1 Introduction

Nuclear genomes of vertebrates are composed of mosaic structures having different GC content (GC content is the molar ratio of guanine + cytosine in nucleotide sequence). These structures are called isochores, and each segment length is about 300kb [1]. Previous research has shown that different GC region of chromosome is correlated with several properties: the staining intensity of chromosomal bands [2] and gene density in GC-rich regions and GC-poor regions of isochores [3]. Then, most of the genes expressed ubiquitously in various tissues as a housekeeping gene are in the GC-rich regions [4]. It is considered that the isochores are related to the gene expression [3]. In addition, the GC content of gene coding region is related to the GC content of surrounding region of genome, for example GC-rich genes tend to present in GC-rich region of isochores and vice versa [5]. These facts indicate that GC content is one of the most important factors for considering the meaning of genes in terms of the expression in cells.

In the accompany paper, we showed that some of proteins coded in the total genome have charge periodicity of 28 residues (PCP28). PCP28 were obtained by using the autocorrelation function of electric charges of amino acid sequences and had the significant peak at 28 residues periodicity in the plot of the function. More PCP28 were found in the eukaryotic genomes than in the prokaryotic genomes. Most of PCP28 are nuclear proteins which play important roles, such as gene regulation, genome replication, and so on. For more understanding of the role of PCP28, we investigated the GC content of PCP28.

In this work, we analyzed the human genome focusing on PCP28, GC content, and gene expression. First, we investigated the population of PCP28 from the human genome as a function of the GC content. The result indicated that the GC content for PCP28 located in the nucleus was very low while that for PCP28 in other localization showed broad distribution. The low GC distribution is usually correlated with the low expression level of PCP28 [3]. Therefore, we investigated the expression pattern of PCP28 in several types of cells by using the published human genes expression data of microarray (http://symatlas.gnf.org).

2 Materials and Methods

We obtained the data of human genome from NCBI (BUILD 36.1). GC content of each gene was calculated from nucleotide sequence of its coding region (exon only) in human genome. Genes were divided into five classes by their locating band classes in chromosomes. We used the classification of chromosomal available in published data, in which bands were divided into five classes based on their staining intensity [2]. For the determination of PCP28, we used the auto correlation function of charge. Prediction of nuclear localized protein was performed with the global and local physicochemical properties of amino acid sequence. Human genes expression data of microarray chip was obtained from the web site of Genomics institute of the Novartis Research Foundation (http://symatlas.gnf.org). For the accurate analysis, we adopted gene which reproducibly expressed in identical experiments.
3 Results and Discussion

When the population of whole genes and PCP28 was plotted as a function of GC contents, the distribution of whole genes was broad, whereas that of PCP28 was skewed toward low GC with a peak around the GC content of 0.42. Possible explanation for the characteristic distributions is that they are composed of some GC content groups. Therefore, we first investigated the relationship between GC contents of genes and chromosomal Giemsa band where the genes coded. Chromosomal band regions were classified into five groups by their staining intensity. Then, genes were divided into these band groups by their locations on chromosome. Distributions of band groups were formed different shapes; especially, the most intense staining group (positive100) and no staining group were significantly different. In addition, the shape of distribution of positive 100 was partially similar to that of PCP28. From these results, it appears that the distribution of PCP28 was composed of the low GC content group and other GC content groups.

For the determination of GC content groups in PCP28, the distribution of PCP28 were fitted with three distributions which were simulated by changing GC occurrence at codon 3rd positions and maintaining the amino acid sequence. In the end of fitting, we obtained the well fitted distribution with three simulated distributions which GC occurrences are random, high (0.89) and low (0.34), respectively. We thought that these groups were related to the role of these proteins in the living cell. Then, we predicted the nuclear localized proteins from PCP28 by using physicochemical properties of amino acid sequences, and these distributions of nuclear and other proteins were fitted with above-mentioned three simulated distributions. As for the nuclear proteins, low GC groups mainly contributed to the formation of the skewed distribution. On the other hand, the distribution of other proteins consisted of mainly random GC group.

If the GC content of gene is correlated with its expression level in the living cell, the GC content groups for the fitting should show the different expression patterns; for example, simulated low GC groups may be rarely expressed in the living cell. Thus we investigated the gene expression pattern in the living cell by using the microarray gene expression data of human which was obtained through the web. In this analysis, we focused on the expression pattern of the genes which are expressed in the particular stage of the development. First, we extracted the fetal-specific (expressed during the fetal period) genes and adult-specific (expressed during the adult period) genes from the three kinds of tissues (whole brain, liver and lung) by comparing the expression patterns of fetal with that of adult on same kind of tissue. Next, we plotted the histograms. In the nuclear proteins, distribution of fetal-specific genes differed from that of adult-specific genes; the shape of distribution of fetal-specific genes skewed toward low GC, whereas that of adult-specific genes skewed toward high GC. These results indicate the correlation between expression level of gene and its GC content, because the fetal-specific genes are expressed during the fetal period that is shorter than adult period. In the other proteins, the shapes of distributions of fetal-specific and adult-specific genes were almost same and broad as a simulated random distribution, which indicates the less correlation between expression level of gene and its GC content. In our opinion, expression of nuclear proteins of PCP28 should be regulated by some systems related to GC content as isochores structures of chromosome, whereas the expression of the other proteins should be regulated by other systems no related to GC content.

References


