Compositional Analysis of Homogeneous Regions in Human Genomic DNA

Eric C. Rouchka[1] and David J. States[2]
WUCS-2002-2

March 19, 2002

[1]Department of Computer Science
Washington University
Campus Box 1045
One Brookings Drive
Saint Louis, MO 63130-4899

[2]University of Michigan School of Medicine
Medical Science Building II, Room 5622A
Ann Arbor, MI 48109

ecr@sapiens.wustl.edu; dstates@umich.edu

Compositional Analysis of Homogeneous Regions in Human Genomic DNA

Eric C. Rouchka[1] and David J. States[2]
ecr@sapiens.wustl.edu; dstates@umich.edu

[1]Department of Computer Science, Washington University, St. Louis, Missouri, 63110, USA
[2]University of Michigan School of Medicine, Ann Arbor, Michigan, 48109, USA

Abstract

Due to the increased production of human DNA sequence, it is now possible to explore and understand human genomic organization at the sequence level. In particular, we have studied one of the major organizational components of vertebrate genome organization previously described as isochores (Bernardi, 1993), which are compositionally homogeneous DNA segments based on G+C content. We have examined sequence data for the existence of compositionally differing regions and report that while compositionally homogeneous regions are present in the human genome, current isochore classification schemes are too broad for sequence-level data.

Keywords: isochores, sequence homogeneity, G+C content

Introduction

It has been proposed that vertebrate genomes, including human, are made up of compositionally homogeneous DNA segments based on G+C content (Bernardi, 1993). These regions, known as isochores, have been studied experimentally using density gradient centrifugation on mechanically sheared DNA in the range of 50-100 kb (Bernardi, 1993) since their discovery in the mid '70s (Macaya, et. al., 1976). Isochores are biologically interesting due to the association between increasing G+C content and high gene density (Mouchiroud, et al., 1991; Gardiner, 1996; Zoubiak et al., 1996).

According to Bernardi's theories, there are five families of isochores, each having a different level of cytosine and guanine (C and G, respectively) as described in Table 1. There are two G+C-poor isochore families L1 and L2 that make up approximately 60% of the human genome. The isochore family L1 is defined to be regions corresponding to less than 37% G+C content; L2 is defined to be regions containing between 37% and 41% G+C. The isochore family H1 forms 24% of the human genome and corresponds to regions between 41% and 46% G+C. The other G+C rich isochore family H2 forms 7.5% of the human genome and corresponds to those regions containing between 46% and 53% G+C. The final isochore family, H3 forms almost 5% of the genome and corresponds to those very G+C rich regions which are greater than 53% G+C. Since the overall composition of the human genome is approximately 60% AT and 40% GC, the L1 and L2 families correspond to isochore regions containing less than average G+C content while the H1, H2, and H3 families correspond to isochore regions containing higher than average G+C content. The availability of human genomic sequence makes it possible to explore and understand human DNA composition at a sequence level. We attempted to correlate Bernardi's isochore family definition to sequence data.

Isochore Class	Range	Percent of Genome
L1	0-37% GC	60[A]
L2	37-41% GC	
H1	41-46% GC	24
H2	46-53% GC	7.5
H3	53-100% GC	4.7

Table 1: Isochore classifications. Indicated are the GC ranges for each of the five isochore classifications as defined by Bernardi (2000). The remaining 3.8% of human genomic DNA corresponds to satellite repeats and ribosomal sequences (Bernardi , 2000). [A]Note that the L1 and L2 isochore classes together represent 60 percent of the human genome.

Methods

Analyzing Homogeneous Segments

In order to study the validity of Bernardi's definitions on a sequence level and to examine more properties of the homogeneous regions found in human sequence data, we took the contig sequences for each chromosome available

in the April 2001 release of UCSC's Goldenpath (Kent and Haussler, 2001). For each of these chromosomes, we examined the effect of varying the fragment size. This was accomplished by segmenting each chromosome into all possible fragments of 1 kb, 5 kb, 10 kb, 20 kb, 50 kb, 75 kb and 100 kb. For each fragment size, there are 101 possible bins into which each fragment could be placed. Each bin represents a G+C percentage, from 0 to 100. We calculated the G+C percentage for each fragment, and then increased the total counts for the appropriate bin. The histograms were compared to determine the effect of variable fragment size and compositional variation from one chromosome to another. Chi-squared analysis was applied in order to compare the G+C distributions among the chromosomes. In addition, we calculated the frequency of the dinucleotide CG within each bin in order to test whether or not a correlation exists between G+C content and the occurrence of CpG dinucleotides.

An attempt to validate Bernardi's classifications was made by calculating where isochore boundaries should be based on the percentage of the genome that belongs to each of his classifications. This was accomplished by calculating which histogram bin represents the first 60% of the genome, the next 24%, the next 7.5%, and the next 4.7%.

Sequence Homogeneity

The term "isochore" implies a level of high sequence homogeneity. In order to test the validity of this point, we examined 80 different contigs greater than 10 MB in length available through the August 2001 Goldenpath human genome assembly (Kent and Haussler, 2001). The total sequence length of these contigs is over 2 GB in length, representing nearly 2/3 of the human genome. At 1 KB intervals, we calculated the G+C percentage for a surrounding 1 KB, 10 KB, 50 KB, 100 KB, 500 KB, 1 MB and 3 MB window. The variation in the G+C content was calculated and reported. In addition, random sequences were generated corresponding to the lengths of each of the contigs with the following frequencies: A = 0.30, C = 0.20, G = 0.20 and T = 0.30. The same tests in variation were tested for the randomized sequences.

Results

Isochore Classifications

Chi-squared analysis was performed on the seven different window sizes (1 kb, 5 kb, 10 kb, 20 kb, 50 kb, 75 kb and

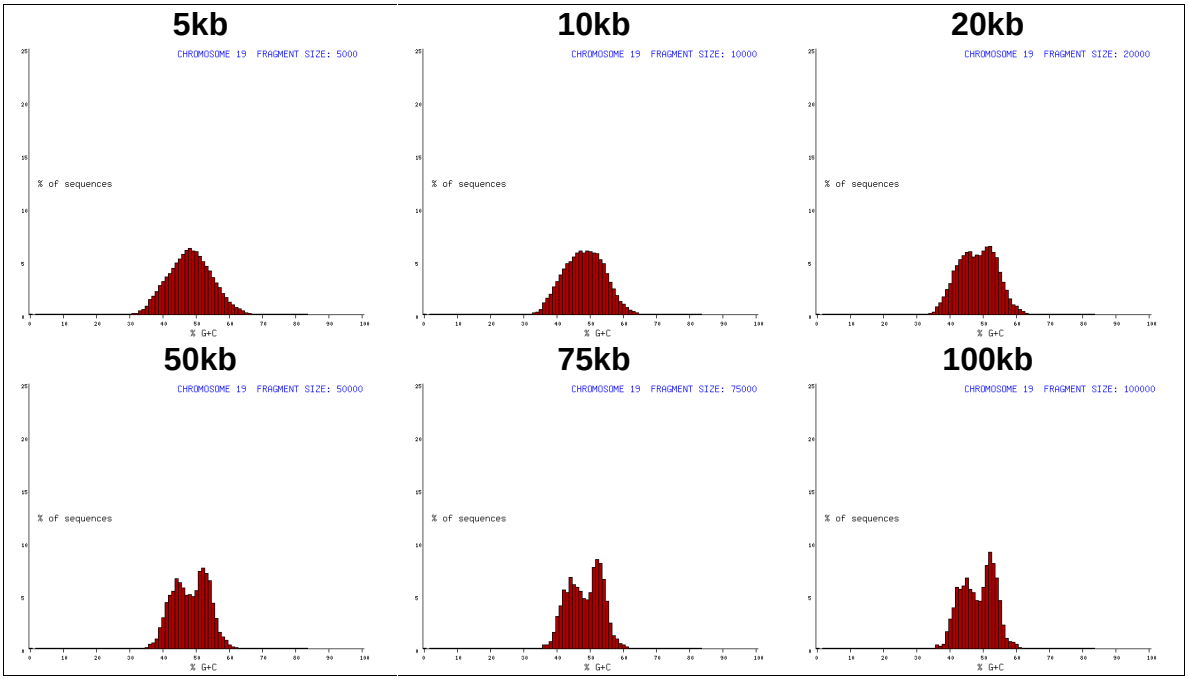


Figure 1: Chromosome 19 C+G histograms. Shown in this figure from top left to bottom right are the resulting C+G histograms for chromosome 19 (extracted from the Goldenpath April 2001 release) using 5kb, 10kb, 20kb, 50kb, 75kb, and 100kb fragments. This graph illustrates that the distribution of C+G within a particular chromosome is dependent on the fragment sizes that are used.

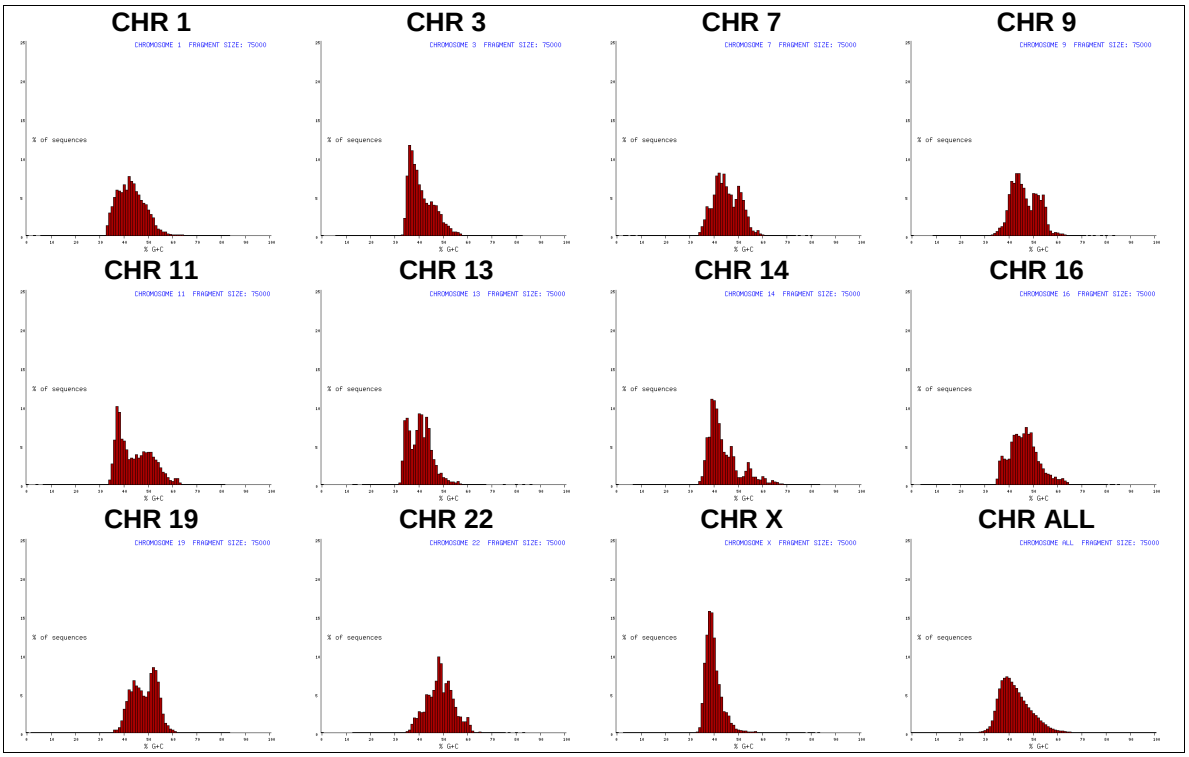


Figure 2: Chromosomal histograms for 75kb fragments. Shown in this figure are the resulting G+C histograms for the following chromosomes: Row 1: (left to right): 1, 3, 7, 9. Row 2: 11, 13, 14, 16. Row 3: 19, 22, X, ALL. The X-axis represents the G+C content, and the Y-axis represents the percentage of fragments falling within a given G+C content. These histograms were created using the April 2001 Goldenpath release (http://genome.ucsc.edu).

100 kb) for each chromosome in a pair-wise fashion. In each case, the null hypothesis that the distributions of G+C fragments are independent of the window size can be rejected (results not shown). Thus, the isochore classification schemes are highly dependent on the fragment sizes being studied. In the case of the five-class system, the results were skewed towards fragments in the range of 50 kb to 100 kb due to the use of density gradient centrifugation. Figure 1 graphically illustrates a dependence on window size with chromosome 19. By looking at this figure, it can be seen that when a smaller fragment size (5 kb) was used when studying chromosome 19, a unimodal distribution of G+C fragments is observed. When the window size was increased (50 kb - 100 kb), a bimodal distribution of G+C fragments can be seen.

In order to determine whether or not G+C content distribution is chromosome specific, Chi-squared analysis was performed (results not shown). The distributions of G+C fragments using 75 kb windows was compared for each pair of chromosomes. The null hypothesis that the G+C content distribution of any two given chromosomes is similar was rejected, no matter which two chromosomes were compared. Displayed in figure 2 is the distribution of G+C fragments using a 75 kb window for eleven different chromosomes and the genome as a whole. As this figure shows, there are vast differences in the G+C fragment distribution among chromosomes. Some chromosomes, such as 1 and X, appear to have a distinct unimodal distribution of fragments at the 75 kb window level. Other chromosomes, such as 9, 11 and 19 seem to have distinct bimodal distributions in the G+C fragments. However, in none of the cases were there more than two distinct peaks in the distribution of G+C fragments. Our results show the difficulty of defining isochore boundaries based on sequence data alone. We do see, however, that there does appear to be two distinct isochores that were observable: the majority that are in low G+C, and those that are high in G+C. Further division of these two major groups based on sequence data appears to be a difficult, if not impossible, task.

According to the density gradient centrifugation experiments performed by Bernardi, 60% of the human genome falls into an L1+L2 isochore classification, 24% is H1, 7.5% is H2, and 4.7% is H3. Table 2 was created using these guidelines to split the histograms for 75 kb fragments for the various chromosomes into densities of 60%, 84%, and 91.5%, which would theoretically find the isochore boundaries. Not surprisingly, we see that when all of the chromosomal data was inspected, 60% of the histograms lie

Isochore Boundary locations based on total percent of all fragments

Chromosome	60% of all fragments L2-H1 Boundary	84% of all fragments H1-H2 Boundary	91.5% of all fragments H2-H3 Boundary
BERNARDI	42% G+C	47% G+C	53% G+C
1	44% G+C	49% G+C	51% G+C
2	44% G+C	47% G+C	49% G+C
3	41% G+C	47% G+C	49%G+C
4	40% G+C	43% G+C	45% G+C
5	41% G+C	44% G+C	46% G+C
6	39% G+C	43% G+C	45% G+C
7	46% G+C	51%G+C	52% G+C
8	42% G+C	45% G+C	49% G+C
9	47% G+C	53% G+C	54% G+C
10	44% G+C	48% G+C	49% G+C
11	46% G+C	52% G+C	55% G+C
12	44% G+C	48% G+C	50% G+C
13	41% G+C	44% G+C	47% G+C
14	43% G+C	51% G+C	55% G+C
15	43% G+C	46% G+C	47% G+C
16	47% G+C	51% G+C	55% G+C
17	49% G+C	52% G+C	54% G+C
18	41% G+C	44% G+C	46% G+C
19	51% G+C	54% G+C	55% G+C
20	47% G+C	50% G+C	53% G+C
21	50% G+C	55% G+C	56% G+C
22	50% G+C	54% G+C	56% G+C
X	40% G+C	43% G+C	45% G+C
Y	39% G+C	42% G+C	43% G+C
ALL	43% G+C	48% G+C	51% G+C

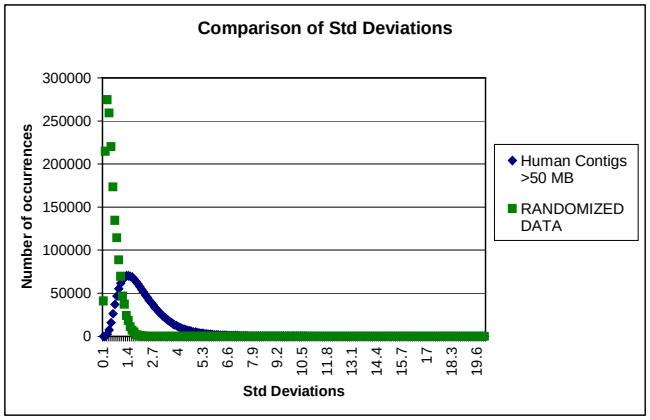
Table 2: Boundary locations based on total percent of all fragments. Shown in column 1 is the chromosome label. Column 2 indicates the breakpoint where 60% of all 75 kb fragments for the given chromosome lie. Column 3 indicates the breakpoint under which 84% of all 75kb fragments lie. Column 4 indicates the breakpoint under which 91.5% of all 75 kb fragments lie. Note that the breakpoints of 60%, 84%, and 91.5% indicate breakpoints for the defined isochore classes L2-H1, H1-H2, and H2-H3 (Bernardi, 2000).

at 43% G+C or less, which is just above the cutoff for the L2-H1 isochore boundaries. 84% of the histograms lie at 48% G+C or less, which is just above the cutoff for H1-H2 isochores. 91.5% of the histograms lie at 51% G+C, or slightly less than the H2-H3 isochore cutoff of 53% G+C. However, Table 2 also shows that these cutoffs do not correlate with isochore boundaries for all chromosomes. Some chromosomes, such as chromosomes 9, 11, 14, 16, 17, 19, 21 and 22 have more fragments that are G+C rich, while other chromosomes such as 4, 5, 6, 13, 18, X and Y have more fragments that are G+C poor. These results suggest that calculating the isochore boundaries based on the fragment density is not valid when applied to individual chromosomes.

Sequence Homogeneity

Figure 3 illustrates the distribution of standard deviations in G+C content for every 1000th base in both the randomly generated contigs and Goldenpath contigs greater than 10 MB in length. The mean was computed by calculating the G+C content for windows of 1 KB, 10 KB, 50 KB, 100 KB, 500 KB, 1 MB and 3 MB. As figure 3 shows, the distribution of standard deviations for the random sequence is much tighter and closer to zero than the distribution of standard deviations for the actual human sequence. Figure 4 shows the calculated cumulative percentage of standard deviations. Examination of this data indicates that in random sequence data, 50% of the points examined have a standard deviation in G+C content of ± 0.4%, while for the real sequence data this number is ± 1.8%. 75% of all random points have a standard deviation of ± 0.7% or less, while this number grows to ± 2.6% in the real sequence data. 95% of all random fragments have a standard deviation of ± 1.2%. This number grows to ± 4.5% in the real sequence. In fact, only 24% of all real sequence data has a standard deviation of ± 1.2% or less. These results indicate that the human genome is much more heterogeneous than the theories of Bernardi (1993) lead one to believe.

A) Distribution of standard deviations from a mean G+C content



B) Cumulative percentage of standard deviations from a mean G+C content

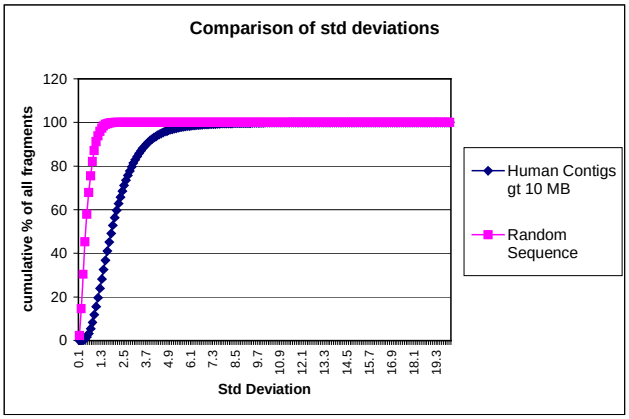


Figure 3: Distribution of standard deviations from a mean G+C content. Shown in A) is the count of each standard deviation calculated for every 1000th base in human and randomized contigs using window sizes of 1 KB, 10 KB, 50 KB, 100 KB, 500 KB, 1 MB and 3 MB. B) shows the cumulative percentage of standard deviations from figure 3 falling under a certain percentage.

Discussion

In order to understand the concept of a 5-class isochore system as proposed by Bernardi, it is important to revisit the experimental procedures performed over 25 years ago. In the article where isochores were first described (Cuny et al., 1981), human genomic DNA was found to be fractionated into five major components using CsCl density profiles. Each component represents a set of DNA segments that sediments differently based on different buoyant densities. The results presented are based on earlier analyses of the composition of eukaryotic genomes (Thiery, Macaya and Bernardi, 1976). Thiery et al. (1976) looked at the separation of human DNA using thirteen different density gradients. What results are thirteen different Gaussian distributions of absorbance, each representing a different distribution of genomic DNA based on G+C content. Three main observations of the experimental work are discussed.

First of all, the decision to choose five major components (later given the label "isochores" by Cuny, et al., 1981) seems somewhat arbitrary. In fact, examination of Figure 1 of Thiery, et al. (1976) indicates that any of the thirteen different results could have been chosen as major components. In addition, if more than thirteen different density gradients were examined, a different distribution of major components could potentially result.

The second critique is that the Gaussian distributions resulting for each of the labeled major components are overlapping. This means, for instance, that a fragment of human genomic DNA containing an average G+C content of 47% could potentially wind up belonging to multiple major components, or isochore families. This is a major problem when looking at a sequence level comparison. It is a necessary requirement that each individual sequence fragment be assigned to a single classification, or at most, belong to an unknown area between two breakpoints.

The final critique is that density gradient centrifugation experiments can only allow for the fractionation of DNA based on the overall G+C content of any segment. It does not seem to be in any way possible to determine the homogeneity. In fact, the only means by which homogeneity can be discerned is by looking at finished sequence data.

The density gradient centrifugation experiments are important in that they indicate that there are larger regions of the human genome with a conserved low or high G+C content. However, the previous school of thought of a five-class isochore system for the human genome with strict boundaries appears to be out-of-date in light of the availability of sequence data.

Our results have shown the difficulty of defining isochore boundaries based solely on sequence data. This is supported by failed attempts of window-based sequence segmentation resulting in arguments against strict definitions of isochore classes (IHGSC, 2001; Nekrutenko and Li, 2000; Häring and Kypr, 2001). We do see, however, that there does appear to be two different classes of isochores that can be observed: the majority that are low in G+C, and those that are high in G+C. Further breakdown of these two major groups based on the sequence data appears to be a difficult task.

Acknowledgements

We wish to thank Zhengyan Kan and Warren Gish for helpful input in preparation of the manuscript. Financial support for this work was provided in part through grants from the National Institutes of Health (HG-R01-01391; 5-T32-HG00045), the Department of Energy (DE-FG02-94ER61910), and the Merck Foundation for Genome Research (grant #225).

References

Bernardi, G. (1993) "The isochore organization of the human genome and its evolutionary history -- a review." *Gene*, **135**:57-66.

Cuny, G., Soriano, P., Macaya, G., Bernardi, G. (1981) "The Major Components of the Mouse and Human Genomes." *European Journal of Biochemestry*, **115**:227-233.

Gardiner,K. (1996) "Base composition and gene distribution: critical patterns in mammalian genome organization." *Trends in Genetics*, **12**(12):519-524.

Häring, D., Kypr, J., (2001) "No isochores in human chromosomes 21 and 22?" *Biochem. Biophys. Res. Comm*. **280**(2):567-573.

International Human Genome Sequencing Consortium (IHGSC), (2001) "Initial sequencing and analysis of the human genome." *Nature*, **409**:860-921.

Kent, J.W., Haussler, D., (2001) "Assembly of the Working Draft of the Human Genome with GigAssembler." *Genome Research*, **11**(9):1541-1548.

Macaya, G., Thiery, J.P., Bernardi, G. (1976) "An approach to the organization of eukaryotic genomes at a macromolecular level." *Journal of Molecular Biology*, **108**(1): 237-254.

Mouchiroud, D., D'Onofrio, G., Aissani, B, Macaya, G., Gautier, C. Bernardi, G. (1991) "The distribution of genes in the human genome." *Gene*, **100**:181-187.

Nekrutenko, A., Li, W.H. (2000) "Assessment of compositional heterogeneity within and between eukaryotic genomes." *Genome Research*, **10**(12):1986-1995.

Thiery, J.-P., Macaya, G., Bernardi, G. (1976) "An Analysis of Eukaryotic Genomes by Density Gradient Centrifugation." *Journal of Molecular Biology*, **108**:219-235.

Zoubak, S., Clay, O., Bernardi, G. (1996) "The gene distribution of the human genome." *Gene*, **174**:95-102.