Bayesian classification for promoter prediction in human DNA sequences

J.-F. Bercher[*], P. Jardin[†] and B. Duriez[**]

[*] *Équipe Signal et Information,*
Dept. Modélisation et Simulation
ESIEE, Noisy-le-Grand, France

[†] *Équipe Signal et Information,*
Dept. Signaux et Télécoms
ESIEE, Noisy-le-Grand, France

[**] *Molecular and cellular bases of genetic diseases,*
INSERM-U654,
Créteil, France

Abstract.
Many Computational methods are yet available for data retrieval and analysis of genomic sequences, but some functional sites are difficult to characterize. In this work, we examine the problem of promoter localization in human DNA sequences. Promoters are regulatory regions that governs the expression of genes, and their prediction is reputed difficult, so that this issue is still open. We present the Chaos Game representation (CGR) of DNA sequences which has many interesting properties, and the notion of 'genomic signature' that proved relevant in phylogeny applications. Based on this notion, we develop a (naïve) bayesian classifier, evaluate its performances, and show that its adaptive implementation enable to reveal or assess core-promoter positions along a DNA sequence.

Key Words: DNA sequences, bayesian classification, promoter prediction, Chaos Game Representation

INTRODUCTION

Recent availability of several mammalian genome sequences has allowed whole genome analyses to unravel their functional properties. One of the challenge of the mammalian genomics is to understand how genomes are transcribed. Specifically, the genes are small sequences spread out along the genomes which, after being transcribed in mRNA, are translated in proteins turning out to be functional units of the cells. Although several sequences located within the genes or in their closed vicinity control their specificity of expression, there are typical regions, the promoters, that define the *Transcription Start Site* (TSS) of the genes. One specific gene may have several promoters and each give rise to a specific mRNA. Detection of the promoters all along the genomes is of crucial interest since it will enable identification of primarily the transcribed sequences, and afterwards the tied sequences accountable for, at least partly, expression specificity due to transcription factor binding.

The minimum functional part of promoters is defined as the core promoter, and it lies several hundred bases around the TSS. Until now, computational tools developed to predict core promoter locations are based on various criteria: biological data, presence of CpG islands closed to the TSS in more than half promoters, detection of specific or highly concentrated transcription binding factor sites (TFBS), homology with orthologous sequences or statistical properties of core promoters compared with other genomic sequences [1]. The approach we followed rests on this last criterion. It is based on the principle of genomic signature which consists in determining frequencies of all the 2-8 oligonucleotides in a given sequence. These frequencies may be figured by pictures known as chaos game representation (CGR) initially described by J. Jeffrey [2]. Comparisons of results issued from different sequences proved to be highly relevant in phylogeny applications [3, 4]. Indeed, genomic signature had shed light on the species -specific oligonucleotide frequencies [3, 5, 6].

Although Gentles and Karlin noticed that dinucleotide relative abundances are remarkably constant across human chromosomes and within the DNA of a particular species [7], we investigated whether functional sequences, especially core promoters, may have a specific genomic signature. Using this CGR we have put in evidence the nonstationarity of the genome: coding, promoter or genomic regions of DNA result in different CGR matrices. In particular we observe the fractal depletion in CG for genomic regions (that is under -representation of CG words) and CG "islands" in about 80% of promoters.

In order to analyse DNA sequences, references probabilities of the genomic, coding and promoters background are built using data from public databases. We also estimate "local" probability distribution functions, using a sliding window, and a forgetting factor.

We built a naïve bayesian classifier for promoter detection, by testing the likelihood ratio promoter/genomic or promoter/coding of the sequence at hand. Results show that performance is interesting when the window is located near the TSS , and the window length is less than 200 bases. Such a classifier has already be useful for classifying species as in [6].

CHAOS GAME REPRESENTATION (CGR)

The Chaos Game Representation is derived from Chaos Theory and presents several interesting properties: the source sequence can be recovered uniquely from the CGR transcription and the distance between CGR position measures similarity between corresponding sequences. There is an established link between CGR and Markov models [10], and an extension to arbitrary discrete sequences, leading to the Universal Sequence Mapping (USM) technique [11].

The CGR is an independent scale representation which maps in a iterative way a nucleotide sequence in the [0,1]x[0,1] square. We choose to consider this square in the complex plane because this allows a mono-dimensional description of the sequence which can be useful for further signal processing. We first assign to each nucleotide $S \in \{C, A, G, T\}$ a value $z(S)$ (position in the complex square) according to: $z(C) = 0 + j0$, $z(A) = 1 + j0$, $z(G) = 0 + j1$, $z(T) = 1 + j1$. Then if we consider a DNA sequence $(S_1, S_2, \ldots, S_n, \ldots, S_N)$ of N nucleotides, the CGR value along this sequence is defined by:

$$CGR(n) = \frac{1}{2}\,CGR(n-1) + \frac{1}{2}z(S_n) \ \ \text{for } n = 1, \ldots, N$$

Figure 1 illustrates this notion of CGR trajectory for sequence 'ATCGT' sequence.

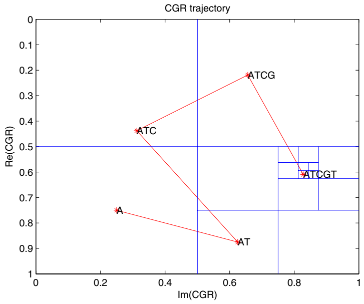


FIGURE 1 - Example of CGR trajectory for sequence 'ATCGT'

We can observe the fractal nature of this representation, since the last nucleotide determines the main quadrant of the (final) CGR position, the previous nucleotide one the sub-quadrant of that main quadrant and so on.

The $CGR(n)$ value can be written as a complex binary number

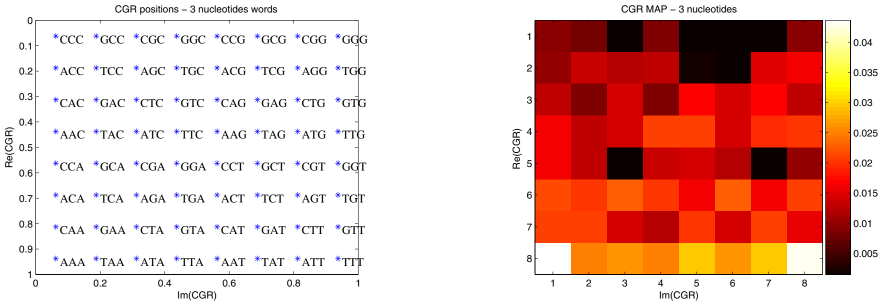
$$CGR(n) = c_1 2^{-1} + c_2 2^{-2} + \ldots + c_n 2^{-n} \ \left(+2^{-(n+1)}\right)$$

with=$c_i = b_i + jb'_i$. This value represents the complex binary code on $2n$ bits of the sequence until nucleotide S_n. The MSB $b_1 + jb'_1$ corresponds to the value $z(S_n)$ of the last nucleotide, while LSB $b_n + jb'_n$ corresponds to the value $z(S_1)$ of the first nucleotide. With a finite precision of k bits, the $CGR(n)$ value codes the "word" of k nucleotides ending by the nucleotide S_n. Each of the 4^k positions in the CGR square corresponds to one possible word. Hence the nucleotide sequence can be directly (and uniquely) recovered from its CGR transcription.

Figure 2 give the repartition of words in the case $k = 3$ of a CGR map, and the representation of a genomic sequence. Figure 2a gives the values $CGR(3)$ for all the words of k=3 nucleotides. Figure 2b represents the CGR(3) map for a genomic database (10 sequences of 100k nucleotides have been used for this representation). The CGR(3) map represents then the frequency matrix of the words of 3 nucleotides. We can observe that some words such as 'AAA' or 'TTT' are overrepresented and others (all the words including 'CG') are underrepresented.

Nonstationarity of the human genome.

Any sequence presents varying statistical properties along the sequence and this phenomenon is accessible through different characteristics such as the local (short term) mean of CGR or the local (short term) entropy. Different regions of sequence (genomic,



(a) CGR(3) values

(b) Example of CGR(3) map for genomic sequences

FIGURE 2 - CGR(3) values and example

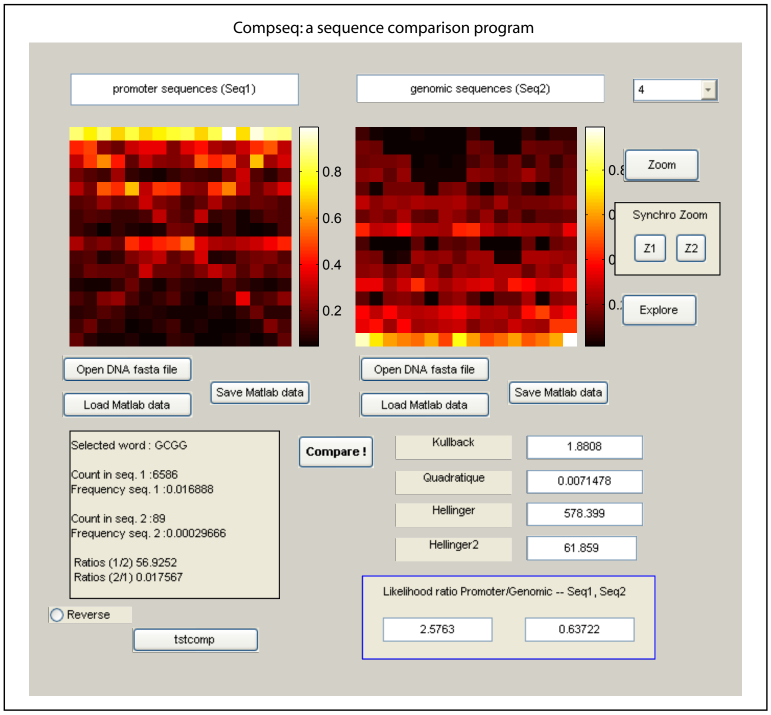
coding or promoter regions) in fact exhibit different global (long term) statistics. This is easily demonstrated by computing CGR maps on large available databases. We used the Eukaryotic Promoter Database [12] http://www.epd.isb-sib.ch/, that contains 1871 human promoter sequences, and a random extraction of sequences on the human genome. Figure 3 presents the result obtained with a software we developed. This clearly shows that promoter and genomic maps exhibit a different "signature", mainly because of the known and characteristic depletion in "CG" words observed in the human genome. However, it is important to note that that the high CG content is not a definitive discriminant feature since promoters from the EPD can be separated into a class with high CG content (1487 sequences which represent about 4/5 of the EPD sequences) and a second class of weak CG promoters (383 sequences or about 1/5 of the data base). We have then extracted from these classes two sets of test sequences representing 1/8 of each class and kept the others (7/8) as learning databases. CGR maps for promoters with high and low CG content are given in Figure 4.

Differences between the sequences can be quantified using several distances, as shown on the lower part of the figure 3, and sequences may be classified according to some similarity to a reference model, as will be discussed now.

NAÏVE BAYESIAN CLASSIFIER

In order to detect or predict promoter regions, we applied a simple naïve bayesian classifier. This kind of classifiers are based on probability models, derived using Bayes' theorem, that incorporate strong independence assumptions. These assumptions may often be obviously false, and the classifier deliberately naïve. Despite these simplifications, naïve Bayes classifiers often work much better in many complex real-world situations than might be expected, and the method proves successful and surprisingly efficient [13, 14]. The overall classifier seems robust enough to bypass deficiencies in its underlying naïve probability model.

The aim of the naïve Bayesian classifier is, given a sequence, S, to predict its most probable origin, and decide if the sequence belongs to a promoter 'Pro' class, or to the



Compseq: a sequence comparison program

promoter sequences (Seq1)

genomic sequences (Seq2)

4

Selected word : GCGG

Count in seq. 1 :6586
Frequency seq. 1 :0.016888

Count in seq. 2 :89
Frequency seq. 2 :0.00029666

Ratios (1/2) 56.9252
Ratios (2/1) 0.017567

○ Reverse

Kullback 1.8808

Quadratique 0.0071478

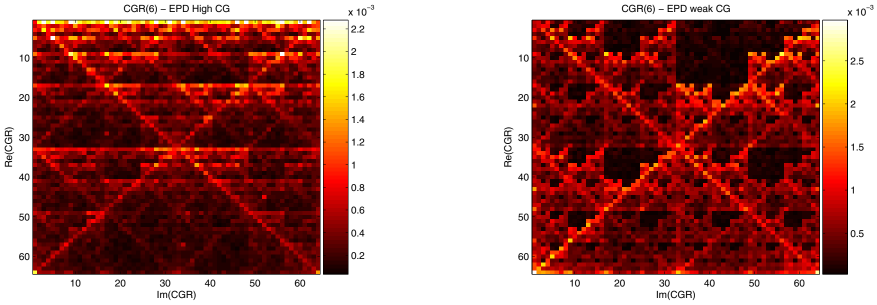
Hellinger 578.399

Hellinger2 61.859

Likelihood ratio Promoter/Genomic -- Seq1 , Seq2

2.5763 0.63722

FIGURE 3 - Output of CompSeq sofware – Comparison of CGR(4) maps for genomic and promoter sequences



(a) CGR(6) high CG promoter map

(b) CGR(6) low CG promoter map

FIGURE 4 - Comparison of CGR(6) maps for high and low CG content promoters.

genomic ‘Geno’ class. We consider a sequence of N nucleotides defined as a succession of $N-(k-1)$ overlapping words of length k, and simply express the probability of finding sequence S in a class C as the product of the $N-(k-1)$ probabilities of finding each word w_i in C:

$$P(S|C) = \prod_{i=1}^{N-(k-1)} P(w_i|C). \tag{1}$$

This assumes that the different words are independent of each other. This is clearly false, at least because of the overlapping between successive words (transition probabilities).

Bayes' rule enable to express the probability of a class C given the sequence S as

$$P(C|S) = \frac{P(S|C)P(C)}{P(S)}. \tag{2}$$

Therefore, we obtain the ratio of posterior probabilities for classes Pro and Geno as

$$\frac{P(Pro|S)}{P(Geno|S)} = \frac{P(Pro)}{P(Geno)} \frac{\prod_{i=1}^{N-(k-1)} P(w_i|Promo)}{\prod_{i=1}^{N-(k-1)} P(w_i|Geno)}, \tag{3}$$

that can be further factorized in

$$\frac{P(Pro|S)}{P(Geno|S)} = \frac{P(Pro)}{P(Geno)} \prod_{i=1}^{N-(k-1)} \frac{P(w_i|Promo)}{P(w_i|Geno)}, \tag{4}$$

thus involving the likelihood ratios $P(w_i|Promo)/P(w_i|Geno)$. Taking the logarithm, we have

$$\log \frac{P(Pro|S)}{P(Geno|S)} = \log \frac{P(Pro)}{P(Geno)} + \sum_{i=1}^{N-(k-1)} \frac{P(w_i|Promo)}{P(w_i|Geno)} \tag{5}$$

and the decision rule is simply

$$\log \frac{P(S|Pro)}{P(S|Geno)} = \sum_{i=1}^{N-(k-1)} \log \frac{P(w_i|Promo)}{P(w_i|Geno)} \underset{Geno}{\overset{Promo}{\gtrless}} \log \frac{P(Geno)}{P(Pro)} \tag{6}$$

Let us observe that the log-likelihood can be expressed as the difference between a Kullback-Leibler divergence and an entropy involving the empirical distribution:

$$\frac{1}{\bar{N}} \sum_{i=1}^{\bar{N}} \log P(w_i|Promo) = \sum_{j=1}^{W} \frac{n_j}{\bar{N}} \log \frac{P(w_i|Promo)}{n_j/\bar{N}} + \frac{n_j}{\bar{N}} \log \frac{n_j}{\bar{N}} \tag{7}$$

with $\bar{N} = N - (k-1)$, W the number of different words and n_j the count of a given word w_j. From (7), we recognize that

$$\frac{1}{\bar{N}} \log P(S|Promo) = D(\hat{P}(S)||P(S|Promo)) - H(\hat{P}(S)), \tag{8}$$

with $D(P||Q)$ the Kullback-Leibler divergence from P to Q, and $H(P)$ the Shannon entropy. Finally, we obtain that

$$\frac{1}{\bar{N}} \log \frac{P(S|Promo)}{P(S|Geno)} = D(\hat{P}(S)||P(S|Promo)) - D(\hat{P}(S)||P(S|Geno)). \tag{9}$$

Relation (6) gives a decision rule that can be implemented in order to classify sequences. The last relation (9) also indicates the interest of studying Kullback -Leibler divergences in this context. Table 1 reports the detection performance obtained with our databases (1/8 was reserved for evaluation) when testing sequences of length 600. HCG and LCG denotes High and Low CG content promoters respectively and FA stands for False Alarm. These results are clearly interesting for HCG promoters and more mitigated for LCG that are more difficulty to discriminate from the genomic background. The practical results can be improved, to a little extent and at a price of a higher complexity, by considering conditional probabilities or ‘nonstationary’ reference distributions.

TABLE 1 - Detection performances for Genomic (Geno), Promoters with high and low content, HCG and LCG respectively.

Test Class	Detection %	FA LCG %	FA HCG %	FA Geno %
LCG	55		22	23
HCG	95	5		0
Geno	75	20	5	

In order to ‘localize’ potential promoter sites, we designed an ‘adaptive’ version. we first estimate local probability distributions, either using a sliding window (typicaly of length 200) or using a forgetting factor. Then, we evaluate the log-likelihood ratio, the entropy and Kullback-Leibler divergences along the sequence, and therefore localize potential sites. In fact, so doing, we compare the distribution of the sequence at hand to reference distributions, the ‘genomic signatures’.

This is illustrated in Figure 5, where we explore a 960 Kb region of chromosome 7 including several annotated genes. We report the established mapping at GenBank, and study the log-likelihood ratio along the sequence (Kullback-Leibler divergences and entropy are not reported here to save space). Several unambiguous peaks emerge that correspond to the different genes. Furthermore, the LOC646531 predicted gene, that has no reported in vivo evidence, is also detected. Another peak preceding the SRPK2 Serine kinase gene has no direct correspondence with a known structure, so that is either a false detection or indicates a potential alternative promoter.

Hence, this shows the interest and usefulness of this approach that extends the notion of ‘genomic signature’. But of course, there are still areas of improvement; by incorporating biological knowledge (concensus sequences), improving the statistical model, and look for ‘long term’ dependence or geometric constraints.

REFERENCES

1. VB. Bajic, SL Tan, Y. Suzuki, S. Sugano S, *Promoter prediction analysis on the whole human genome*. Nat Biotechnol vol 22:1467-73, (2004)
2. H. J. Jeffrey, *Chaos Game Representation of gene structure*. Nucleic Acids Research. 18:2163-2170 (1990)
3. Deschavanne, P.J., Giron, A., Vilain, J., Fagot, G. and Fertil, B., *Genomic signature: characterization and classification of species assessed by chaos game representation of sequences*, Mol. Biol.Evol., 16(10):1391–1399, (1999)
4. J. Joseph, R. Sasikumar, *Chaos game representation for comparison of whole genomes*. BMC Bioinformatics vol 7:243, (2006)

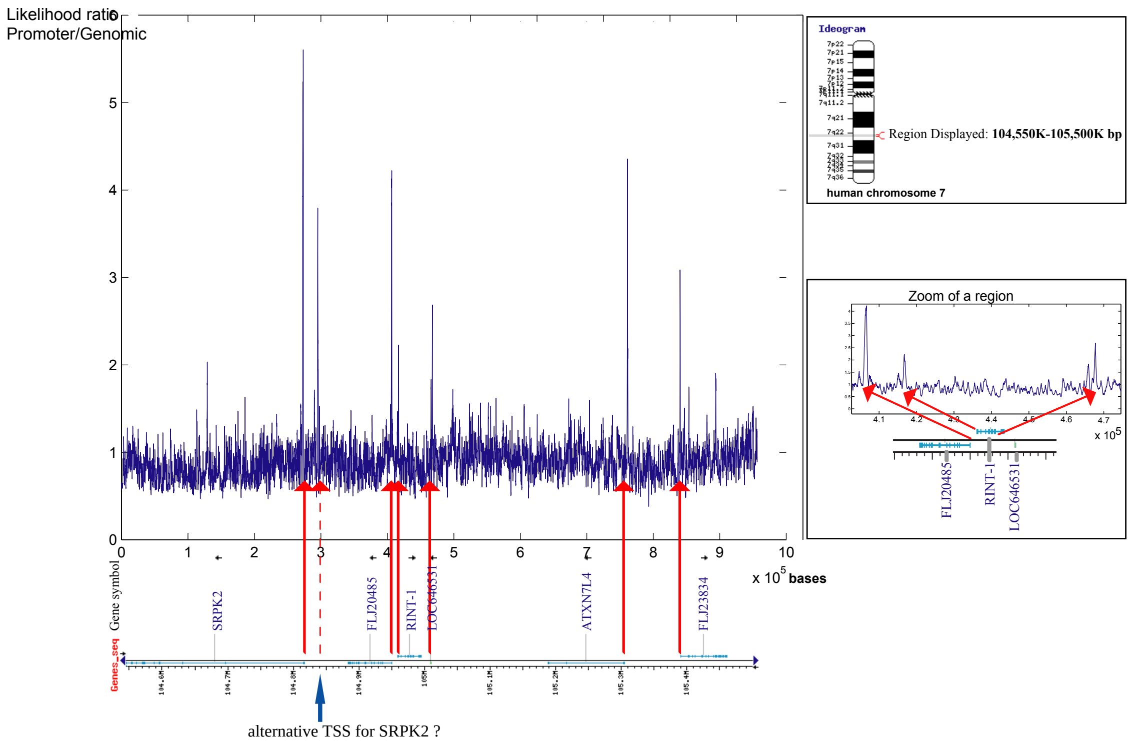


FIGURE 5 - Likelihood ratio Promoter/Genomic computed along the sequence, using a sliding window of length 200. Peaks reveal and correspond to known positions of promoters, assess predicted genes, and may indicate alternative promoters.

5. Karlin S, Campbell AM, Mrazek J., *Comparative DNA analysis across diverse genomes*, Annual Review of Genetics, 32:185-225 (1998)
6. R. Sandberg, G. Winberg, C.-I. Bränden, A. Kaske, I. Ernberg and J. Cöster, *Capturing Whole-Genome Characteristics in Short Sequences Using a Naïve Bayesian Classifier*, Genome research, Vol. 11, 8:1404-1409, (2001)
7. A.J. Gentles, S. Karlin, *Genome-scale compositional comparisons in eukaryotes*, Genome Research, 11(4):540-6 (2001)
8. S. Hannenhalli and S. Levy, *Promoter prediction in the human genome*, Bioinformatics, vol. 17, suppl. 1, pp S90-S96, (2001)
9. M.Q. Zhang, *Prediction*, *Annotation and Analysis of Human Promoters* - CSHL Quantitative Biology Symposium 68, pp. 217-225, (2003)
10. J.S. Almeida, J.A. Carriço, A Maretzk, P.A. Noble, M. Fletcher. “Analysis of genomic sequences by Chaos Game Representation”, Bioinformatics, Vol.17, 5:429-437, (2001)
11. J.S. Almeida, S. Vinga. “Universal sequence Map (USM) of arbitrary discrete sequences”, BMC Bioinformatics, (2002)
12. C.D. Schmid, R. Perier, V. Praz, P. Bucher, *EPD in its twentieth year: towards complete promoter coverage of selected model organisms*, Nucleic Acids Research, vol 34, D82-85 (2006)
13. P. Domingos and M. Pazzani, “On the optimality of the simple Bayesian classifier under zero-one loss” Machine Learning, 29:103–137, (1997)
14. D.J. Hand and K. Yu, “Idiot’s Bayes - not so stupid after all?” International Statistical Review, Vol 69:385-399, (2001)