Compositional heterogeneity and GC-content domains in animal genomes

Eran Elhaik and Dan Graur

Department of Biology & Biochemistry, University of Houston, Houston, TX 77204-5001

Introduction

Animal genomes are not uniform in their long-range sequence composition. They are composed of a mosaic of sequence stretches of variable lengths that differ widely in their GC compositions. The isochore theory, proposed three decades ago (Macaya et al. 1976. *J. Mol. Biol.* 108:237-254), depicts the mammalian genome as a mosaic of long, fairly homogeneous genomic regions called isochores, each with a characteristic GC content. Whether or not these sequence stretches meet the criteria of isochores is a matter of debate (Cohen et al. 2005. Mol. Biol. Evol. 22:1260-1272).

In all animals studied so far, the distribution of GC-content domain lengths (plotted on a log-log scale) was found to follow a heavy-tail distribution with power-law decay exponents ranging from –1.5 to –2.5. Here, we compared the composition of homogeneous segments among the sequenced genomes of *Apis mellifera* (honeybee), *Strongylocentrotus purpuratus* (sea urchin), *Ciona savignyi* (sea squirt), *Caenorhabditis elegans* (nematode), *Drosophila melanogaster*, *Drosophila pseudoobscura*, *Drosophila simulans*, *Drosophila yakuba* (four species of fruit flies), *Anopheles gambiae* (mosquito), *Danio rerio* (zebrafish), *Tetraodon nigroviridis* (pufferfish), *Gallus gallus* (chicken), *Mus musculus* (mouse), *Rattus norvegicus* (rat), *Canis familiaris* (dog), *Bos taurus* (cow), *Monodelphis domestica* (opossum), *Pan troglodytes* (chimpanzee), and

***Homo sapiens* (human). *Saccharomyces cerevisiae* (yeast) was used as an outgroup.**

Methods

Genome Sequences

We downloaded fully sequenced eukaryotic genomes from NCBI ftp website (ftp://ftp.ncbi.nlm.nih.gov/genomes/).

Partition of Genomic Sequences into Segments that have Characteristic GC Contents and Differ Significantly from the GC Contents of Adjacent Segments

Several methods have been proposed in the literature for identifying segments with characteristic GC content. In this study, we partitioned the genomic sequences into segments by the binary recursive segmentation procedure, DJS, proposed by Bernaola-Galván et al. (1996. Phys. Rev. 53E:5181-5189). In this procedure, the chromosomes are recursively segmented by maximizing the difference in GC content between adjacent subsequences (Figure 1). The process of segmentation is terminated when the difference in GC content between two neighboring segments is no longer statistically significant (Cohen et al. 2005. Mol. Biol. Evol. 22:1260-1272).

Measuring Power Laws

We used logarithmic binning in the construction of Figure 2, which is why the points representing the individual bins appear equally spaced (Newman. 2005. Contemp. Phys. 46:323-351). To extract the exponent we employed the equation:

$$\alpha = 1 + n\left[\sum_{i=1}^{n} \ln \frac{x_i}{x_{\min}}\right]^{-1} \qquad (1)$$

Homogeneous Segment Number and Length

We defined four cutoff size classes for homogeneous segment lengths (0-10 Kb, 10-100 Kb, 100-300 Kb, >300 Kb). We classified all the homogeneous segments into these size classes, calculated their mean values and studied their distribution among all size classes (Table 1).

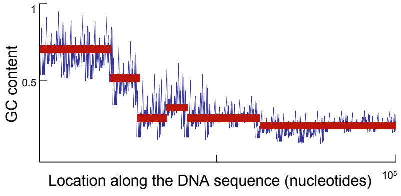

Figure 1. An illustration of the spatial distribution of GC content of non-overlapping 32 bp windows along a simulated sequence 1-million nucleotides long (blue). The segmentation algorithm yielded seven segments (red bars), including one longer than 300 Kb.

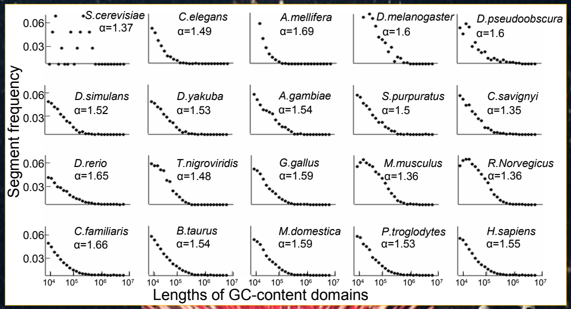

Figure 2. Length distributions of GC-content domain lengths in yeast and 19 animals and the alpha exponent of a Log-Log graph(Eq. 1).

	Number of segments in size class				Segment frequency				Mean length of segment in class			
	0-10 Kb	10-100 Kb	100-300 Kb	>300 Kb	0-10 Kb	10-100 Kb	100-300 Kb	>300 Kb	0-10 Kb	10-100 Kb	100-300 Kb	>300 Kb
S.cerevisiae	19	11	2	13	0.422	0.244	0.044	0.289	5,701	41,263	235,456	855,567
C.elegans	389	567	88	46	0.357	0.520	0.081	0.042	4,720	33,077	166,203	1,115,882
A.mellifera	4,881	4,378	283	19	0.511	0.458	0.030	0.002	4,928	29,348	161,743	426,814
D.melanogaster	926	1,444	227	47	0.350	0.546	0.086	0.018	4,955	34,297	169,234	485,182
D.pseudoobscura	1,627	1,444	227	47	0.486	0.432	0.068	0.014	4,929	33,060	167,451	554,839
D.simulans	841	908	187	65	0.420	0.454	0.093	0.032	5,003	33,386	169,360	574,392
D.yakuba	1,127	1,226	215	77	0.426	0.464	0.081	0.029	4,924	34,301	168,057	709,888
A.gambiae	2,196	2,352	390	116	0.435	0.465	0.077	0.023	4,968	34,221	166,065	680,656
S.purpuratus	1,987	15,404	529	11	0.111	0.859	0.030	0.001	6,063	28,161	152,617	369,138
C.savignyi	79	397	121	60	0.120	0.604	0.184	0.091	5,744	42,069	173,238	693,041
D.rerio	10,998	6,571	1,189	612	0.568	0.339	0.061	0.032	3,741	33,497	169,654	957,918
T.nigroviridis	1,336	1,844	354	75	0.370	0.511	0.098	0.021	4,098	36,860	165,998	469,537
G.gallus	12,772	10,574	1,646	434	0.502	0.416	0.065	0.017	3,471	34,967	168,729	529,918
M.musculus	5,884	9,133	3,589	1,839	0.288	0.447	0.176	0.090	3,219	43,073	175,805	651,922
R.norvegicus	5,577	9,379	3,612	1,788	0.274	0.461	0.177	0.088	3,221	43,098	174,943	659,975
C.familiaris	36,501	26,986	3,236	898	0.540	0.399	0.048	0.013	3,748	31,892	165,190	599,003
B.taurus	20,071	20,899	3,298	1,190	0.442	0.460	0.073	0.026	4,125	33,956	167,435	590,197
M.domestica	12,749	10,537	1,645	434	0.503	0.415	0.065	0.017	3,469	34,975	168,755	529,918
P.troglodytes	20,281	21,256	3,478	1,192	0.439	0.460	0.075	0.026	3,954	35,308	166,978	609,568
H.sapiens	28,304	27,077	4,075	1,337	0.466	0.445	0.067	0.022	3,836	34,359	165,784	615,123

Table 1. Number, frequency, and mean length of homogeneous segments in four size classes. Lowest and highest values are marked in red and blue, respectively.

Results

S.cerevisiae had the lowest number of homogeneous segments (45) and most of them were classified to the first and second size classes. *A.mellifera* had a relatively high abundance of short segments in the first size class (51%) and the second size class (46%), and the lowest abundance of long segments among all the tested genomes in the two other size classes (3% and 0.2%, respectively). Similarly to *A.mellifera*, *D.rerio*, *G.gallus*, and *C.familiaris* posses a high proportion of short segments in the first two size classes and low proportion of large segments of the two other size classes. In contrast, *M.musculus* and *R.norvegicus* have, on average, the shortest segments in the first size class (3.2 Kb) and the largest segments in the next two size classes (43 Kb and 175 Kb, respectively). They also had a high proportion (9%) of segments in the fourth.

On average, *S.purpuratus* had the longest segments in the first size class (6 Kb) and the shortest segments in the other three size classes. However, when looking at the frequency of segments in *S.purpuratus*, the second size class segments had the highest proportion among all species (86%) and the other size classes had the lowest proportion among all species (11%, 3%, 0.1%). On average, *C.elegans* posses the longest segments in the fourth size class (1,116 Kb).

A comparison of the distributions of GC-content lengths among yeast and 19 animal genomes is shown in Figure 2. We used a G goodness-of-fit test to determine that none of the distributions of segment length are similar to one another. In all animals studied, the distribution of GC-content domain lengths (plotted on a log-log scale) was found to follow a heavy-tail distribution with power-law decay exponents ranging from 1.35 to 1.69. In yeast, no such behavior was observed. The compositionally homogeneous segments in all the animal sequences do not have a characteristic length. Rather, there is an abundance of short segments and only a small number of long ones.

Conclusions

The segments of *A.mellifera* and *S.purpuratus* had a high abundance of short segments in the first two size classes. However, it is reasonable to assume, that these findings are affected by the quality of the assembly that still contains many scaffolds.

We used a rigorous method to calculate the power-law exponent and found that the distribution of GC-content domain lengths in all animals follow a narrow heavy-tail distribution with power-law decay exponents ranging from –1.36 to –1.69. This range is narrower than the range described by Cohen et al. (2005. Mol. Biol. Evol. 22:1260-1272), who used a less rigorous method. The power-law exponents cluster according to taxa, such as human-chimpanzee, mouse-rat and insects (fruit flies, mosquito, and honeybee). The power law property has so far been found in all multicellular organisms but not in unicellular organisms. We therefore conclude that genomic GC composition may obey similar rules in all metazoans.

Reference

Bernaola-Galván, P., R. Roman-Roldan, and J. L. Oliver. 1996. Compositional segmentation and long-range fractal correlation in DNA sequences. Phys. Rev. E. 53:5181-5189.

Cohen N, Dagan T, Stone L, Graur D (2005) GC composition of the human genome: in search of isochores. Mol Biol Evol 22:1260–1272

Newman M. E. J (2005) Power laws, Pareto distributions and Zipf's law. Contemporary Physics. 46: 323-351

Sodergren, E., Weinstock, G. M., ... Elhaik, E., Graur, D... ...Thorn, R., and Wright, R. 2006. The genome of the sea urchin *Strongylocentrotus purpuratus*. Science 314:941-952

Weinstock G. M., Robinson, G. E., ... Elhaik, E., Graur, D... ...Villasana, D., and Wright, R. 2006. Insights into social insects from the genome of the honeybee *Apis mellifera*. Nature 443: 931-949.

Acknowledgement

This work was supported by NSF grant DBI-0543342.
We thank Drs. Christine G. Elsik, and Justin Reese (Texas A&M University) for introducing us to the genome sequencing frame of mind.