K was set to five. Euclidian distances were used to calculate differences of DNA sequences. Results for different values of parameter K were similar. For each of the three figures, gray level images (left) and gray values (matrix at right) were juxtaposed.

No matter which feature of autosomes was compared, our results show that intra-species autosomal distances were less than that between species. Hence, in terms of functional classification of genes, amino acid sequences, and complete DNA sequences, eukaryotic autosomes resemble each other within a species while that from different organisms have the inter-species difference. However, further investigations should be made when more data become available.

Based on our results and that of other authors [19] (see also http://www.nature.com/nsu/000525/000525-11.html), we propose a hypothesis (named intra-species autosomal random shuffling) as an explanation for our observations. During the evolutionary histories of eukarya, many events (e.g., duplication, transposition, recombination, inversion, etc.) have influenced sequence composition, chromosomal structure and genomic organization [19–22]. Over long time, segments of eukaryotic autosomes seem to have been shuffled randomly between autosomes within an organism. Meanwhile, each organism has its special ecological niche as well as species-specific properties, such as GC content, mutation bias, etc. Thus, autosomes within an organism tend to be similar in sequence composition and functional classification whereas greater differences developed between different species. For intra-species homogeneity in functional classification of autosomes, duplications (including duplications of genes, large segments and chromosomes) might be the primary mechanism among those exploited by intra-species autosomal random shuffling [23, 24]. However, evidence is scarce and details are still unclear. When more accurate annotations of proteins for eukaryotic genomes and more genomic data are available, functional distributions of genes on autosomes can be analyzed more thoroughly.

It is estimated that 1.5–14.5% of genes in a prokarya are related to LGT [25]. LGT imports functional units of exogenous species and reduces the discrepancy between species. If LGT was a very frequent event in the evolution of eukarya, the coding area of two organisms would be very similar. Based on our results, we suggested that LGT did not occur very frequently during the evolutionary histories of eukarya.

The mutation rate of nucleotide acid sequences in mammals is so low that statistically only 2.2×10^{-9} nucleotides change per year per site [26]. Hence, we speculate that genomic structure changes have at least the same influence as sequence mutations on the evolution of eukarya. Autosomal shuffling changed genomic structure and facilitated formation of new genes and gene orders, thus speeding up the evolutionary process of eukarya.

5. Acknowledgements

We thank Geir Skogerbo for helpful discussions. We also thank anonymous reviewers for helpful comments on improving the manuscript. This work was supported by the Chinese National "863" High-tech Program (Grant No. 2002AA231031), and Knowledge Innovation Program of the Chinese Academy of Sciences (Grant Nos. KSCX2-2-07 and KJCX1-08).

References

1. Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J. and Sutton, G.G.: The Sequence of the Human Genome, *Science* **291** (2001), 1304–1351.
2. Zhang, M.Q.: Computational Prediction of Eukaryotic Protein-Coding Genes, *Nat. Rev. Genet.* **3** (2002), 698–709.
3. Pellegrini, M., Marcotte, E.M., Thompson, M.J., Eisenberg, D. and Yeates, T.O.: Assigning Protein Functions by Comparative Genome Analysis: Protein Phylogenetic Profiles, *Proc. Natl. Acad. Sci. U.S.A.* **96** (1999), 4285–4288.
4. Pennacchio, L.A. and Rubin, E.M.: Genomic Strategies to Identify Mammalian Regulatory Sequences, *Nat. Rev. Genet.* **2** (2001), 100–109.
5. Belle, E.M., Smith, N. and Eyre-Walker, A.: Analysis of the Phylogenetic Distribution of Isochores in Vertebrates and a Test of the Thermal Stability Hypothesis, *J. Mol. Evol.* **55** (2002), 356–363.
6. Bernardi, G.: Isochores and the Evolutionary Genomics of Vertebrates, *Gene* **241** (2000), 3–17.
7. Sankoff, D.: Rearrangements and Chromosomal Evolution, *Curr. Opin. Genet. Dev.* **13** (2003), 583–587.
8. Sankoff, D. and Nadeau, J.H.: Chromosome Rearrangements in Evolution: From Gene Order to Genome Sequence and Back, *Proc. Natl. Acad. Sci. U.S.A.* **100** (2003), 11188–11189.
9. Stuart, G.W., Moffett, K. and Leader, J.J.: A Comprehensive Vertebrate Phylogeny Using Vector Representations of Protein Sequences from Whole Genomes, *Mol. Biol. Evol.* **19** (2002), 554–562.
10. Lin, J. and Gerstein, M.: Whole-Genome Trees Based on the Occurrence of Folds and Orthologs: Implications for Comparing Genomes on Different Levels, *Genome Res.* **10** (2000), 808–818.
11. Ling, L., Wang, J., Cui, Y., Hi, W. and Chen, R.: Proteome-Wide Analysis of Protein Function Composition Reveals the Clustering and Phylogenetic Properties of Organisms, *Mol. Phylogenet. Evol.* **25** (2002), 101–111.

12. Fang, W.: The Disagreement Degree of Multi-Person Judgments in Additive Structure, *Math. Soc. Sci.* **25** (1994), 85–111.
13. Fang, W.: On a Global Optimization Problem in the Study of Information Discrepancy, *J. Global Optim.* **11** (1997), 387–408.
14. Tatusov, R.L., Koonin, E.V. and Lipman, D.J.: A Genomic Perspective on Protein Families, *Science* **278** (1997), 631–637.
15. Pearson, W.R. and Lipman, D.J.: Improved Tools for Biological Sequence Comparison, *Proc. Natl. Acad. Sci. U.S.A.* **85** (1988), 2444–2448.
16. Hogenesch, J.B., Ching, K.A., Batalov, S., Su, A.I., Walker, J.R. and Zhou, Y.: A Comparison of the Celera and Ensembl Predicted Gene Sets Reveals Little Overlap in Novel Genes, *Cell* **106** (2001), 413–415.
17. Gatesy, J., Desalle, R. and Wheeler, W.: Alignment-Ambiguous Nucleotide Sites and the Exclusion of Systematic Data, *Mol. Phylogenet. Evol.* **2** (1993), 152–157.
18. Wheeler, W.C., Gatesy, J. and Desalle, R.: Elision: A Method for Accommodating Multiple Molecular Sequence Alignments with Alignment-Ambiguous Sites, *Mol. Phylogenet. Evol.* **4** (1995), 1–9.
19. Fischer, G., James, S.A., Roberts, I.N., Oliver, S.G. and Louis, E.J.: Chromosomal Evolution in Saccharomyces, *Nature* **405** (2000), 451–454.
20. Ejima, Y. and Yang, L.: Trans Mobilization of Genomic DNA as a Mechanism for Retrotransposon-Mediated Exon Shuffling, *Hum. Mol. Genet.* **12** (2003), 1321–1328.
21. Zdobnov, E.M., von Mering, C., Letunic, I., Torrents, D., Suyama, M. and Copley, R.R.: Comparative Genome and Proteome Analysis of Anopheles gambiae and Drosophila melanogaster, *Science* **298** (2002), 149–159.
22. Gilbert, W.: Why Genes in Pieces?, *Nature* **271** (1978), 501.
23. Bowers, J.E., Chapman, B.A., Rong, J. and Paterson, A.H.: Unravelling Angiosperm Genome Evolution by Phylogenetic Analysis of Chromosomal Duplication Events, *Nature* **422** (2003), 433–438.
24. Wolfe, K.H. and Shields, D.C.: Molecular Evidence for an Ancient Duplication of the Entire Yeast Genome, *Nature* **387** (1997), 708–713.
25. Garcia-Vallve, S., Romeu, A. and Palau, J.: Horizontal Gene Transfer in Bacterial and Archaeal Complete Genomes, *Genome Res.* **10** (2000), 1719–1725.
26. Kumar, S. and Subramanian, S.: Mutation Rates in Mammalian Genomes, *Proc. Natl. Acad. Sci.*_$U.S.A.$ **99** (2002), 803–808.