Available online at www.sciencedirect.com

SCIENCE DIRECT

Gene 358 (2005) 127 – 138

GENE
SECTION
FUNCTIONAL GENOMICS

www.elsevier.com/locate/gene

The properties of CpG islands in the putative promoter regions of human immunoglobulin (Ig) genes

Guang B. Liu [a,e,*], Hong Yan [b], Ya F. Jiang [c], Rong Chen [b], John D. Pettigrew [a], Kong-Nan Zhao [b,d,*]

[a] *Vision, Touch and Hearing Research Centre, Faculty of Biomedical Sciences, The University of Queensland, Brisbane, Qld 4072, Australia*
[b] *Department of Health Statistics, School of Medicine, Xi'an Jiaotong University, Xi'an, PR China*
[c] *School of Nursing, Peking Union Medical College, Ba Da Chu Road, Beijing, PR China*
[d] *Centre for Immunology and Cancer Research, The University of Queensland, Princess Alexandra Hospital, Brisbane, Qld 4102, Australia*
[e] *Department of Biological and Physical Sciences, Faculty of Science, The University of Southern Queensland, Toowoomba, Qld 4350, Australia*

Received 4 March 2005; received in revised form 25 May 2005; accepted 9 June 2005
Available online 19 August 2005
Received by W. Makalowski

Abstract

CpG island is a GC-rich motif occurred in gene promoter region, which can play important roles in gene silencing and imprinting. Here, we present a set of discriminant functions that can recognize the structural and compositional features of CpG islands in the putative promoter regions (PPRs) of human and mouse immunoglobulin (Ig) genes. We showed that the PPRs of both human and mouse Ig genes irrespective of gene chromosomal localization are apparently CpG island poor, with a low percentage of the CpG islands overlapped with the transcription start site (TSS). The human Ig genes that have CpG islands in the PPRs show a very narrow range of CpG densities. 47% of the Ig genes fall in the range of 3.5–4 CpGs/100 bp. In contrast, the non-Ig genes examined have a wide range of the density of CpG island, with 10.5% having the density of 8.1–15 CpGs/100 bp. Meantime, five patterns of the CpG distributions within the CpG islands have been classified: Pat A, B, C, D, and E. 21.6% and 10.8% of the Ig genes fall into the Pat B and Pat D groups, respectively, which were significantly higher than the non-Ig genes examined (8.2% and 3.8%). Moreover, the length of CpG islands is shorter in human Ig genes than in non-Ig genes but is much longer than in mouse orthologues. These findings provide a clear picture of non-neutral and nonrandom occurrence of the CpG islands in the PPRs of human and mouse Ig genes, which facilitate rational recommendations regarding their nomenclature.

Keywords: Human genome; Mouse genome; CpG density; CpG distribution pattern

1. Introduction

The characterization of promoters and their regulatory elements is one of the major challenges in bioinformatics and functional genomics (Fickett and Hatzigeorgiou, 1997). Different approaches have been developed to detect conserved motifs in different genes (Fickett and Hatzigeorgiou, 1997; Fickett and Wasserman, 2000). Although the in silico approaches seem promising, unambiguous identification of regulatory elements is not straightforward (Fickett and Hatzigeorgiou, 1997). One problem is that the in silico approaches limit the putative promoter regions (PPRs) to an arbitrary number of base pairs upstream of the gene

Abbreviations: PPR, putative promoter region; TSS, transcription start site; Chr, chromosome (Chr22 = chromosome 22); Ig, immunoglobulin; SW, starting window; O/E, observed/expected value; C, V, J, D, K and L, constant, variable, joining, diversity, kappa and lambda Ig genes; IGD, inter-gene distance.

* Corresponding authors. Guang B. Liu is to be contacted at Vision Touch and Hearing Research Centre, Faculty of Biomedical Sciences, The University of Queensland, Brisbane, Qld 4072, Australia. Tel.: +43 617 3365 4072; fax: +43 617 3365 4522. Kong-Nan Zhao, Centre for Immunology and Cancer Research, The University of Queensland, Princess Alexandra Hospital, Brisbane, Qld 4102, Australia. Tel.: +43 617 3240 5282; fax: +43 617 3240 5946.

E-mail addresses: g.liu@uq.edu.au (G.B. Liu), knzhao@cicr.uq.edu.au (K.-N. Zhao).

0378-1119/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.gene.2005.06.002

transcription start site (TSS). Ideally, this number should be chosen based on a functionally defined PPR because the length of the PPRs may differ considerably. Gene expression may be influenced not only by the regulatory sequences existed in the upstream and downstream flanking regions of a gene but also by the factors located in a remote distance (Shimada et al., 1989; Clegg et al., 1996). For example, the imprinting centre within human 15q11–q13 functions to co-regulate imprinted genes over a 2-Mb domain (Saitoh et al., 1996). X inactivation, as a regulatory mechanism related to gene expression, might influence a domain as far as >150 kb (Heard, 2004). Thus, to study the regulatory sequences of genes, a comprehensive analysis covering different numbers of base pairs upstream of the TSS of genes may be informative.

CpG islands are about 200-bp stretches of DNA that have a significantly higher concentration of CpG dinucleotides than the bulk of the genome (Davuluri et al., 2001; Ohlsson and Kanduri, 2002). CpG islands located in the gene PPRs play important roles in the reorganization of chromatin during mammalian spermiogenesis (Kundu and Rao, 1999) and in gene silencing during processes such as X-chromosome inactivation, imprinting, and silencing of intragenomic parasites (Takai and Jones, 2002). CpG islands are identified at the 5′ end of approximately 60% of human genes and so are important genomic landmarks (Cross et al., 2000). Study of the occurrence and characteristics of the CpG islands has gained great interests. Whole genome CpG island libraries have been prepared for human (Cross et al., 1994), chicken (McQueen et al., 1996), mouse (Cross et al., 1997) and pig (McQueen et al., 1997). These libraries provide a normalized set of sequences for the 5′ end of CpG island-associated genes. Studies using these libraries have revealed that, in each species, CpG islands are not randomly distributed but are concentrated in particular regions (Cross et al., 2000). However, the mechanism of the CpG island in the regulation of gene expression remains unclear (Antequera, 2003). Little is known about whether and how density and organization of the CpG islands in gene promoter regions are gene-specific in human genome.

In vertebrates, antibody responses are one of the two classes of the immune responses to protect them from infection by microorganisms and parasites. The antibody responses involve the production of antibodies, which are proteins called immunoglobulins. An immunoglobulin (Ig) protein consists of two light chains (called as kappa (K) and lambda (L) chains, respectively) and two heavy (H) chains (Klein, 1997). The synthesis of the three Ig chains involves multigene families from four gene groups: variable (V), diversity (D), joining (J) and constant (C), each one with unique characteristics (Giudicelli et al., 2005). Thus, the molecular genetics of the Ig genes is complex and unique in the genome of vertebrates (LeFranc and LeFranc, 2001). According to a recent report, the comprehensive IMGT (the international ImMunoGeneTics) genome database has contained 431 Ig genes from human and 459 from mouse, which represents the complete set of both human and mouse Ig genes for all three Ig loci: IgH, IgL and IgK (Giudicelli et al., 2005). The IgH locus contains four gene groups: IgHV, IgHD, IgHJ and IgHC, but the IgL locus contains only three gene groups: IgLV, IgLJ and IgLC, so does the IgK locus: IgKV, IgKJ and IgKC (Giudicelli et al., 2005). Most of the Ig genes for the three Ig loci that express the three Ig chains are located on three chromosomes of both human and mouse genome. In human, they are located on Chr2, Chr14 and Chr22, while in mouse on Chr6, Chr12 and Chr16. Most of the Ig genes on human Chr2 and mouse Chr6 express the kappa chain and on human Chr22 and mouse Chr16 produce the lambda chain, while majority of the Ig genes from the human Chr14 and mouse Chr12 encodes the Ig heavy chain. Recently, taking the in silico approach, we observed that CpG islands occur in the PPRs of the Ig genes on human Chr22 with very low frequency (Liu, unpublished data). Thus, we wonder if the Ig genes from human Chr2 and Chr14 also have a low frequency of the CpG islands occurred in their PPRs and whether the occurrence of the CpG islands in the PPRs is associated with the gene products (e.g., heavy, kappa and lambda chains of the Ig). As little is known about whether and how CpG islands occur in the PPRs of the human Ig genes, we carried out a comprehensive analysis for the CpG islands in the PPRs of these genes. Here we describe the density, distribution pattern and organization of the CpG islands in the PPRs of the human and mouse Ig genes. Our results reveal that the characteristics of CpG islands in the PPRs of the Ig genes are highly different from that of the non-Ig genes on Chr22 in human genome.

2. Materials and method

Both human and mouse gene-including DNA sequences were downloaded from the website of National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov/mapview/) on and before 23rd December 2004.

2.1. Sequence length

Each gene-including DNA sequence downloaded includes 5000 bp upstream from the transcription start site (TSS), the full length of the gene (extron+intron) and the 1000 bp downstream from the 3′ end of the gene (Liu et al., submitted for publication). If the inter-gene distance (the distance between the TSS of the gene being downloaded and the 3′ end of the upstream gene) is less than 5000 bp, then the downloaded DNA sequence includes the actual sequence between the TSS of the gene and the 3′ end of the upstream gene, the full length of the gene and 1000 bp downstream from the 3′ end of the gene. In each gene-including DNA sequence, the sequence upstream of the TSS is termed as a putative promoter region (PPR) and the