CpGP Dynamics – The Dynamics of CpG Island and Promoter to Validate Nucleosomal Gene Expression

S. Prasanth Kumar

Department of Bioinformatics, Alagappa University, Karaikudi- 630003, India
prasanthbioinformatics@gmail.com

Abstract

Computational prediction of nucleosome positioning relies upon in vitro and in vivo experimental outcome such as sequence positioning and exclusion signatures, structural thermodynamic details, histone-DNA interaction models, etc. On the other hand, CpG island and promoter prediction programs are available which depends upon the algorithm built by the predictive power of trained experimental datasets from sequencing projects. "CpGP dynamics – The dynamics of CpG island and promoter to validate nucleosomal gene expression" is a web-based program which predicts the nucleosome- positioning (NP) and exclusion (NE) signatures in the user provided nucleotide sequence and presents a graphical output. It also utilizes the sequence positions of CpG island and promoter predicted by third-party programs as input to generate graphical sequence output. These two graphical outputs can be merged to discriminate the more accurate sequence positions of CpG island and promoter from a number of likelihood predictions. The program is freely accessible at http://www.cpgpdynamics.webs.com.

Keywords: *Nucleosome Positioning and Exclusion, CpG island, Promoter, Bioinformatics*

1. Introduction

Eukaryotic genomic DNA is packaged into highly compacted nucleosome arrays known as chromatin. A 147 base pair (bp) stretch of DNA wraps around the histone protein octamer making up a nucleosome and such interacting DNA are referred as nucleosomal DNA [1]. Neighbouring nucleosomes are connected by about 10-50 bp of DNA called linker DNA. The tendency of nucleosomal DNA to interact with histone is highly dependent on specific DNA sequence which renders them to bend sharply every helical repeat [2]. Likewise, it also occludes them in order to interact with DNA binding proteins such as polymerase (Pol), transcription factors (TFs), regulatory, repair and recombination complexes [3]. Numerous i*n vitro* and *in vivo* experiments confirmed these intrinsic sequence preferences and helped to decipher "genomic code for nucleosome positioning and exclusion" and have been implicated in gene expression studies [4]. Bioinformatic tools are currently available which predicts the likelihood of DNA being nucleosome positioning (occupancy) and/or occlusion (exclusion). Predictions are made from position weight matrices which take into account the periodic patterns of dinculeotides derived from about 200 nucleosomal DNA sequences [4], identification of binding sites of known transcription factors [5], model derived from nucleosome formation energetics studied in high throughput sequencing maps [6], probe on DNA bendability matrix of *C. elegans* [7], duration Hidden Markov Model (HMM) in which linker DNA discrimination was modeled [8], degree of DNA flexibility [9], symmetry of curvature of a DNA sequence [10], etc.

Many statistical measures have been applied to detect promoters in the DNA sequence such as trained time-delay neural network [11], decision tree consisting of a set of quadratic discriminant functions [12], combined approach of genetic algorithms and neural networks [13], scoring homologies with putative eukaryotic Pol II promoter sequences [14], etc whereas the combination of base composition with the statistical descriptor plays a significant role to identify CpG islands. Gardiner and Frommer sequence criterion to classify CpG islands depicts that a genomic region that complies with three conditions (i). GC content above 50%, (ii). Ratio of observed-to-expected number (CpG o/e) of CpG dinucleotides above 0.6, and (iii). Length greater than 200 bp; can be considered as a potential CpG island [15].

The present work deals about the development of a program with a web-interface named as "CpGP dynamics – The dynamics of CpG island and promoter to validate nucleosomal gene expression" which predicts the nucleosome- positioning (NP) and exclusion (NE) signatures derived from literatures in the user provided DNA sequence. It also considers the sequence positions of CpG island and promoter regions as input to generate a graphical sequence output. If both the graphical results are compared, it will help the user to identify the most accurate predictions of sequence positions corresponding with CpG island and promoter as the concerned programs provides the users with a list of likelihood predictions for a DNA sequence. The tactics behind the graphical results comparison is explained in details with an example predicted by the present program in the Results and discussion section. The program's main objective is to consider the NP and NE signals in identifying the CpG island and promoter regions embedded in the gene sequence from a number of predictions provided with a range of probabilities which are only based on models developed from known experimental datasets. CpGP dynamics is freely accessible over World Wide Web at http://www.cpgpdynamics.webs.com.

2. Materials and Methods

2.1 Extraction of NP and NE from Bibliographic Literatures

Two NP signatures were used to identify the regions of DNA prefer to wrap around nucleosomes computationally. The motif, $(A/T)_3NN(G/C)_3NN$ (where N = A/T/G/C as applicable) was proven to be superior in nucleosome formation as disclosed in the *in vitro* experiments [16]. *In vitro* investigation on nucleosome-DNA interaction model showed that a distinctive sequence motif which recurs periodically at the helical repeat (~10 bp) is due to sharp bending of DNA across nucleosomes. This motif contains AA/TT/TA dinculeotides which are 10 bp apart. Additionally, it consists of a GC dinucleotide centrally positioned [4]. Combining these information, another motif, $GCNN(A/T)_2NNGC$ favoring nucleosomal wrapping was developed. However, genomic regions with high GC/CG dinucleotides density will tend to accommodate in CpG island as per Gardiner and Frommer sequence criterion.

Four NE signatures were utilized to recognize the regions of DNA sequence having nucleosome occlusion preferences. A DNA repeat, $(G/C)_3NN(G/C)_3NN(G/C)_3NN$ was shown to avert NP and act as NE sequence motif [17, 18]. Other three motifs, homopolymers of A and T nucleotides and poly (dA:dT) tract are known to constitute unusual- structural, dynamic, and mechanical properties, and also resist sharp bending which makes them inappropriate for stable DNA-histone interaction [19, 20]. NXSensor, a web-based tool which predicts the later exclusion signals having equal to- or greater than 10 motif length [9]. According to Dechering *et al*, 1998 studies on distinct frequency-distributions of homopolymeric DNA tracts in different genomes, they proposed a length of 10-20 bp or even greater [21]. For example, i*n vivo* and *in vitro* studies on nucleosome occupancy showed that AAAAA (5-mer) was reported with the lowest occupancy [22]. Besides, AT-rich oligomers were often observed in non-nucleosomal DNA of several organisms [18]. Hence, the NE signature relating to above 3 motifs was extended to 8 bp as minimum length and no limit over its maximum length.

2.2 Computational Development of CpGP Dynamics

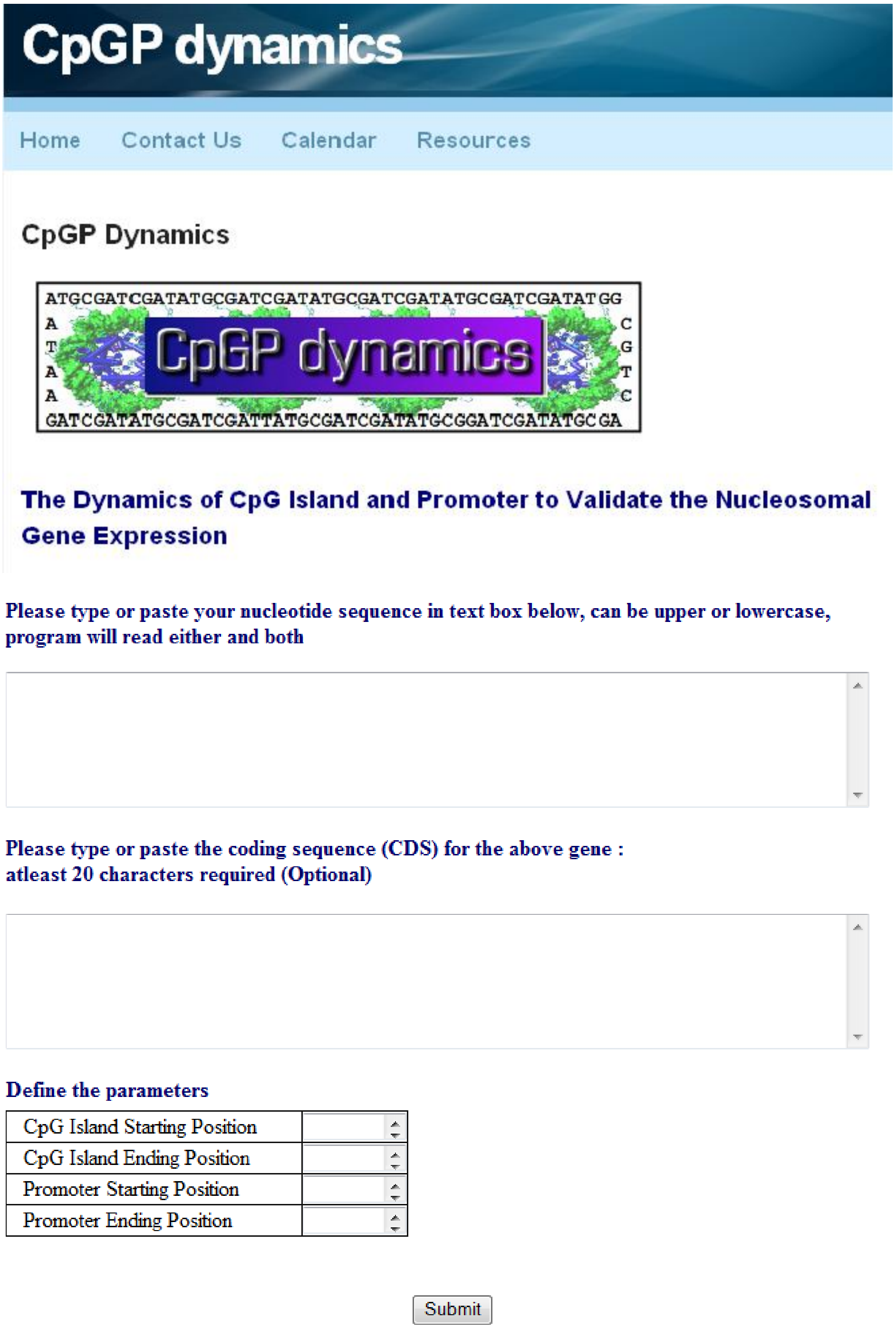
The above mentioned NP and NE signatures were coded in the form of regular expression. Due care was taken to meet the standards of scripting languages in order to run in Java Script-enabled web browsers [23]. Program's compatibility was checked with Microsoft Internet Explorer 7, Mozilla Firefox 3.6 and Google Chrome. The program page can be saved in personal computers and can customize the coding using text editor/website creator (access/modification to the program should be complied with the standards of Creative Common (CC) attribution). The program is hosted at http://www.cpgpdynamics.webs.com.

3. Results and Discussion

3.1 Web Components of CpGP Dynamics

The web-interface of CpGP dynamics (Figure 1) has two types of inputs: text and numerical. Text based fields considers the input of nucleotide sequence and CDS (coding sequence). Numerical fields take into account the sequence positions of CpG island and promoter region predicted by third-party programs. The nucleotide sequence is essential for running the program and it is the only mandatory field requires to be provided whereas other fields are optional depending upon the context of usage (Figure 2). Users can directly access the web-interface using a Java Script enabled web browser. If required, the webpage can be downloaded in their computers to run the program. The program was built using JavaScript scripting language and can be easily modified and optimized.

Besides, CpGP dynamics have an option to provide the sequence of CDS. The start of the CDS may correspond to translational start site but gives no clue about the localization of promoter. If the user provides CDS (atleast 20 bp) then the program will give two insights. First, the CDS start site in the gene sequence which helps the user to identify how many bps of DNA is upstream to CDS start site i.e. 3' untranslated region (UTR) plus promoter and regulatory regions. For example, the length of the nucleotide sequence is 1000 bp, the CDS start sequence is CATGCGGCATCGTTAGCCAT, the program will search for the specified CDS in the nucleotide sequence and reports the CDS start site in the gene sequence, say 699. Hence, it can be concluded that the given gene sequence has a 698 bp length 3' UTR along with either whole or part of the (overlapping) promoter and regulatory regions. Second, if the user had specified the starting and ending position of promoter, in that case, the distance between promoter end site and CDS start site can be studied. If a positive value comes out, then the promoter and coding sequence is not overlapping and if the vice versa condition prevails, then the promoter and the coding sequence is overlapping, thereby three outcomes are possible. The promoter prediction may be inaccurate and/or the predicted promoter region (defined in CpGP dynamics) has a high sequence range and/or experimentally proven otherwise. For instance in the above example, if the user specifies the promoter end site as 512, the program will search for specified CDS (as described above) and reports the distance between the promoter end site and CDS start site as 187 (positive value). It means that a DNA stretch of 187 bp is found between promoter and coding region representing both the genomic elements are spaced adequately. Lest, if the user defined 710 as promoter end site, then this descriptor will give a value of -11 (negative value). Thus, sketches out a scenario of overlapping promoter and coding sequence. The algorithm was developed to search for first instance of 'ATG' codon (translation start site) if the user does not mention about the CDS but only defines the numerical parameters. It also generates a graphical overlay of ATG codon found about 500 bp (approximately) downstream of the promoter end site.


CpGP dynamics
Home
Contact Us
Calendar
Resources
ATGCGATCGATATGCGATCGATATGCGATCGATATGCGATCGATATGG
A T A A
CpGP dynamics
C G T C
GATCGATATGCGATCGATTATGCGATCGATATGCGGATCGATATGCGA
Submit

CpGP Dynamics

The Dynamics of CpG Island and Promoter to Validate the Nucleosomal Gene Expression

Please type or paste your nucleotide sequence in text box below, can be upper or lowercase, program will read either and both

Please type or paste the coding sequence (CDS) for the above gene : atleast 20 characters required (Optional)

Define the parameters

CpG Island Starting Position	
CpG Island Ending Position	
Promoter Starting Position	
Promoter Ending Position	

Figure 1. Webshot of CpGP Dynamics Program

3.2 The Background of Graphical Output Comparison

The strategy behind the graphical results comparison can be put forth as follow: (i). NP in promoters can be categorized into two types: (a). Open promoters have a long nucleosome depleted region (NDR; ~150 bp) upstream of transcription start site (TSS) and these promoters are generally TATA-less suggesting the presence of NDRs in the genomic regions around TSS to facilitate recognition and binding of promoter sequences by the TFs for transcription [24]. Hence, NE signatures can be found in these regions (Figure 3 A), (b). Occupied (closed) promoters typically have TATA element and are found to be either partly or almost fully occupied by nucleosomes. Notably, the TATA box is found to be localized at the nucleosome edge (~25-125 bp from TSS) and contain a gradient of increasing nucleosome occupancy downstream of the TSS [24]. Therefore, if the DNA sequence consists of a TATA box, then, the presence of NE in the regions around TATA box and/or TSS will be less and there is an equal chance of finding a number of NP signatures downstream to it (Figure 3 B), (ii). Daenan *et al*, 2008 presented an intriguing hypothesis that nucleosomes position themselves (*in vivo*) making the cognate sites for TFs exposed through a relaxed, open chromatin structure whereas cryptic TFs binding sites scattered throughout the genome appears to be masked and wrapped around nucleosome particles [5], (iii). Genomic regions corresponding to TATA box was reported with a nucleosome occupancy having a probability of only about ~0.5 (this probability value could not be used as a good discriminator of being positioned or excluded from nucleosomes) as revealed in array intensity and nucleosome calls measured by Whitehouse *et al*, 2007 [25] and Lee *et al*, 2007 [26], respectively. Additionally, *in vivo* studies conducted by Segal *et al*, 2006 in yeast genome indicated that TATA box is found outside a stably positioned nucleosome [4]. In the context of point number (ii) and (iii), there is an equal possibility of finding a NE signature just upstream of TSS and/or a NE

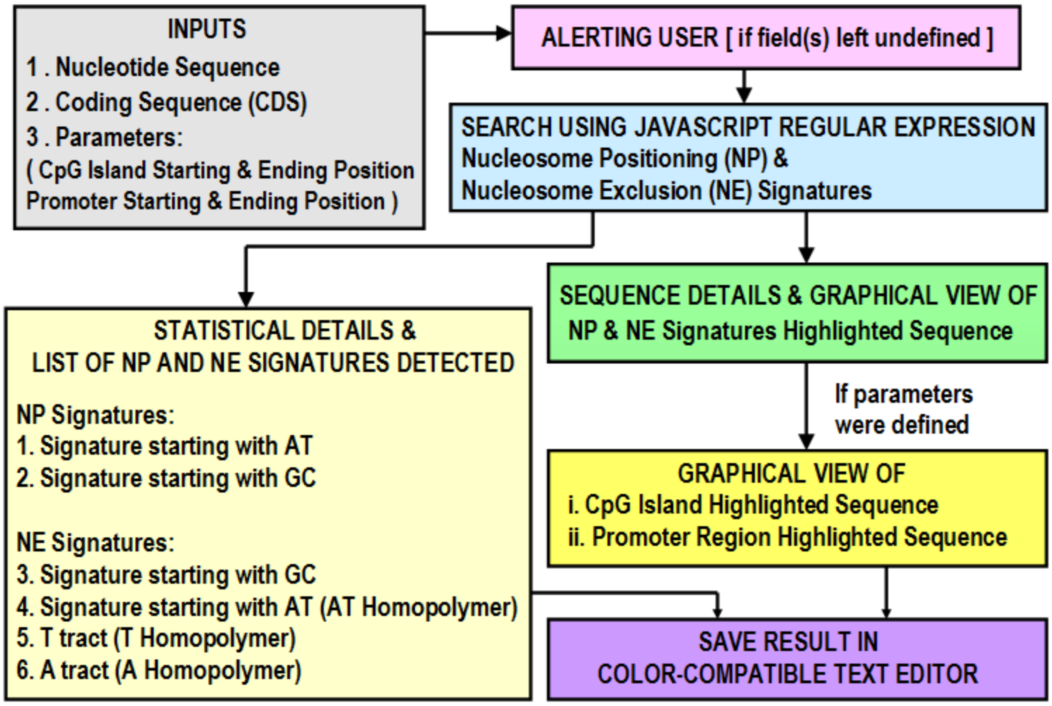

INPUTS
1 . Nucleotide Sequence
2 . Coding Sequence (CDS)
3 . Parameters:
(CpG Island Starting & Ending Position
Promoter Starting & Ending Position)
ALERTING USER [if field(s) left undefined]
SEARCH USING JAVASCRIPT REGULAR EXPRESSION
Nucleosome Positioning (NP) &
Nucleosome Exclusion (NE) Signatures
SEQUENCE DETAILS & GRAPHICAL VIEW OF
NP & NE Signatures Highlighted Sequence
STATISTICAL DETAILS &
LIST OF NP AND NE SIGNATURES DETECTED
NP Signatures:
1. Signature starting with AT
2. Signature starting with GC
NE Signatures:
3. Signature starting with GC
4. Signature starting with AT (AT Homopolymer)
5. T tract (T Homopolymer)
6. A tract (A Homopolymer)
If parameters
were defined
GRAPHICAL VIEW OF
i. CpG Island Highlighted Sequence
ii. Promoter Region Highlighted Sequence
SAVE RESULT IN
COLOR-COMPATIBLE TEXT EDITOR

Figure 2. Computational Framework of CpGP Dynamics

signature immediately downstream of TSS, (iv). Choi's research on H2A.Z containing nucleosomes in resting T cells revealed that promoter containing a CpG island tend to remain nucleosome-free [27]. In other words, CpG island corresponds to NDRs and (v). An analysis on lipopolysaccharide inducible genes in macrophages illustrated that CpG island is a nucleosome destabilizing element which enables the transcriptional activator Sp1 to gain access to promoter binding sites in the uninduced (in the absence of stimulating signals) state without the need for nucleosome remodeling [28]. Hence, CpG island may compose NE signatures.

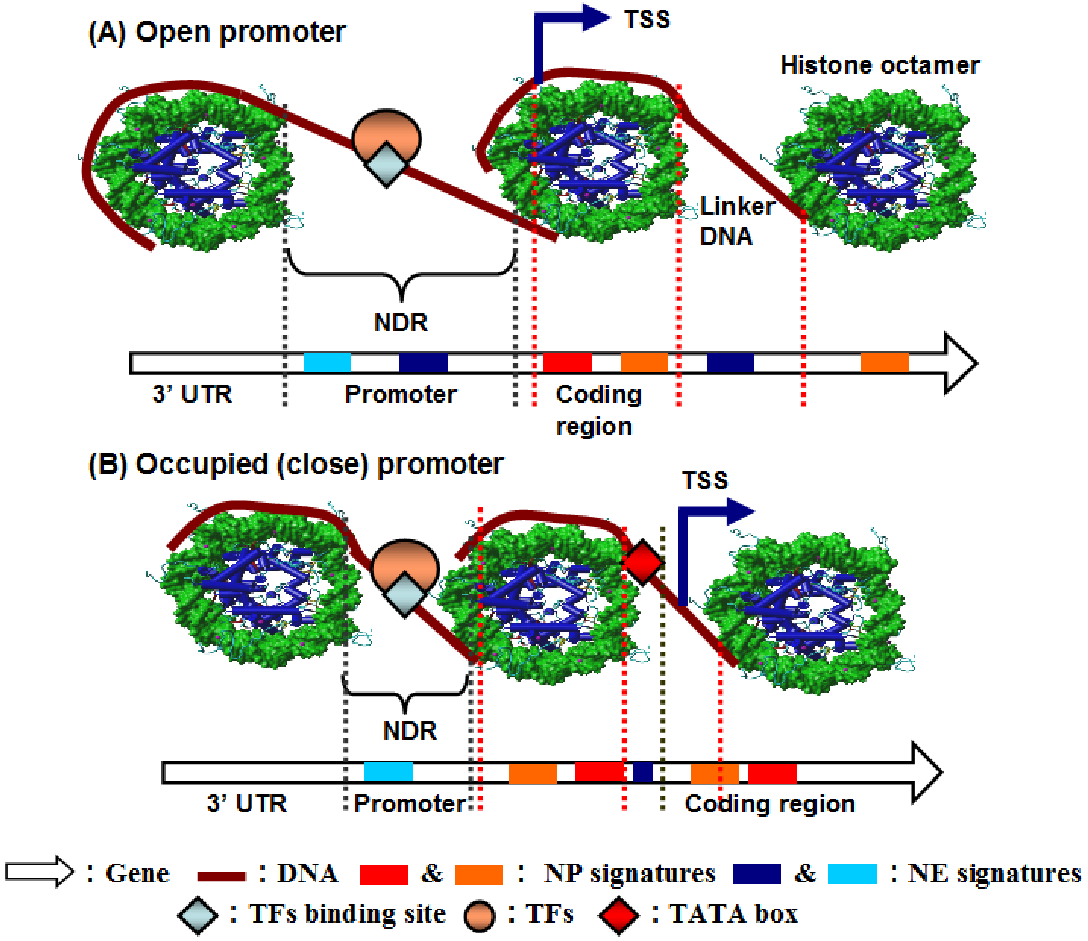


Figure 3. (A). Open promoters do have a long stretch of DNA containing nucleosome depleted region (NDRs) where nucleosome exclusion (NE) signatures can be observed while (B). Occupied (close) promoter contains a short NDR (and also positions TATA box at the nucleosome edge) in which the chance of finding NE signals are relatively low but increasing magnitude of NP signatures can be observed downstream

If the graphical sequence outputs of predicted signatures and promoter/CpG islands are compared, the occurrence of NP signatures can be related to coding and/or exonic part of gene whereas NE gives idea about the regulatory, the regions around TSS (both TATA and TATA-less promoters) as well as the 3'UTR. A graphical illustration (Figure 4) demonstrates the need of such comparison (nucleotide sequence, CDS and parameters given as per Table 1 vide S. No. 1). It is clear from the output that a NE signature was observed just upstream of TSS (marked with +1), the promoter and CpG islands were not overlapping and there was no signature (in the 236 bp upstream to TSS) found on the promoter region indicating a stable nucleosome.

3.3 Benchmarking CpGP Dynamics

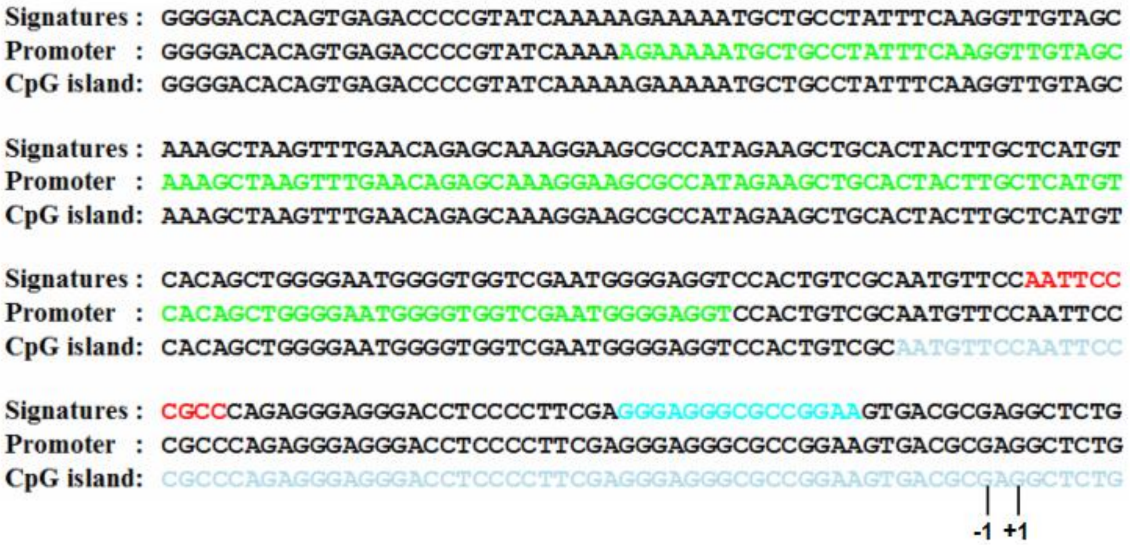
In order to benchmark and validate the presented program, experimentally known nucleosome formation sequences (NFSs) were retrieved from Nucleosome Positioning Region Database

Table 1. Result of DNA Sequence-based Programs to drive CpGP Dynamics

S.No	Program	Result obtained (Predictions made for 1000 bp (-700 to +299) of HMBS gene)	CpGP dynamics parameters / Comments	NP and NE signatures predicted by CpGP dynamics Total NP: 7 NE: 6
1	Promoter 2.0	TSS : 600 Score : 1.125 (Highly likely prediction)	CSP: 635 CEP: 950 PSP: 500 PEP: 625	Sign. on TSS: - Sign. on PR: - Sign. on CGI: NP 4, NE 2
2	Neural Network Promoter Prediction (NNPP)	Overall score cutoff: 0. 80 PSP: 322 PEP: 372 TSS : 363 Score : 0.96	CSP: 635 CEP: 950 PSP: 322 PEP: 372	Sign. on TSS: - Sign. on PR: NP 1, NE 1 Sign. on CGI: NP 4, NE 2
		PSP: 766 PEP: 816 TSS : 807 Score : 0.82	CSP: 635 CEP: 950 PSP: 766 PEP: 816	Sign. on TSS: NP Sign. on PR: NP 1, NE 1 Sign. on CGI: NP 4, NE 2
3	FirstEF	PSP: 306 PEP: 875 P(promoter): 0.9688 CpG window: 672 – 873 Exon: 806 – 876 P (exon): 1.0000	CSP: 672 CEP: 873 PSP: 306 PEP: 875	Sign. on TSS: - Sign. on PR: NP 3, NE 4 Sign. on CGI: NP 2, NE 2
4	ProScan Version 1.7	PR: 3378 – 3628 Promoter score: 53.71 Promoter cutoff: 53.0000	No prediction in the first 1000 bp. Hence, discarded from the analysis	-
5	CpGProD	CpG island associated promoter region: 462 – 1590 G+C frequency: 0.5958 CpG o/e ratio: 0.6664	CSP: 462 CEP: 1000 PSP: 462 PEP: 1000	Sign. on TSS: - Sign. on PR / CGI : NP 4, NE 2

6	CPGPLOT (EMBOSS program)	G+C frequency: >50.00 % CpG o/e ratio: >0.60 Length: 635 - 950	This length value has been specified as CSP and CEP for analysis of programs vide. S. no. 1 and 2.	-

Abbreviations: TSS - transcription start site, CSP - CpG island starting position, CEP - CpG island ending position, PSP - promoter starting position, PEP - promoter ending position, Sign. - signature, PR - promoter region, CGI - CpG island, NP - nucleosome positioning, NE - nucleosome exclusion.



```
Signatures : GGGGACACAGTGAGACCCCGTATCAAAAAGAAAAATGCTGCCTATTTCAAGGTTGTAGC
Promoter   : GGGGACACAGTGAGACCCCGTATCAAAAAGAAAAATGCTGCCTATTTCAAGGTTGTAGC
CpG island:  GGGGACACAGTGAGACCCCGTATCAAAAAGAAAAATGCTGCCTATTTCAAGGTTGTAGC

Signatures : AAAGCTAAGTTTGAACAGAGCAAAGGAAGCGCCATAGAAGCTGCACTACTTGCTCATGT
Promoter   : AAAGCTAAGTTTGAACAGAGCAAAGGAAGCGCCATAGAAGCTGCACTACTTGCTCATGT
CpG island:  AAAGCTAAGTTTGAACAGAGCAAAGGAAGCGCCATAGAAGCTGCACTACTTGCTCATGT

Signatures : CACAGCTGGGGAATGGGGTGGTCGAATGGGGAGGTCCACTGTCGCAATGTTCCAATTCC
Promoter   : CACAGCTGGGGAATGGGGTGGTCGAATGGGGAGGTCCACTGTCGCAATGTTCCAATTCC
CpG island:  CACAGCTGGGGAATGGGGTGGTCGAATGGGGAGGTCCACTGTCGCAATGTTCCAATTCC

Signatures : CGCCCAGAGGGAGGGACCTCCCCTTCGAGGGAGGGCGCCGGAAGTGACGCGAGGCTCTG
Promoter   : CGCCCAGAGGGAGGGACCTCCCCTTCGAGGGAGGGCGCCGGAAGTGACGCGAGGCTCTG
CpG island:  CGCCCAGAGGGAGGGACCTCCCCTTCGAGGGAGGGCGCCGGAAGTGACGCGAGGCTCTG
```

Figure 4. The need for comparing graphical outputs. Shown is an excerpt of HMBS gene in which 236 bp of DNA sequence upstream to experimental TSS (marked with +1). A NE signal (light blue colored texts) in the 'Signatures' row was observed just upstream of TSS. The TSS, NP signal (red colored texts) were observed in CpG island (pale blue colored texts) whereas promoter region (green colored texts) were found 10 bp apart from CpG island indicating both are closely regulated

(NPRD) [29]. NPRD is the first curated NFS-oriented database which compiles available experimental data on locations and characteristics of NFSs. Random selection of fifty human NFSs were recovered (NPRD sequence accession numbers are provided in Table 2) and each NFS was submitted one by one to CpGP dynamics (no other fields were filled during the submission) to predict the underlying NP and NE signature in the sequence. Ideally, the NFSs should not contain any NE signatures. Upon counting the number of positioning signals, a sum of 49 signatures were found in 50 NFSs with few exceptions. It was also seen that 62 exclusion expressions were predicted in the benchmarked dataset. Notably, these NFSs (a total of 50) having a sequence length of 6071 bp was comprised with NE signatures of length

715 bp (total of NE signatures length in the dataset) contributing 11.78 % of exclusion distribution. Subsequently, NE signatures were also manually checked to determine any reports of false-positives. It was confirmed that there is no such incorrect predictions made boosting up the accuracy of prediction. The occurrence of NE signatures may be due to the additional protein factors associated with the nucleosome in the dataset.

Table 2. NPRD Dataset used in Benchmark

S. No	NPRD sequence accession number	NP signatures predicted	NE signatures predicted	S. No	NPRD sequence accession number	NP signatures predicted	NE signatures predicted
1	N00060	1	2	26	N00668	0	2
2	N00061	1	1	27	N00669	0	1
3	N00062	0	2	28	N00670	2	2
4	N00063	2	0	29	N00671	0	0
5	N00064	1	4	30	N00672	0	1
6	N00648	0	0	31	N00673	0	1
7	N00649	2	1	32	N00674	0	1
8	N00650	2	1	33	N00675	0	0
9	N00651	2	1	34	N00050	2	2
10	N00652	2	1	35	N00059	2	0
11	N00653	0	0	36	N00688	0	6
12	N00654	1	1	37	N00689	2	5
13	N00655	0	2	38	N00690	1	4
14	N00656	2	0	39	N00691	1	4
15	N00657	2	0	40	N00692	3	0
16	N00658	0	0	41	N00693	3	1
17	N00659	0	1	42	N00916	0	0
18	N00660	0	1	43	N00917	2	0
19	N00661	1	0	44	N00012	0	0
20	N00662	0	0	45	N00027	1	2
21	N00663	2	1	46	N00028	2	1
22	N00664	2	2	47	N00029	0	2
23	N00665	1	0	48	N00093	1	4
24	N00666	1	1	49	N00094	0	0
25	N00667	1	0	50	N00607	1	1

Abbreviations: NPRD - nucleosome positioning region database,
NP - nucleosome positioning, NE - nucleosome exclusion.

3.4 Comparison with DNA Sequence based Nucleosome Position Prediction Programs

CpGP dynamics makes its predictions by searching over the experimentally reported signatures (regular expressions in programming sense) in the user given nucleotide sequence. It was compared with online nucleosomes position prediction tool developed at Segal's Lab of Computational Biology with Kaplan *et al*, 2008 (version 3.0) [30] as selected model and NuPoP build up by Northwestern University, USA [8]. The former program is based on probabilistic model derived from sequence preferences studied using *in vitro* map while the latter considers duration HMM concentrated on linker length distribution for making out its predictions. Hydroxymethylbilane synthase (HMBS) gene of *Homo sapiens* was retrieved from National Center for Biotechnology Information (NCBI) Gene database [31] with an accession number NM_000190 as random input to accomplish this comparison. This 10 kb gene contains 15 exons [32] and a manual consensus search of TATA and GC box revealed its locations at -369 to -361 and -19 and -14, respectively. Upon examining the RefSeq annotation of HMBS gene in Eukaryotic Promoter Database [33] (entry: EP26007), the TSS, promoter region and the sequence positions were calibrated accordingly in our tested gene sequence and these elements position were also authenticated by cross-referencing the literature information published by Whatley *et al*, 2000 [32].

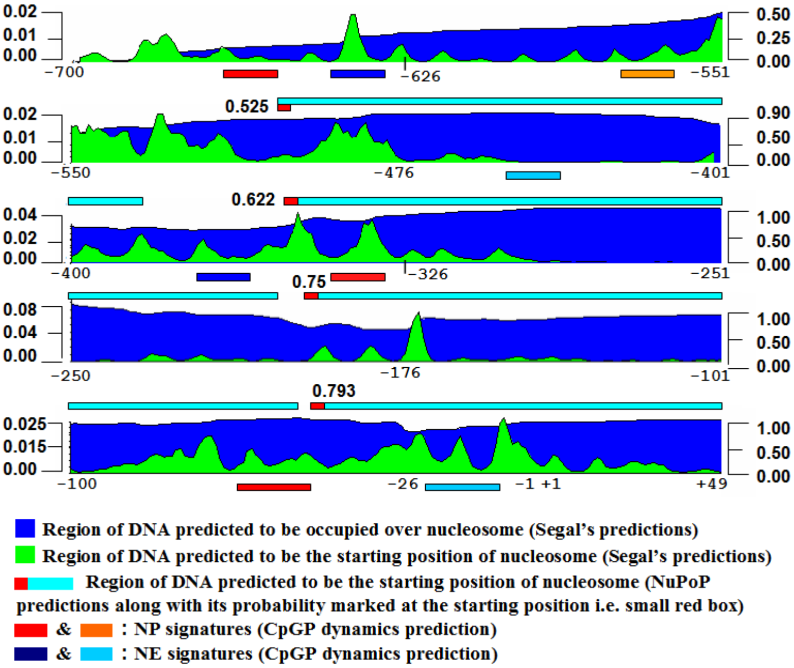
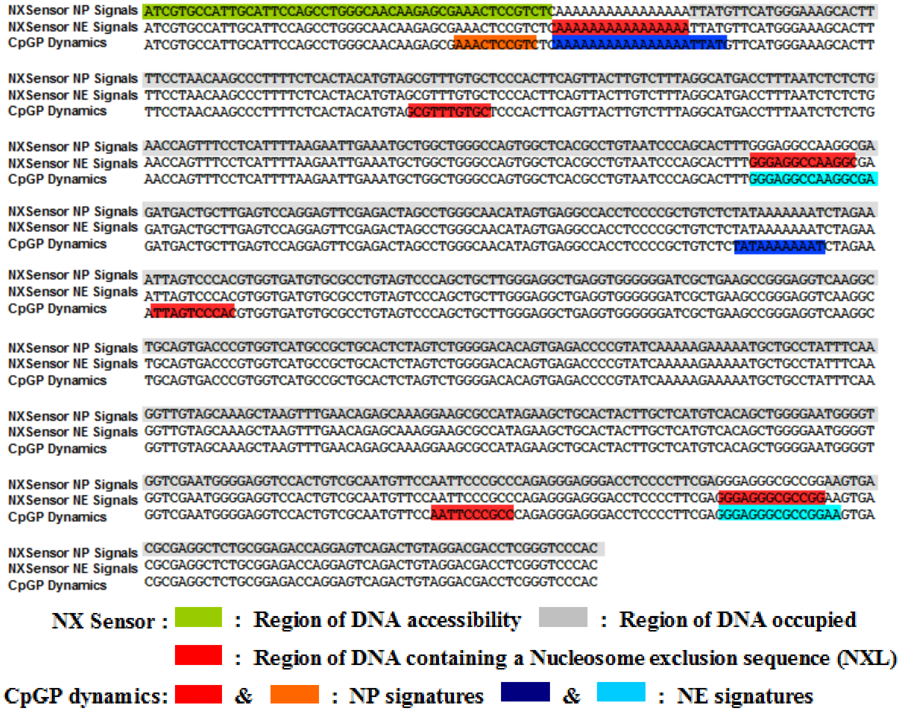


Figure 5. A graphical representation of predictions made at the promoter region of HMBS gene (1000 bp shown with experimentally known sequence positions at the x-axis) using Segal's tool, NuPoP and CpGP dynamics. Segal's prediction was represented as areas (vertical scale at the left side indicates the probability of being occupied over nucleosome and at the right side ranges the probability of starting position of nucleosome as predicted by Segal's tool), NuPoP results were shown appropriately at the upper part of the area and CpGP dynamics detected signatures were represented as boxed regions at the bottom of the area graph

A graphical sketch was presented (Figure 5) which represents the predictions made by above mentioned two programs with the CpGP dynamics predictions along the horizontal axis (x-axis). The regions of DNA predicted to be occupied over nucleosomes (blue colored area; predicted by Segal's tool) around the TSS (consider -20 to +20) wasn't in good correspondence with the anticipations as one might expects a valley because of lower occupancy (drop in area) in the TSS position and a higher occupancy around TSS. But the region around TSS was marked only with a linear surface area bounded by a probability of 0.025 (refer left vertical scale). Undoubtedly, CpGP dynamics predicted a NE signature (light blue colored box along the x-axis) at -23 to -9 position range which is just upstream of TSS indicating a lower occupancy and NDR. Fortunately, the nucleosome starting position represented in the form of peak as predicted by Segal's tool with a probability of 0.97 (refer right vertical scale) was observed at -9 site and supports the prediction

of NE signature predicted by the presented program. Further, comparison with NuPoP results (p-start score: probability of starting position of a nucleosome; showed on the upper part of the surface area) showed that an exclusion signal predicted by CpGP dynamics was immediately followed by a nucleosome starting position predicted by NuPoP with a probability of 0.622. Moreover, a slighter downstream region was predicted with a signal of positioning (refer 3rd lane



```
NXSensor NP Signals ATCGTGCCATTGCATTCCAGCCTGGGCAACAAGAGCGAAACTCCGTCTCAAAAAAAAAAAAAAAATTATGTTCATGGGAAAGCACTT
NXSensor NE Signals ATCGTGCCATTGCATTCCAGCCTGGGCAACAAGAGCGAAACTCCGTCTCAAAAAAAAAAAAAAAATTATGTTCATGGGAAAGCACTT
CpGP Dynamics       ATCGTGCCATTGCATTCCAGCCTGGGCAACAAGAGCGAAACTCCGTCTCAAAAAAAAAAAAAAAATTATGTTCATGGGAAAGCACTT

NXSensor NP Signals TTCCTAACAAGCCCTTTTCTCACTACATGTAGCGTTTGTGCTCCCACTTCAGTTACTTGTCTTTAGGCATGACCTTTAATCTCTCTG
NXSensor NE Signals TTCCTAACAAGCCCTTTTCTCACTACATGTAGCGTTTGTGCTCCCACTTCAGTTACTTGTCTTTAGGCATGACCTTTAATCTCTCTG
CpGP Dynamics       TTCCTAACAAGCCCTTTTCTCACTACATGTAGCGTTTGTGCTCCCACTTCAGTTACTTGTCTTTAGGCATGACCTTTAATCTCTCTG

NXSensor NP Signals AACCAGTTTCCTCATTTTAAGAATTGAAATGCTGGCTGGGCCAGTGGCTCACGCCTGTAATCCCAGCACTTTGGGAGGCCAAGGCGA
NXSensor NE Signals AACCAGTTTCCTCATTTTAAGAATTGAAATGCTGGCTGGGCCAGTGGCTCACGCCTGTAATCCCAGCACTTTGGGAGGCCAAGGCGA
CpGP Dynamics       AACCAGTTTCCTCATTTTAAGAATTGAAATGCTGGCTGGGCCAGTGGCTCACGCCTGTAATCCCAGCACTTTGGGAGGCCAAGGCGA

NXSensor NP Signals GATGACTGCTTGAGTCCAGGAGTTCGAGACTAGCCTGGGCAACATAGTGAGGCCACCTCCCCGCTGTCTCTATAAAAAAATCTAGAA
NXSensor NE Signals GATGACTGCTTGAGTCCAGGAGTTCGAGACTAGCCTGGGCAACATAGTGAGGCCACCTCCCCGCTGTCTCTATAAAAAAATCTAGAA
CpGP Dynamics       GATGACTGCTTGAGTCCAGGAGTTCGAGACTAGCCTGGGCAACATAGTGAGGCCACCTCCCCGCTGTCTCTATAAAAAAATCTAGAA

NXSensor NP Signals ATTAGTCCCACGTGGTGATGTGCGCCTGTAGTCCCAGCTGCTTGGGAGGCTGAGGTGGGGGATCGCTGAAGCCGGGAGGTCAAGGC
NXSensor NE Signals ATTAGTCCCACGTGGTGATGTGCGCCTGTAGTCCCAGCTGCTTGGGAGGCTGAGGTGGGGGATCGCTGAAGCCGGGAGGTCAAGGC
CpGP Dynamics       ATTAGTCCCACGTGGTGATGTGCGCCTGTAGTCCCAGCTGCTTGGGAGGCTGAGGTGGGGGATCGCTGAAGCCGGGAGGTCAAGGC

NXSensor NP Signals TGCAGTGACCCGTGGTCATGCCGCTGCACTCTAGTCTGGGGACACAGTGAGACCCCGTATCAAAAAGAAAAATGCTGCCTATTTCAA
NXSensor NE Signals TGCAGTGACCCGTGGTCATGCCGCTGCACTCTAGTCTGGGGACACAGTGAGACCCCGTATCAAAAAGAAAAATGCTGCCTATTTCAA
CpGP Dynamics       TGCAGTGACCCGTGGTCATGCCGCTGCACTCTAGTCTGGGGACACAGTGAGACCCCGTATCAAAAAGAAAAATGCTGCCTATTTCAA

NXSensor NP Signals GGTTGTAGCAAAGCTAAGTTTGAACAGAGCAAAGGAAGCGCCATAGAAGCTGCACTACTTGCTCATGTCACAGCTGGGGAATGGGGT
NXSensor NE Signals GGTTGTAGCAAAGCTAAGTTTGAACAGAGCAAAGGAAGCGCCATAGAAGCTGCACTACTTGCTCATGTCACAGCTGGGGAATGGGGT
CpGP Dynamics       GGTTGTAGCAAAGCTAAGTTTGAACAGAGCAAAGGAAGCGCCATAGAAGCTGCACTACTTGCTCATGTCACAGCTGGGGAATGGGGT

NXSensor NP Signals GGTCGAATGGGGAGGTCCACTGTCGCAATGTTCCAATTCCCGCCCAGAGGGAGGGACCTCCCCTTCGAGGGAGGGCGCCGGAAGTGA
NXSensor NE Signals GGTCGAATGGGGAGGTCCACTGTCGCAATGTTCCAATTCCCGCCCAGAGGGAGGGACCTCCCCTTCGAGGGAGGGCGCCGGAAGTGA
CpGP Dynamics       GGTCGAATGGGGAGGTCCACTGTCGCAATGTTCCAATTCCCGCCCAGAGGGAGGGACCTCCCCTTCGAGGGAGGGCGCCGGAAGTGA

NXSensor NP Signals CGCGAGGCTCTGCGGAGACCAGGAGTCAGACTGTAGGACGACCTCGGGTCCCAC
NXSensor NE Signals CGCGAGGCTCTGCGGAGACCAGGAGTCAGACTGTAGGACGACCTCGGGTCCCAC
CpGP Dynamics       CGCGAGGCTCTGCGGAGACCAGGAGTCAGACTGTAGGACGACCTCGGGTCCCAC
```

NX Sensor : 🟩 : Region of DNA accessibility ▨ : Region of DNA occupied
🟥 : Region of DNA containing a Nucleosome exclusion sequence (NXL)
CpGP dynamics: 🟥 & 🟧 : NP signatures 🟦 & 🟦 : NE signatures

Figure 6. Graphical illustration of predictions made at the promoter region of HMBS gene (1000 bp shown) by CpGP dynamics and NXSensor programs

i.e., sequence positions from -400 to -326) describing an overall stable predictions (nucleosome) at this region. Likewise, when compared at annotated TSS site, no such good discrimination was observed. Concisely, the approach of probabilistic model and sequence based pattern search has its own advantages and limitations and henceforth, if all the predictions were compared, this will give an outline suggesting regions of low and overwhelmed predictions.

3.5 Comparison with Program Employing Similar Algorithm

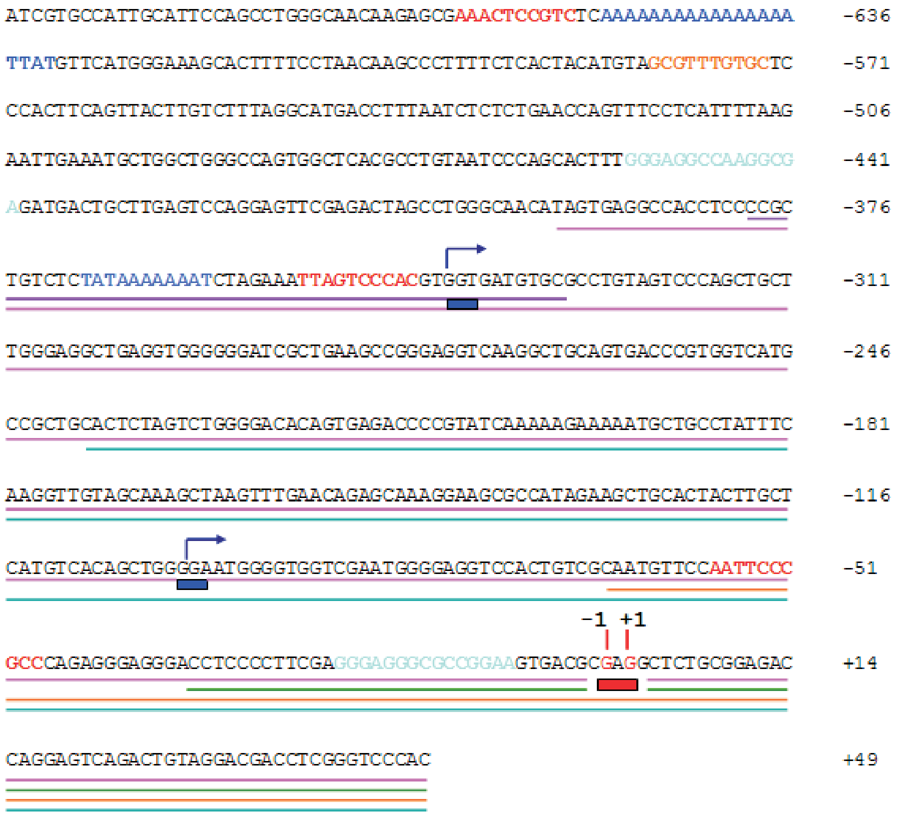
Furthermore, the results of CpGP dynamics were compared with NXSensor which only searches the nucleosome exclusion sequences (NXLs) in the DNA sequence built from DNA bending and flexibility experimental observations [9]. The above mentioned gene sequence was utilized to test the predictions. Three NXLs were reported by NXSensor and CpGP dynamics predicted the NE signatures exactly at the same position (Figure 6). Unfortunately, NXSensor was unsuccessful to identify a NE (sequence: TATAAAAAAAT at the 4th lane; the non-inclusion of poly (dA:dT) tract as NXL was the exact reason for this insignificance) but CpGP dynamics discriminated well. This unpredicted NE was representing a TATA box. The main advantage of using the presented program is that it also predicts the NP signatures (4 counts in the sequence provided) thereby enhancing the prediction accuracy.

3.6 The Protocol of CpGP Dynamics

It has been described elsewhere that CpGP dynamics also considers the input of numerical parameters (CpG island starting and ending position and promoter starting and ending position) predicted by third-party programs. To demonstrate the usage of presented program, HMBS gene was considered as testing sequence (only 1000 bp, sequence range: -700 to +299; it was ensured that an experimentally known TSS, promoter and CpG island were comprised within this range) and provided as input to predict promoter regions and/or the TSS position (programs: Promoter 2.0 [13], Neural Network Promoter Prediction (NNPP) [11], FirstEF (First Exon Finder) [12], ProScan version 1.7 [34]; CpG island associated promoter prediction: CpGProD [35]) and the CpG island location (programs: EMBOSS CPGPLOT [36], CpGProD [35]) in the gene sequence. CpGP dynamics was employed solely to detect NP and NE signatures in the testing gene sequence and without any knowledge of numerical parameters. It predicted 7 positioning and 6 exclusion signals.

CPGPLOT of EMBOSS suite was used to scan the putative CpG island(s) in the gene (Gardiner and Frommer sequence criterion was obeyed) and showed that positions from 635 to 950 (the sequence positions were calibrated appropriately with the experimentally annotated positions) were distinguished as CpG island (Figure 7). This CpG island position was used as one among the numerical parameter for specification in CpGP dynamics when Promoter 2.0 and NNPP prediction results were submitted individually. Promoter 2.0 predicted a TSS at 600 with a score, 1.125 (identified as highly likely prediction). A major drawback of computational prediction of eukaryotic promoters is that the length (sequence range) of predicted region is comparatively high insisting the need to depend precisely on experimental identification. According to Genomatix [37], it has been recommended to restrict the promoter region to about 30 to 1000 bp upstream of TSS for initial *in silico* screening. As Promoter 2.0 only provided the TSS position, a sequence window of 125 bp with TSS was constructed, i.e. promoter starting position: 500, TSS: 600 and ending position: 625. NNPP identified 2 promoter regions with an overall score cutoff of 0.80 along with the information about the TSS, promoter range and validated by its region based score (regions scored more than 0.80 was taken). These promoter range (i.e. 322-372 and 766-816) and CPGPLOT predictions was specified in numerical fields. Next, FirstEF yielded information on promoter and CpG island window along with first exon prediction. However, positions relevant to numerical parameters were only considered for submission. ProScan version 1.7 predicted no elemental positions in the first 1000 bp. An attempt was carried out to lower its threshold upto 53 and no such predictions were observed. Hence, it was decided to discard its predictions for comparing CpG dynamics results. Another program, CpGProD developed with intentions of identifying CpG island associated promoters and thereby didn't discriminate both of the elements: CpG island and promoter, but provided a putative sequence range (462-1000). Hence, the starting positions of CpG island as well as promoter was defined as 462 while 1000 was mentioned as ending positions.

Upon individually examining the graphical outputs of promoter and CpG island (generated due to numerical parameters specification) with that of signatures embedded graphical output of CpGP dynamics (similarly as shown in Figure 4), Promoter 2.0 predicted a TSS at -100 position (Figure 7), a 100 bp upstream of experimentally known TSS site (+1). NNPP made its first prediction at the position -379 to -330 in which an exclusion and positioning signals were found. This NE signature was predicted due to the presence of TATA box in that region. On the other hand, NNPP's second prediction was observed very downstream to known TSS. FirstEF results were in good correspondence with the CpGP dynamics outcome as the predicted region was composed of 5 positioning and 6 exclusion signals (NP: 3 and NE: 4 in promoter region and NP: 2 and NE: 2 in CpG island, refer Table 1 vide S. No. 3). CpGProD provides information only


ATCGTGCCATTGCATTCCAGCCTGGGCAACAAGAGCGAAACTCCGTCTCAAAAAAAAAAAAAAAA -636
TTATGTTCATGGGAAAGCACTTTTCCTAACAAGCCCTTTTCTCACTACATGTAGCGTTTGTGCTC -571
CCACTTCAGTTACTTGTCTTTAGGCATGACCTTTAATCTCTCTGAACCAGTTTCCTCATTTTAAG -506
AATTGAAATGCTGGCTGGGCCAGTGGCTCACGCCTGTAATCCCAGCACTTTGGGAGGCCAAGGCG -441
AGATGACTGCTTGAGTCCAGGAGTTCGAGACTAGCCTGGGCAACATAGTGAGGCCACCTCCCCGC -376
TGTCTCTATAAAAAAATCTAGAAATTAGTCCCACGTGGTGATGTGCGCCTGTAGTCCCAGCTGCT -311
TGGGAGGCTGAGGTGGGGGGATCGCTGAAGCCGGGAGGTCAAGGCTGCAGTGACCCGTGGTCATG -246
CCGCTGCACTCTAGTCTGGGGACACAGTGAGACCCCGTATCAAAAAGAAAAATGCTGCCTATTTC -181
AAGGTTGTAGCAAAGCTAAGTTTGAACAGAGCAAAGGAAGCGCCATAGAAGCTGCACTACTTGCT -116
CATGTCACAGCTGGGGAATGGGGTGGTCGAATGGGGAGGTCCACTGTCGCAATGTTCCAATTCCC -51
-1 +1
GCCCAGAGGGAGGGACCTCCCCTTCGAGGGAGGGCGCCGGAAGTGACGCGAGGCTCTGCGGAGAC +14
CAGGAGTCAGACTGTAGGACGACCTCGGGTCCCAC +49

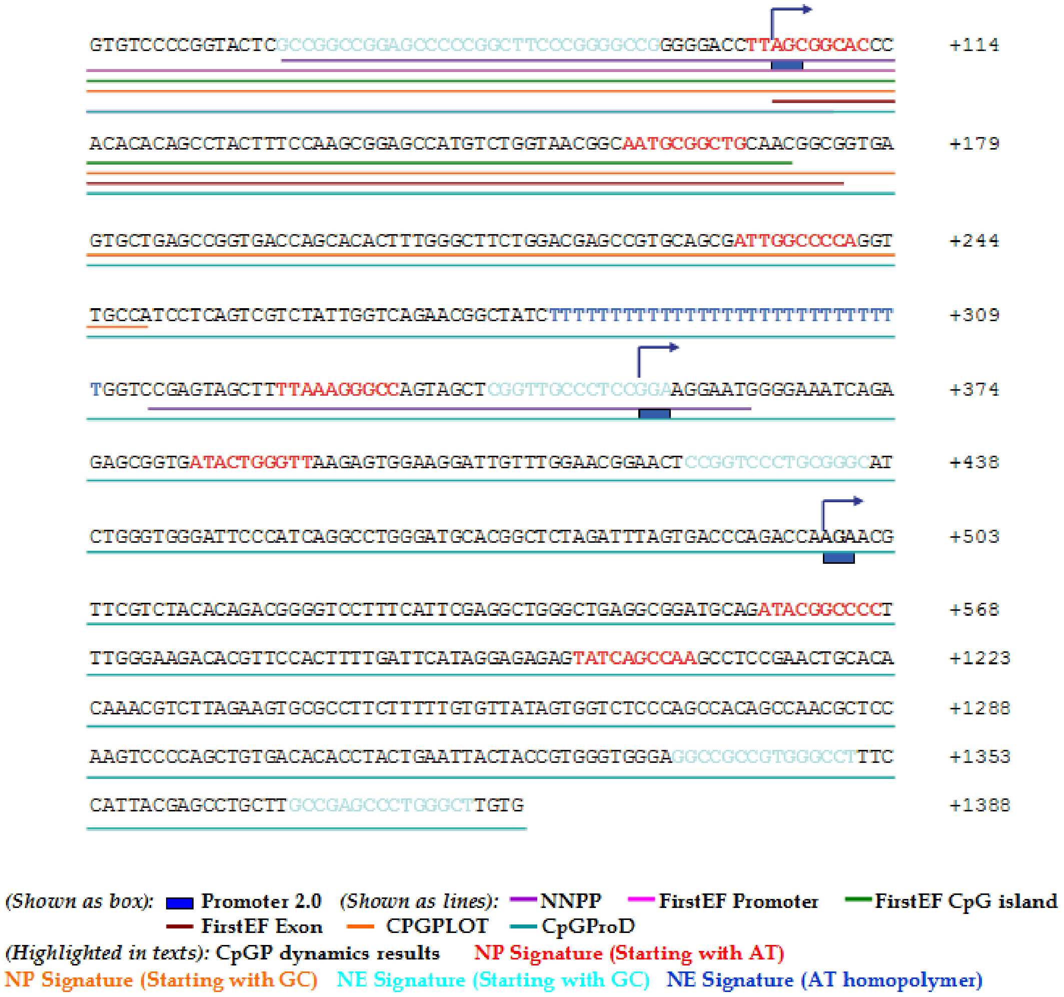

GTGTCCCCGGTACTCGCCGGCCGGAGCCCCCGGCTTCCCGGGGCCGGGGGACCTTAGCGGCACCC +114
ACACACAGCCTACTTTCCAAGCGGAGCCATGTCTGGTAACGGCAATGCGGCTGCAACGGCGGTGA +179
GTGCTGAGCCGGTGACCAGCACACTTTGGGCTTCTGGACGAGCCGTGCAGCGATTGGCCCCAGGT +244
TGCCATCCTCAGTCGTCTATTGGTCAGAACGGCTATCTTTTTTTTTTTTTTTTTTTTTTTTTT +309
TGGTCCGAGTAGCTTTTAAAGGGCCAGTAGCTCGGTTGCCCTCCGGAAGGAATGGGGAAATCAGA +374
GAGCGGTGATACTGGGTTAAGAGTGGAAGGATTGTTTGGAACGGAACTCCGGTCCCTGCGGGCAT +438
CTGGGTGGGATTCCCATCAGGCCTGGGATGCACGGCTCTAGATTTAGTGACCCAGACCAAGAACG +503
TTCGTCTACACAGACGGGGTCCTTTCATTCGAGGCTGGGCTGAGGCGGATGCAGATACGGCCCCT +568
TTGGGAAGACACGTTCCACTTTTGATTCATAGGAGAGAGTATCAGCCAAGCCTCCGAACTGCACA +1223
CAAACGTCTTAGAAGTGCGCCTTCTTTTTGTGTTATAGTGGTCTCCCAGCCACAGCCAACGCTCC +1288
AAGTCCCCAGCTGTGACACACCTACTGAATTACTACCGTGGGTGGGAGGCCGCCGTGGGCCTTTC +1353
CATTACGAGCCTGCTTGCCGAGCCCTGGGCTTGTG +1388
(Shown as box): Promoter 2.0
(Shown as lines): NNPP
FirstEF Promoter
FirstEF CpG island
FirstEF Exon
CPGPLOT
CpGProD
(Highlighted in texts): CpGP dynamics results
NP Signature (Starting with AT)
NP Signature (Starting with GC)
NE Signature (Starting with GC)
NE Signature (AT homopolymer)

Figure 7. Overlay of predicted promoter and CpG island with CpGP dynamics signature predictions describing the requirement for integration of various knowledge

about the CpG island associated promoter with a sequence range (-37 to +1590: 1628 bp). It was also unsuccessful in encompassing the CpG island positions as normally predicted by CpG island searching programs (in this case, CPGPLOT and FirstEF CpG window prediction). Hence, the comparison of CpGProD with the presented program cannot be achieved. A histogram describing the predictions of various signatures found in the sequence window provided by sequence-based promoter and CpG island computational programs was shown (Figure 8). If the provided promoter and CpG island composes an equal distribution of NP and NE signals, then it can be one amongst the more accurate predictions and so, eliminates the hits made by marginal predictions.

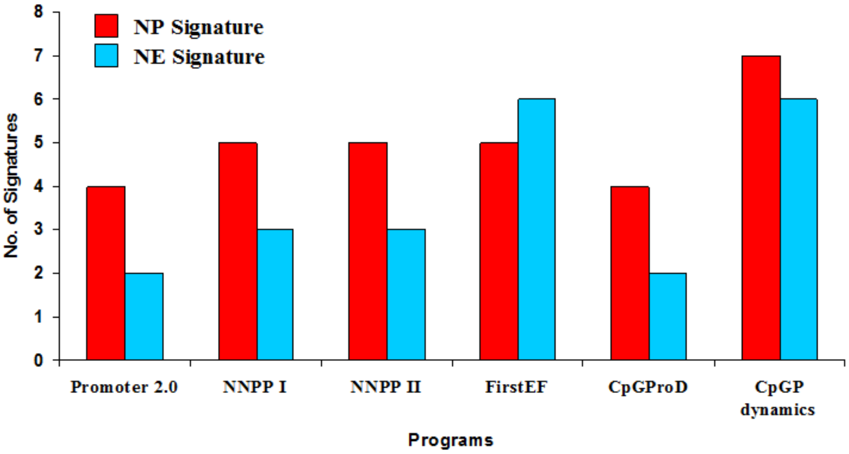


Figure 8. Comparison of CpGP Dynamics with Sequence-based Promoter and CpG Island Computational Programs

4. Conclusion

Intrinsic sequence dependencies dictate nucleosome positioning and occlusion partially which paved way for developing empirical rules. Computational analysis of genomic regions revealed its close relationship with gene regulation. The present work was aimed to develop a bioinformatic program which predicts positioning and occlusion signals upon which the knowledge of CpG island and promoter region locations was applied to discriminate the more accurate predictions from a list of likelihood usually generated for a DNA sequence. There is a tremendous requirement for integrating the pattern of statistical positioning and relate its locations with genomic elements.

Acknowledgements

The author acknowledged with thank for generous support from The Department of Science and Technology, Innovation in Science Pursuit for Inspired Research (INSPIRE) Scheme, Govt. of India. Thanks were due to 3 anonymous reviewers for their critical assessment and valuable suggestions. The author is grateful to Mrs. Sukanya Rawal for analyzing the program codes to comply with the standards of non-server based web programs and for introducing some of the excerpts.

References

[1] Richmond TJ and Davey CA, Nature, vol. 423, **(2003)**.

[2] White JH and Bauer WR, Cell, vol. 56, **(1989)**.

[3] Widom J, Q. Rev. Biophys, vol. 34, **(2001) .**

[4] Segal E, Mittendorf YF, Chen L, Thastrom AC, Field Y, Moore IK, Wang JPZ and Widom J, Nature, vol. 442, no. 7104, **(2006)**.

[5] Daenen F, Roy FV and De Bleser PJ, BMC Genomics, vol. 9, no. 332, **(2008)**.

[6] Lockea G, Tolkunova D, Moqtaderib Z, Struhlb K and Morozova AV, Proc. Natl. Acad. Sci., vol. 107, no. 49, **(2010)**.

[7] Gabdank D, Barash D and Trifonov EN, J. Biomol. Str. Dyn., vol. 26, no. 4, **(2009)**.

[8] Xi L, Mittendorf YF, Xia L, Flatow J, Widom J and Wang JP, BMC Bioinformatics, vol. 11, no. 346, **(2010)**.

[9] Luykx P, Bajic IV and Khuri S, Nucl. Acids Res. (Web server issue), vol. 34, **(2006)**.

[10] Munteanu MG, Vlahovicek K, Parthasarathy S, Simon I and Pongor S, Trends Biochem. Sci., vol. 23, **(1998)**.

[11] Reese MG, Comp. & Chem., vol. 26, **(2001)**.

[12] Davuluri RV, Grosse I and Zhang MQ, Nat. Gen., vol. 29, **(2001)**.

[13] Knudsen S, Bioinformatics, vol. 15, no. 5, **(1999)**.

[14] Web Promoter Scan Service developed at BioInformatics and Molecular Analysis Section (BIMAS), Center for Information Technology, National Institute of Health, http:// www.bimas.cit.nih.gov/molbio/proscan.

[15] Gardiner-Garden M and Frommer M, J. Mol. Biol., vol. 196, **(1987)**.

[16] Shrader TE and Crothers DM, Proc. Natl. Acad. Sci., vol. 86, **(1989)**.

[17] Wang YH and Griffith JD, Proc. Natl. Acad. Sci., vol. 93, **(1996)**.

[18] Satchwell SC, Drew HR and Travers AA, J. Mol. Biol., vol. 191, **(1986)**.

[19] Suter B, Schnappauf G and Thoma F, Nucl. Acids Res., vol. 28, **(2000)**.

[20] Segal E and Widom J, Curr. Opin. Struct. Biol., vol. 19, no. 1, **(2009)**.

[21] Dechering KJ, Cuelenaere K, Konings RN and Leunissen JA, Nucl. Acids Res., vol. 26, **(1998)**.

[22] Anderson JD and Widom J, Mol. Cell. Biol., vol. 21, **(2001)**.

[23] Duffy S, in How to do everything with JavaScript, McGraw-Hill/Osborne Publishers, California USA, vol. 1, **(2003)**, pp. 179-205.

[24] Arya G**,** Maitra A and Grigoryev SA, J. Biomol. Str. Dyn., vol. 27, no. 6, **(2010)**.

[25] Whitehouse W, Rando OJ, Delrow J and Tsukiyama T, Nature, vol. 450, **(2007)**.

[26] Lee W, Tillo D, Bray N, Morse RH, Davis RW, Hughes TR and Nislow C, Nat. Genet., vol. 39, **(2007)**.

[27] Choi JK, Genome Biol., vol. 11, no. R70, **(2010) .**

[28] Ramirez-Carrozzi VR, Braas D, Bhatt DM, Cheng CS, Hong C, Doty KR, Black JC, Hoffmann A, Carey M and Smale ST, Cell, vol. 138, no. 1, **(2009)**.

[29] Levitsky VG, Katokhin AV, Podkolodnaya OA, Furman DP and Kolchanov NA, Nucl. Acids Res. (Database issue), vol. 33, **(2005)**.

[30] Kaplan N, Moore IK, Fondufe-Mittendorf Y, Gossett AJ, Tillo D, Field Y, LeProust EM, Hughes TR, Lieb JD, Widom J and Segal E, Nature, vol. 458, no. 7236, **(2009)**.

[31] McEntyre J and Ostell J, Editors, The NCBI handbook [Internet], National Library of Medicine U.S.A. **(2002)**, Accessible at http://www.ncbi.nlm.nih.gov/books/NBK21105/.

[32] Whatley SD, Roberts AG, Llewellyn DH, Bennett CP, Garrett C and Elder GH, Hum Genet., vol. 107, **(2000)**.

[33] Schmid CD, Perier R, Praz V and Bucher P, Nucl. Acids Res. (Database issue), vol. 34, **(2006)**.

[34] Prestridge DS, J. Mol. Biol., vol. 249, **(1995)**.

[35] Ponger L and Mouchiroud D, Bioinformatics, vol. 18, no. 4, **(2002)**.

[36] Rice P, Longden I and Bleasby A, TIG, vol. 16, no. 6, **(2000)**.

[37] Genomatix - Collection of programs for promoter identification, genome mapping, identification of transcription factor binding sites and design of mutation experiments, http://www.genomatix.de/.

Authors

S. Prasanth Kumar

S. Prasanth Kumar is currently working as faculty in Bioinformatics Laboratory, Gujarat University. He pursued his post-graduation from Department of Bioinformatics, Alagappa University, India. He has been awarded DST INSPIRE fellowship from Ministry of Science and Technology, Government of India to pursue doctoral study. He was a recipient of Young Scientist Award from Association of Biotechnology and Pharmacy, India and Gujarat University in the year 2009 and 2010, respectively. His research interest includes developing algorithms and tools pertaining to structural bioinformatics, molecular modeling and sequence analysis.