NXSensor web tool for evaluating DNA for nucleosome exclusion sequences and accessibility to binding factors

Peter Luykx, Ivan V. Bajić1 and Sawsan Khuri2,*

Department of Biology, University of Miami, Coral Gables, FL 33124, USA, 1School of Engineering Science, Simon Fraser University, Burnaby, B.C. V5A 1S6, Canada and 2The Dr. John T. Macdonald Foundation Center for Medical Genetics, University of Miami Miller School of Medicine, Miami, FL 33101 USA

Received February 10, 2006; Revised March 7, 2006; Accepted March 20, 2006

ABSTRACT

Nucleosomes, a basic structural unit of eukaryotic chromatin, play a significant role in regulating gene expression. We have developed a web tool based on DNA sequences known from empirical and theoretical studies to influence DNA bending and flexibility, and to exclude nucleosomes. NXSensor (available at http://www.sfu.ca/~ibajic/NXSensor/) finds nucleosome exclusion sequences, evaluates their length and spacing, and computes an ‘accessibility score’ giving the proportion of base pairs likely to be nucleosome-free. Application of NXSensor to the promoter regions of housekeeping (HK) genes and those of tissue-specific (TS) genes revealed a significant difference between the two classes of gene, the former being significantly more open, particularly near transcription start sites (TSSs). NXSensor should be a useful tool in assessing the likelihood of nucleosome formation in regions involved in gene regulation and other aspects of chromatin function.

INTRODUCTION

Transcription factors, enhancer-binding proteins and chromatin-remodeling factors all play a role in regulating gene expression. In addition, nucleosomes, the basic structural units of chromatin, are now thought to also be involved in this process ([1,2]; see also (3,4)]. The dissociation or displacement of nucleosomes from DNA by the action of other protein factors, such as DNA- and histone-modifying enzymes (methylases, kinases, acetylases and deacetylases) likely gives the basal transcription apparatus access to the promoter regions of genes (5–7). A means for assessing the distribution of nucleosomes along stretches of regulatory DNA would therefore be a useful addition to our understanding of gene regulation.

BACKGROUND AND DESCRIPTION OF NXSensor

DNA sequences chosen as NXSs were the following, based on experimental observations (14–16):

\[
[(G/C)_{3}N_{2}]_{3} \text{; e.g. GGCAACGCTTGGGTA, } \text{ GGCCGCGCAGGGGCT}
\]

\[
A_{10} \text{; e.g. AAAAAAAAA, TTTTTTTTTT}
\]

Other studies lend support to the unfavorability of these sequences for nucleosome formation because...
of their hindering of DNA bending or flexibility (12,13,17).

A NXS was defined as one of the DNA sequences above, a contiguous non-overlapping DNA sequence long enough to comprise one full turn of the DNA double helix. The NXSensor program annotates individual DNA regions by marking NXSs and then examining their spacing to find sequences between NXSs that are <147 bp in length, on the assumption that nucleosomes are unlikely to form in such regions. The remaining segments of DNA are those where nucleosomes may be located.

VALIDATION AND APPLICATION

To test the general validity of the NXSs chosen, and to test the ability of the NXSensor program to find and annotate such sequences, the DNA sequences of positioned nucleosomes listed in the nucleosome positioning region database (NPRD) were examined (18). Ideally, in these nucleosome-associated DNA sequences there should be no NXSs as defined here. We found that only nine (8%) of the 112 positioned nucleosomes contained any NXSs. These sequences (a total of 12 NXSs) comprised 199 bp in a total of 16 829 nucleosome-associated base pairs (1.18%), which represents reasonable agreement with expectation. Exceptions in which a NXS is accommodated within a nucleosome may be the result of the presence of additional protein factors associated with nucleosomal DNA.

As a further test of NXSensor, the SV40 viral sequence was examined because it is known to contain a nucleosome-free segment of about 400–500 bp within its 5243 bp genome (19–21). The results showed a close correspondence between the experimentally-determined location of the nucleosome-free region and the nucleosome-free region predicted by NXSensor’s analysis of DNA sequence (Supplementary Figure 1).

In a final comprehensive test, NXSensor was used to analyze the promoter regions of two sets of human genes, housekeeping (HK) and tissue-specific (TS) genes. These two sets of genes were chosen in order to assess the potential distribution of nucleosomes in promoters of genes that are used differently in different tissues, to substantiate the conclusions of other researchers who have recently investigated this question using different approaches, and to demonstrate further the usefulness of the NXSensor tool.

One hundred genes from each set, HK and TS, were selected on the basis of the tissue expression patterns given in the Novartis Research Foundation’s Genomic Institute ‘SymAtlas’ (http://symatlas.gnf.org/) (22). The genes selected were at the two extremes of tissue expression patterns: HK genes showing significant expression in all 73 normal tissues of the SymAtlas, and TS genes showing, according to gene-specific non-cross-hybridizing probe sets (those with ‘_at’ suffixes), significant expression in only one or two tissues. To avoid possible bias in selecting HK genes, we included as wide a variety of basic cell functions as possible, and in selecting TS genes, as wide a variety of cell- and tissue types as possible. Genes with only one region of transcription initiation were used instead of those with widely-spaced alternative transcription start sites (TSSs), as indicated in the UCSC Genome Browser (http://genome.ucsc.edu/), and double-checked in the Database of Transcriptional Start Sites (DBTSS, http://dbtss.hgc.jp/). The list of genes is available at http://www.sfu.ca/~ibajic/NXSensor/200genes.xls.

To examine the region around the TSS of each gene, we used the sequence from 1000 bases upstream (which would include the complete promoter of most genes) to 1000 bases downstream from the TSS (−1000 to +1000), and called this 2000 bp stretch the ‘promoter region,’ centered on the TSS. The TSS was either the predominant site in, or a site near the middle of, the cluster of oligo-capped cDNA clones shown in the DBTSS (23). In cases where no oligo-capped cDNA clones were available from the DBTSS (21 of 100 HK genes, 36 of 100 TS genes), the RefSeq (NCBI) annotated TSSs were used. All sequences tested were datamined from the UCSC Genome Browser.

The basic ‘nucleosome segments’ option of NXSensor annotates the submitted sequence by highlighting in gray the segments available for nucleosome binding, leaving open the segments unlikely to be bound to nucleosomes by virtue of the number and positions of NXSs (Figure 1 and Supplementary Figure 1). The annotated promoter region sequences were used to construct two basic measures. The ‘NXScore’ defined as the number of NXSs in a window of 147 bp at a given position within the 2000 base promoter region, estimates the likelihood that nucleosomes are excluded from promoter regions at different distances from the TSS. This measure is the obverse of the ‘nucleosome formation potential’ based on nucleosome binding sequences used by Levitsky et al. (1,24). The NXScore is averaged over all promoter regions of each class of genes, HK and TS, to generate a graph of average NXScore versus position, centered on a given window (Figure 2).

The second measure was the ‘accessibility score,’ measuring the overall accessibility of the 2000 base promoter region to protein factors. The accessibility score is defined as the proportion of base pairs in the region likely to be free of nucleosomes, according to the number and spacing of the NXSs. Accessibility scores can vary from 0.0 (all sequences potentially occupied by nucleosomes) to 1.0 (no sequences likely to be occupied by nucleosomes). In calculating the accessibility score ACC(10), short exclusion sequences of 10 bp or less flanked by two segments having the potential for nucleosome binding were not considered ‘open’ because of the unlikelihood that any protein would be able to bind effectively such a short sequence between two nucleosomes. Accessibility scores for two examples are given in Figure 1.

A direct comparison of the average number of NXSs, per window of 147 bases, in HK and TS promoter regions is shown in Figure 2. In both sets of promoter regions there were more NXSs nearer the TSS. There was a significant difference between HK-gene promoters and TS-gene promoters in the mean number of NXSs in the specific 147 bp window centered on the TSS (HK, 2.10 and TS, 0.82; P < 0.01, Kolmogorov–Smirnov test). Supplementary Figure 2 gives an illustrative example of 10 000 bases around the TSS of a HK gene, in which NXSensor analysis suggests that essentially only the TSS region is free of nucleosomes.

Figure 2 also shows that the mean number of NXSs in this window for the HK genes was significantly higher than that expected for random-sequence DNA with the same base composition as the HK-gene promoter regions (2.10 for HK
genes, 0.53 for the corresponding random-sequence DNA, \( P = 0.0014 \). In contrast, the mean number of NXSs in this window for the TS genes was hardly different from the NXSs expected in random-sequence DNA having the TS promoter base composition (0.82 for TS genes, 0.22 for the corresponding random-sequence DNA, \( P = 0.06 \)).

From NXSensor output we obtained accessibility scores, ACC(10), for the promoter regions in our dataset. As shown in Figure 3A, there was a significant difference between the distributions of accessibility scores of HK genes and of TS genes, the HK genes having significantly higher accessibility scores than the TS genes (\( P < 0.01 \), Kolmogorov–Smirnov test).

Further analysis was conducted using a modified accessibility score, ACC(50), which includes sequences at least 50 bp in length likely to be nucleosome-free. This is the length of sequence estimated to be occupied by the basal transcription apparatus (25,26). ACC(50) was examined for the DNA region from \(-110\) to \(+60\), the region in which the basal transcription apparatus would have to bind in order to initiate transcription at the majority of TSSs clustered around the modal or central TSS at \(+1\) (27). The comparison of this TSS region (Figure 3B) showed that 58% of the HK genes, but only 22% of the TS genes, had an ACC(50) of 0.5 or higher. Of the HK genes, 45% had an ACC(50) of 1.0, while only 15% of the TS genes did. The overall distributions of HK gene and TS gene ACC(50) scores were significantly different by the Kolmogorov–Smirnov test (\( P < 0.01 \)).

**DISCUSSION**

Work in recent years has shown that the nucleosomes are neither uniformly nor randomly positioned along the DNA double helix. In a number of instances the location of nucleosomes in control regions of DNA was found to be an important feature of gene regulation. The validity of this statement was recognized explicitly by Levitsky et al. (1), and is the basis for the nucleosome positioning region database (NPRD), (http://srs6.bionet.nsc.ru/srs6/) (18). The ‘nucleosome formation potential’ scores have recently been used to demonstrate differences in the intergenic regions (28) and more specifically in the promoter regions (29) of human HK and TS genes.
We have taken a different approach by utilizing a combination of published experimental and observational data to identify DNA sequences unfavorable for nucleosome formation because of their influence on DNA bending and flexibility. Other sequences not used here may also be unfavorable for nucleosome formation. One example is TGGA repeats (30), but these occurred too infrequently in our sample of 200 promoter regions to have any influence on the results.

Using our NXSensor program (http://www.sfu.ca/~ibajic/NXSensor/), we have shown that, as expected, these NXSs are rare in DNA sequences occupied by nucleosomes listed in the NPRD. NXSensor also accurately predicts the location of the nucleosome-free zone in viral SV40 DNA. In a set of HK and TS promoter region DNA sequences, NXSensor analysis has demonstrated that the promoter regions of HK genes are likely to be relatively nucleosome-free, a finding that may help to explain the wide tissue distribution of their expression. This is in contrast to those of TS genes, whose promoter regions contain sequences more favorable for nucleosome formation (28,29), and which are therefore likely to require additional TS transcription factors to modify or displace nucleosomes before the genes can be expressed.

The promoter regions of the two sets of genes differed in base composition and in how many of them had CpG islands (31,32). The GC content of the promoter regions of HK genes was 56.86%, compared with 50.44% for TS genes. CpG islands were characteristic of 92 of 100 HK genes, but only 19 of 100 TS genes, in agreement with the observations of others (32). The greater frequency of NXSs of the [(G/C)₃N₂]₃ type in HK-gene promoter regions compared with TS-gene promoter regions is partly a reflection of the higher GC content of the former (56.86% G + C in HK, 50.44% G + C in TS). However, the average number of [(G/C)₃N₂]₃ sequences in HK-gene promoter regions, 15.4, was even higher than the 7.8 ± 2*2.7 (S.D.) expected in random sequences of the same length and base composition (P < 0.05). In addition, in spite of their higher GC content, HK-gene promoter regions had almost as many nucleosome-unfriendly polyA and polyT tracts as did the TS-gene promoter regions (50 in HK, as compared with 52 in TS). These two observations show that the frequency of both types of NXS in HK-gene promoters is higher than would be expected from their base composition, suggesting a functional significance, possibly related to nucleosome exclusion.

Promoter regions of HK genes are typically GC-rich. Presumably this is the consequence of selection pressure for certain kinds of promoter sequences. It may be speculated that one component of such selection is for sequences that tend to exclude nucleosomes. Another possible component
is selection for GC-rich sequences that bind certain ubiquitous transcription factors, such as Sp-1. These two components of selection might be related, in that the factors that regulate expression of HK genes may have evolved to bind to nucleosome-free regions.

We were unable to find any correlation between accessibility score and expression levels of these genes as given in SymAtlas (22) (see Supplementary Figure 3). It is likely that the lower nucleosome occupancy of promoter regions of HK genes is related not to their expression levels but primarily to their ubiquity of expression in different tissues.

The NXSensor web tool is flexible enough to allow for different settings of regions likely to be free of nucleosomes and more accessible to other protein factors. NXSensor can be set for more stringent exclusion criteria by increasing the number of tandem NXSs required for nucleosome exclusion. The minimum length of ‘nucleosome-free’ sequence can be increased to accommodate the space required for larger protein complexes, an approach we took to show that the region immediately surrounding TSSs of HK genes is likely to be more accessible to the basal transcription apparatus than is the corresponding region of TS genes.

Here we have applied NXSensor to the promoter regions of individual genes and classes of genes. NXSensor may also be used to investigate other control regions farther from coding sequences, such as enhancer and inhibitor regions, as well as sites of methylation, imprinting, recombination, repair, pre-mRNA splicing and indeed any DNA sequences where nucleosome location is likely to be a factor in overall chromatin organization and function (33–38).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR online.

ACKNOWLEDGEMENTS

The authors thank Steven Green (Biology Department, University of Miami) for statistical advice, John Walker and Andrew Su (Genomics Institute of the Novartis Research Foundation) for advice on the interpretation of SymAtlas gene expression data, Terace Fletcher (Department of Biochemistry and Molecular Biology, University of Miami) for helpful discussions, and anonymous reviewers for many cogent comments. Part of the work performed while I.V.B. was with the Electrical and Computer Engineering Department, University of Miami, Coral Gables, Florida 33124, USA. Funding to pay the Open Access publication charges for this article was provided by the Dr. John T. Macdonald Foundation Center for Medical Genetics.

Conflict of interest statement. None declared.

REFERENCES