Genomics 93 (2009) 314–322

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

Discovering sequences with potential regulatory characteristics

Minou Bina [a,*], Phillip Wyss [a], Sheryl A. Lazarus [a], Syed R. Shah [a], Wenhui Ren [b], Wojciech Szpankowski [b], Gregory E. Crawford [c], Sang P. Park [d], Xiaohui C. Song [d]

[a] Department of Chemistry, Purdue University, West Lafayette, IN 47907, USA
[b] Department of Computer Sciences, Purdue University, West Lafayette, IN 47907, USA
[c] Institute for Genome Sciences and Policy, Duke University, Durham, NC, 27708, USA
[d] Rosen Center for advanced computing, Purdue University, West Lafayette, IN 47907, USA

ARTICLE INFO

Article history:
Received 15 October 2007
Accepted 17 November 2008
Available online 30 December 2008

Keywords:
Human genome
Gene regulation
Regulatory signals
Codes in human DNA
Sequence context
Transcription factor binding sites
Genetic vocabulary

ABSTRACT

We developed a computational model to explore the hypothesis that regulatory instructions are context dependent and conveyed through specific 'codes' in human genomic DNA. We provide examples of correlation of computational predictions to reported mapped DNase I hypersensitive segments in the HOXA locus in human chromosome 7. The examples show that statistically significant 9-mers from promoter regions may occur in sequences near and upstream of transcription initiation sites, in intronic regions, and within intergenic regions. Additionally, a subset of 9-mers from coding sequences appears frequently, as clusters, in regulatory regions dispersed in noncoding regions in genomic DNA. The results suggest that the computational model has the potential of decoding regulatory instructions to discover candidate transcription factor binding sites and to discover candidate epigenetic signals that appear in both coding and regulatory regions of genes.

Introduction

A relatively long history supports the idea that genomic DNA represents a text, or a language, and that the order and the location of 'words' in that text would define the genetic information [1–3]. In fact, the determination of the genomic DNA sequences has brought linguistic metaphors to new heights: referring to DNA as a language and to the human genome as the 'book of life' [1,2]. Support for description of information in DNA as a text has emerged from the formulation of codons (words) for specifying the amino acid sequence of proteins [4,5].

Regulatory signals are generally assumed to not occur in the coding regions of genes. However, a theoretical model proposes that in addition to coding for proteins, the exons of genes may include information for other biologically meaningful signals such as binding sites for regulators of transcription [6,7].

A priori one could expect that regardless of their position in genomic DNA, regulatory signals may share common characteristics. For example, functional transcription factor binding sites have been localized not only near the beginning of genes but also in control regions within intronic sequences and in regulatory regions localized far upstream and far downstream of transcription start sites [8]. In addition to binding sites for transcription factors, regulatory segments might also include underlying signals for controlling other aspects of gene expression. For example, the arrangement of A-tracts in regulatory regions might provide signals for bending DNA to influence the three-dimensional architecture of the sequences in these regions of chromosomes [9]. Occurrences of CpG containing elements may provide epigenetic information and signals for methylation of DNA [10,11].

Several high-throughput procedures have been developed for mapping the position of regulatory regions of genes. Localization of DNase I hypersensitive (HS) sites in chromatin has emerged as a powerful experimental tool for mapping the regulatory regions that are poised for activation of gene expression [12–15]. This method has a long history and has withstood the test of extensive validations [14,15].

Computational models have also aimed at predicting the position of regulatory regions in genomic DNA. These models are often based on specific hypotheses. Examples include the hypothesis that clustering of transcription factor binding sites in a given region of genomic DNA reflects the presence of a regulatory segment: for example see [16,17], reviewed in [18]. Another model is based on the hypothesis that functional regulatory sequences could be subject to evolutionary selection, leaving a signature that could be detected in alignments of genomic DNA sequences from several species: for example see [19–21].

Abbreviations: CDSs, Coding sequences; DNase I, Deoxyribonuclease I; HS, Hypersensitive; HSSs, Hypersensitive sites; TFBS, Transcription factor binding sites; TSSs, Transcription start sites.

[*] Corresponding author. Fax: +1 765 494 0239.
E-mail address: Bina@Purdue.edu (M. Bina).

0888-7543/$ – see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.ygeno.2008.11.008

Our hypothesis is that the human genome has evolved to produce a well-defined “language” for conveying regulatory information in the DNA. To explore this idea, previously we examined characteristics of 9-mers collected from proximal-promoters of protein-coding genes [22,23]. Based on experimental data, we assumed that regulatory signals should occur frequently in regions preceding the TSSs. We chose 9-mers because they seemed to be a relatively “good” length for discovering the genomic context of regulatory signals including transcription factor binding sites [22]. We chose constant length DNA in order to reduce computational burden due to short sequences that appeared frequently in genomic DNA [22,24]. The computational model assumes that complementary 9-mers are equivalent. This assumption is based on studies showing that TFBSs can exert control on gene expression irrespective of their orientation in DNA: for example see [25–27].

In this report, we present a computational model (weighted density plots) for identifying the genomic DNA regions that include statistically significant occurrences of 9-mers collected from promoter and coding regions of human genes. We find that, in these plots, specific peaks often correlate with experimentally mapped regulatory regions in genomic DNA.

Results

Sampling of characteristics of 9-mers from human promoter and coding sequences

We determined the frequency of occurrences of complementary 9-mers in three groups of human DNA: a set of promoters, defined with respect to the 5′ end of ESTs [22]; coding regions in cDNAs obtained from GenBank [28]; and sequences corresponding to a draft of total genomic DNA [22]. The initial goal was to use 9-mers from CDSs as a control for 9-mers derived from the promoter regions. However, as detailed in subsequent sections, unexpectedly we found that a subset of 9-mers from CDSs also appeared frequently in regulatory regions of genes.

Furthermore, previous methods for identifying over-represented n-mers for *de novo* pattern detection [29] have not addressed the problem of sequences that appear frequently in genomic DNA. To resolve this problem, we normalized the frequency of the 9-mers in promoter regions and in coding sequences with respect to their corresponding occurrences in total genomic DNA [22].

Ranking of frequencies provides a statistical measure of the relative abundance of 9-mers in promoters or CDSs, with respect to their corresponding occurrences in total genomic DNA. For example, in the statistical scheme, ranking of 1 corresponds to those 9-mers that appear equally in promoters and total genomic DNA. Thus, rankings greater than 1 statistically could be significant: a ranking of 3.08 had a p (or β) value of about 10^{-27}; a ranking of 7 had a β value of about 10^{-50} [22].

The computational model uses the collected 9-mers to produce weighted density plots to predict the position of potential regulatory signals in human genomic DNA. The program employs a specified window to scan the human genomic DNA. The program examines all possible 9-mers in each window and then finds their computed ranks. The program uses the ranks to determine a weighted sum. As the window slides along a genomic DNA, the weighted sums would produce intensity values at each nucleotide position.

In addition to ranks, we wished to apply specific criteria to distinguish the 9-mers that occurred preferentially in non-coding regions from those that appeared frequently in CDSs. Towards this goal, we created three types of density plots (Fig. 1). We constructed a plot (CDS_Hits) to view the weighted density of matched of genomic DNA sequences with 9-mers collected from coding sequences. The ranking procedure excluded 9-mers that appeared frequently in genomic DNA. We imposed specific criteria (see discussion and methods) to construct a plot (Reg_Signal Pred1) to display the weighted density of matches of genomic DNA sequences with 9-mers collected from the promoter regions of genes. Additional criteria (see methods) were imposed to construct a plot (Reg_Signal Pred2) to distinguish the 9-mers with “high” non-coding context from those with a relatively high coding context. In that plot, intensities greater than one reflect normalized values of 9-mers that appear more frequently in promoters than in coding regions of genes.

Analysis of the HOXA locus on human chromosome 7

To evaluate the computational model, we analyzed several relatively long genomic DNA segments selected to include many genes. As an example, we highlight the results obtained for the HOXA cluster of genes on human chromosome 7. The cluster is relatively long and includes HOXA1, HOXA2, HOXA3, HOXA4, HOXA5, HOXA6, HOXA7, HOXA9, HOXA11, HOXA13, and EVX1 (Supplemental Fig. 1). We evaluate the intensities of predicted signals in the context of genomic positions of mapped DNase I HS segments [15,30,31].

Hypersensitivity to DNase I provides a relatively robust measure of the chromosomal regions that have an “open” chromatin structure [13]. A relatively large body of experimental data indicates that the DNase I HS segments in genomic DNA are either nucleosome-free or contain modified nucleosomal structures [10,14]. Accessibility of these segments is thought to expose the control signals in the DNA, for recognition by the regulators of gene expression [14]. Evidence indicates that the results of high-throughput methods are likely to be accurate since the methods have correctly identified the HS segments mapped by conventional techniques [30].

To compare the position of DNase I HS segments to the predictions of the computational model, we display the results in custom tracks in the genome browser at UCSC. Initially, we will qualitatively compare the position of the peaks in the density plots to the mapped HS segments: Fig. 1, tracks appearing under Duke/NHGRI DNase I-hypersensitivity and under UW/Regulome QCP DNase I Sensitivity [15]. A subsequent section provides statistical evidence for the correlations.

Supplemental Fig. 1 gives an overview of the position of HS segments and the organization of the genes in the HOXA locus. We analyzed the entire locus, using a sliding window of 30 bp. Results show that peaks in the weighted density plots are primarily localized within the gene-rich segments (Supplemental Fig. 1). In the custom tracks, intensity of peaks is displayed as pixilated bars, in order to produce condensed plots (for example, see Fig. 1). Fig. 2 shows an example of full-display of density plots with respect to landmarks including potential TFBSs.

Fig. 1 shows an expanded view of the genomic DNA region that includes HOXA1. Three custom tracks display the predictions. The displayed predictions provide typical results showing that statistically ranked 9-mers from promoter regions may appear as clusters in the vicinity of transcription start sites and in sequences further upstream (Fig. 1, track labeled Reg_Signal Pred1). Predictions also provide typical results showing that statistically ranked 9-mers from CDSs appear to cluster not only in exonic regions but also in non-coding sequences: in that example, upstream of TSSs of HOXA1 (Fig. 1, track labeled CDS_Hits).

A comparison of the tracks shown in Fig. 1 reveals correspondence of the positions of predicted regulatory signals to the experimentally determined DNase I HS segments that include the 5′ end of HOXA1. Also, there is a correspondence between predicted regulatory signals in a region upstream of the transcription start site (~3700 bp) and HS segments in chromatin isolated from several cell lines (Fig. 1, tracks labeled GM069, CD4, HeLa, and $CaCO_2$). Fig. 1 shows that the mapped DNase I HS segments may include the transcribed untranslated region and the coding region of a gene.

Supporting information gives additional examples of correlation of predicted regulatory signals within DNase I HS segments in both