Available online at www.sciencedirect.com

SCIENCE DIRECT

Gene 350 (2005) 129–136

GENE
SECTION
EVOLUTIONARY GENOMICS

www.elsevier.com/locate/gene

Genome-wide analysis reveals strong correlation between CpG islands with nearby transcription start sites of genes and their tissue specificity

Riu Yamashita[a,b], Yutaka Suzuki[c], Sumio Sugano[c], Kenta Nakai[a,*]

[a]*Human Genome Center, Institute of Medical Science, University of Tokyo, 4-6-1, Shirokane-dai Minato-ku, Tokyo 108-8639, Japan*
[b]*Undergraduate Program for Bioinformatics and Systems Biology, Faculty of Science, University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan*
[c]*Laboratory of Functional Genomics, Department of Medical Genome Sciences, Graduate School of Frontier Sciences, University of Tokyo, 4-6-1 Shirokane-dai Minato-ku, Tokyo 108-8639, Japan*

Received 11 November 2004; received in revised form 28 December 2004; accepted 24 January 2005
Available online 19 March 2005
Received by T. Gojobori

Abstract

It has been envisaged that CpG islands are often observed near the transcriptional start sites (TSS) of housekeeping genes. However, neither the precise positions of CpG islands relative to TSS of genes nor the correlation between the presence of the CpG islands and the expression specificity of these genes is well-understood. Using thousands of sequences with known TSS in human and mouse, we found that there is a clear peak in the distribution of CpG islands around TSS in the genes of these two species. Thus, we classified human (mouse) genes into 6600 (2948) CpG+ genes and 2619 (1830) CpG− ones, based on the presence of a CpG island within the −100: +100 region. We estimated the degree of each gene being a housekeeper by the number of cDNA libraries where its ESTs were detected. Then, the tendency that a gene lacking CpG islands around its TSS is expressed with a higher degree of tissue specificity turned out to be evolutionarily conserved. We also confirmed this tendency by analyzing the gene ontology annotation of classified genes. Since no such clear correlation was found in the control data (mRNAs, pre-mRNAs, and chromosome banding pattern), we concluded that the effect of a CpG island near the TSS should be more important than the global GC content of the region where the gene resides.

Keywords: CpG islands; Tissue specificity; Housekeeping genes; Isochores

1. Introduction

Most CpG dinucleotides are methylated in vertebrates and most of them are thought to have depleted over evolutionary time by the transition mutation from methylated cytosine to thymine (Bird and Taggart, 1980). However, CpG clusters, called 'CpG islands,' are still found in several vertebrate genomes (Gardiner-Garden and Frommer, 1987), and it has been reported that the dinucleotides in CpG islands are usually unmethylated, thus avoiding the above mutations (Larsen et al., 1992). CpG islands often reside in the promoter region of genes, and the methylation of CpGs in these regions is thought to affect the expression of their downstream genes. For example, several transcription factors are known to exhibit differential activities depending on whether or not their *cis*-elements are methylated (Cross and Bird, 1995; Costello et al., 2000). There have been several reports examining the effect of CpG islands on the tissue specificity of nearby gene expression. The pioneering work by Gardiner-Garden and Frommer (1987) and Larsen et al. (1992) reported that promoters of most housekeeping genes contain CpG islands while promoters of not a few tissue-specific genes lack of them, based on the analysis of GenBank sequences. However, their analyses were based on the very small subset of genes available in those days. The deduced results should be confirmed to hold true for the post-genome-era view of the CpG islands.

Abbreviations: TSS, transcription start site; GO, gene ontology; EST, expressed sequence tag.

* Corresponding author. Tel.: +81 3 5449 5131; fax: +81 3 5449 5133.
E-mail address: knakai@ims.u-tokyo.ac.jp (K. Nakai).

0378-1119/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.gene.2005.01.012

In addition, recent studies examining the relationship between the isochore structures and the tissue specificity have made the situation more complicated (Bernardi, 2000a,b, 2001; Eyre-Walker and Hurst, 2001). GenBank-based sequence analyses have shown that heavy (GC-rich) isochores tend to contain housekeeping genes, while light (GC-poor) isochores are rich in tissue-specific genes (Bernardi, 1995; Pesole et al., 1999). A genome-wide analysis based on the SAGE data also indicates that housekeeping genes tend to be found in the regions of high GC content (Versteeg et al., 2003). Since the CpG islands tend to exist in heavy isochores, this seemed to be in good agreement with the results obtained by Gardiner-Garden and Frommer and Larsen et al. However, there are also some reports denying any correlation between the GC content and the tissue specificity. For example, no significant expression difference was observed between light and heavy isochores in studies based on the EST data (Goncalves et al., 2000; Ponger et al., 2001). Moreover, the correlation may not be evolutionarily conserved: Vinogradov (2003) reported that there is a correlation between the GC content of the third position of codons and their tissue specificity in human, but not in mouse.

To our opinion, the current confusion has been partly derived from the lack of precise information about the CpG islands. The above contradiction should be resolved by considering the presence of precisely described nearby CpG islands. In such analyses, the knowledge of precise transcriptional start sites (TSS) must be very important to find the precise correlation between TSS and CpG islands. Because many cDNA sequences in GenBank are imperfect in their 5′ ends, the estimation of some of their TSS can be severely inaccurate for the presence of unidentified first introns. In fact, a recent comprehensive analysis of about 29,000 CpG islands that were identified in the draft human genome sequence did not show any clear relationship with the positions of TSS (Lander et al., 2001). However, this analysis did not use comprehensive data of precise TSS.

We have constructed a database of TSS (DBTSS) (Suzuki et al., 2002, 2004), which contains a number of human/mouse TSS determined by the oligo-capping (Maruyama and Sugano, 1994) or Cap-trapper methods (Carninci et al., 1997). In this study, the data in DBTSS were used to study the possible correlation between the presence of CpG islands in promoters and their tissue specificity.

2. Materials and methods

2.1. Sequence data

TSS data were obtained from DBTSS version 3 (Suzuki et al., 2002, 2004). The mapping information of their corresponding NCBI RefSeq cDNA sequences onto human or mouse genomes was obtained from the UCSC genome web site (‘refGene.txt’). Because RefSeq contains several splicing variants in the same locus, we could not get an agreement between RefSeq genes and TSS in several cases. Thus, variants were regarded as belonging to the same gene if their LocusLink IDs, obtained from NCBI’s ‘loc2ref’ table, are the same. All DBTSS clones that start downstream of translational start sites were removed, using the ‘refGene.txt’ coding sequence (CDS) information. If multiple TSS were observed in one LocusLink ID, the representative TSS was selected, as shown in Fig. 1. Based on the position of representative TSS (+1), sequences from −2000 to +2000 (denoted like −2000: +2000) were obtained. Sequences that contain more than five “N” characters, which can be an uncertain or unsequenced region, were removed. ‘TSS-flanking regions’ were defined as the −100: +100 region. ‘Mature mRNA sequences’ and ‘pre-mRNA sequences’ were obtained using the ‘refGene.txt’ table (November 2002).

2.2. Detection of CpG islands

For the definition of CpG islands, the method by Gardiner-Garden and Frommer was used; that is, GC content ((#[C]+#[G])/200, where #[*N*] denotes the number of nucleotide “N” within a window) and CpG score (#[CG]/(#[C]*#[G])*200) with a 200 base-window size. Thus, we classified genes to be CpG positive (CpG+) when its GC content in (−100: +100) exceeds 0.5 and when its CpG score in the same region exceeds 0.6; otherwise, they are classified as CpG negative (CpG−) genes. All data are available at ftp://ftp.hgc.jp/pub/hgc/db/dbtss/.

2.3. Estimation of expression profile

Data on the tissue specificity of gene expression were obtained from the NCBI UniGene database (Wheeler et al., 2003): each UniGene entry contains information on the source of EST libraries from which the clones were taken. Thus, the number of library sources for each gene was counted and this number was regarded as the degree of ubiquitous gene expression. The NCBI LocusLink ‘loc2UG’

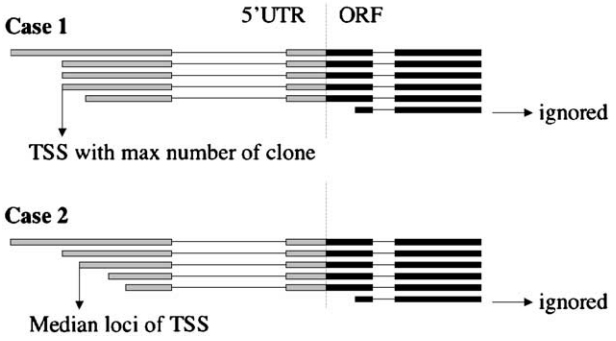

Fig. 1. How to select a representative TSS. Case 1: If there is a single most frequent TSS in a gene, we regarded it (the mode) as a representative. Case 2: Otherwise, we chose the clone corresponding to the median as a representative.