Available online at www.sciencedirect.com
ScienceDirect
Gene 410 (2008) 165–176
GENE
www.elsevier.com/locate/gene

DNA sequence and structural properties as predictors of human and mouse promoters

Pelin Akan, Panos Deloukas *

Wellcome Trust Sanger Institute, Hinxton CB10 1SA Cambridge, UK

Received 23 May 2007; received in revised form 30 November 2007; accepted 5 December 2007
Available online 23 December 2007
Received by O. Clay

Abstract

Promoters play a central role in gene regulation, yet our power to discriminate them from non-promoter sequences in higher eukaryotes is mainly restricted to those associated with CpG islands. Here, we examined *in silico* the promoters of 30,954 human and 18,083 mouse transcripts in the DBTSS database, to assess the impact of particular sequence and structural features (propeller twist, bendability and nucleosome positioning preference) on promoter classification and prediction. Our analysis showed that a stricter-than-traditional definition of CpG islands captures low and high CpG count promoter classes more accurately than the traditional one. We observed that both human and mouse promoter sequences are flexible with the exception of the TATA box and TSS, which are rigid regions irrespective of association with a CpG island. Therefore varying levels of structural flexibility in promoters may affect their accessibility to proteins, and hence their specificity. For all features investigated, averaged values across core promoters discriminated CpG island associated promoters from background, whereas the same did not hold for promoters without a CpG island. However, local changes around − 34 to − 23 (expected position of TATA box) and the TSS were informative in discriminating promoters (both classes) from non-promoter sequences. Additionally, we investigated ATG deserts and observed that they occur in all promoter sets except those with a TATA-box and without a CpG island in human. Interestingly, all mouse promoter sets showed ATG codon depletion irrespective of the presence of a TATA-box, possibly reflecting a weaker contribution to TSS specificity in mouse.

Keywords: Promoter; Human; Mouse; Genome-wide computational analysis; CpG island; TATA-box; DNA bendability; Propeller twist; Nucleosome positioning preference; ATG desert

1. Introduction

Promoters carry the central regulatory information of genes, therefore their in-depth characterization is vital to understand gene function. At present, there is an immense experimental effort for promoter identification and characterisation using techniques such as gene reporter assays, chromatin immunoprecipitation, oligo-capping, cap analysis of gene expression (CAGE) and 5′ end serial analysis of gene expression (SAGE) (Suzuki et al., 2001; Shiraki et al., 2003; Hashimoto et al., 2004; Kim et al., 2005; Cooper et al., 2006). Such methods generate powerful information towards understanding promoter sequence features as well as their activation mechanisms. One limitation though comes from having only a limited number of tools to mark inactive promoters. *Ab initio* prediction can be an alternative approach to locate promoter regions in a given genome sequence. Several algorithms have been designed to predict promoters and / or transcription start sites (TSSs) (reviewed in Pedersen et al., 1999; Hannenhalli and Levy, 2001; Maston et al., 2006). In a comparative study (Bajic et al., 2004) most prediction programs failed to operate at the genome scale, while predictors such as FirstEF (Davuluri et al., 2001), Eponine (Down and Hubbard, 2002) and CpGproD (Ponger

Abbreviations: TBP, TATA-box binding protein; TSS, transcription start site; CGI, CpG island; WCGI, associated with CpG island; WOCGI, not associated with CpG island; DBTSS, Database of Transcription Start Sites; R, purine; Y, pyrimidine; DPE, downstream promoter element; BRE, TFIIB recognition element; LCG, low CpG count; HCG, high CpG count.

* Corresponding author. Tel.: +44 1223 494717.
E-mail address: panos@sanger.ac.uk (P. Deloukas).

0378-1119/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.gene.2007.12.011

and Mouchiroud, 2002) performed reasonably well for certain promoter sets. Promoters associated with CpG islands (WCGI) were relatively easy to predict; most programs could predict ~ 80% of those promoters by making one false prediction for every true prediction (Davuluri et al., 2001). No program was able to predict promoters not associated with CpG islands (WOCGI) satisfactorily. For instance, FirstEF (Davuluri et al., 2001) predicted only 5% of such promoters while making 16 false predictions for every true prediction. WCGI promoters are associated with about 60% of human genes and often direct transcription of house-keeping or widely-expressed genes (Antequera, 2003). On the other hand, promoters not associated with CpG islands are generally tissue specific and often associated with a TATA box (Schug et al., 2005; Yamashita et al., 2005). Several studies investigated promoters in terms of base composition analysis, motif finding, comparative genome analysis and tissue-specificity profiles (Babenko et al., 1999; Louie et al., 2003; Aerts et al., 2004; Bajic et al., 2004; Schug et al., 2005). Overall, analysis showed that promoters are GC-rich, enriched with specific transcription binding site motifs depending on their tissue specificity profiles. Specific sequence motifs (TATA box, Initiator, BRE, DPE) within core promoters were found to be present in particular synergetic combinations, suggesting specific modes of transcription initiation depending on the presence and mutual positioning of such elements (Gershenzon and Ioshikhes, 2005). Several additional sequence motifs have also been found in subsets of promoters (FitzGerald et al., 2004; Marino-Ramirez et al., 2004; Xie et al., 2005; Yang et al., 2007).

The limited power of sequence-related information (i.e. base content, transcription factor binding sites, statistical properties of promoter sequences, and comparative genomics approaches) for deriving general signatures for promoters has motivated studies that employ structural properties of DNA in promoter prediction and classification (Pedersen et al., 1998; Gabrielian et al., 1999; Gardiner et al., 2003; Florquin et al., 2005; Kanhere and Bansal, 2005). Indeed, promoter DNA has to possess the required 3D structure to allow DNA-binding events and accommodate alterations in nucleosome positioning in order to allow the start of transcription. TATA boxes and initiator sequences have been found to comprise distinct flexible and rigid sequences compared to other parts of promoters (Pedersen et al., 1998; Babenko et al., 1999; Fukue et al., 2004; Fukue et al., 2005). Studies to measure promoter strength as a function of DNA flexibility using luciferase reporter assays showed that presence of rigid sequences around the expected position of the TATA-box had a positive influence on transcription (Fukue et al., 2004). However, it is important to note that structural properties of DNA are in part a result of the underlying nucleotide content.

Some promoters have been reported to exhibit depletion of ATG triplets around the TSS (Lee et al., 2005). This promoter subclass, called ATG deserts, is typically not associated with a TATA box, independently of the presence or absence of a CpG island (Lee et al., 2005). Lee et al. postulated that such promoters using multiple TSSs would still generate a single peptide starting with a methionine. This mechanism could be valuable especially for TATA-less promoters, which show a diverse distribution of TSS (Suzuki et al., 2001).

In this study, we examined the potential of three DNA structural properties for predicting WCGI and WOCGI promoters, namely propeller twist angle, bendability, and nucleosome positioning preference, and extended previous structural analyses of promoters. We analysed all human and mouse promoters from DBTSS and investigated the details of GC content change and CpG count in WCGI and WOCGI promoters. We also compared different definitions of CpG islands (CGI). Finally, we investigated ATG deserts to estimate their prevalence in the selected promoter sets and their association with presence of the TATA box in human and mouse promoters.

2. Materials and methods

2.1. Datasets

The datasets of human and mouse promoters in DBTSS (release 5.2.0; (Suzuki, 2001; Yamashita et al., 2006)) were used in this analysis. All 30,964 human promoters in DBTSS were used for analysis, but 940 of 19,023 mouse promoters were not included to the analysis because of repeats and low sequence quality, leaving 18,083 promoters for analysis. We also generated a human and mouse non-promoter sequence set each containing 20,000 random fragments of 1201 bp selected outside exonic regions to avoid the sequence bias of coding DNA.

2.2. CpG islands and CpG counts

The program ‘CpG island searcher’ (Takai and Jones, 2002) was used for identification of CGI, which were defined as regions greater than 500 bp in size, with a GC composition ≥ 55%, and an observed/expected CpG ratio of 0.65. The definition used here is more stringent than the original definition (200 bp in size, with a GC composition ≥ 50%, and an observed/expected CpG ratio of 0.60) (Gardiner-Garden and Frommer, 1987) in order to exclude GC-rich Alu-repetitive elements. Of the 30,964 human promoters, 12,229 (40%) were associated with a CGI. Of the 18,083 mouse promoters, 8415 were associated with a CGI. The original definition of CGI classified 17,456 (56%) human transcripts as associated with CGI.

Normalised CpG count was calculated as in (Saxonov et al., 2006), where observed CpG count is divided by expected CpG count calculated as $(\text{GC content}/2)^2$.

2.3. Positional weight matrix

For the analysis, promoter sequences were aligned relative to the TSS. A positional weight matrix was generated by calculating the frequency of each nucleotide for each position in the alignment. The nucleotide frequencies were then normalised by subtracting the expected frequency of each nucleotide at any random site in the human genome. The expected frequencies