Bayesian classification and entropy for promoter prediction in human DNA sequences

J.-F. Bercher [(1)], P. Jardin [(1)], B. Duriez [(2)]
(1) Équipe Signal et Information, ESIEE, France.
(2) INSERM-U654, Molecular and cellular bases of genetic diseases, Créteil, France.

February 27, 2006

There is now a large amount of genomic data available in databases for researchers. Computational methods are yet available for data retrieval and analysis, including sequences similarity searches, structural and functionnal predictions. Computationnal detection of genes has received an important interest and many accurate methods are available. However, other functionnal sites are more difficult to characterize.

In this work, we examine the potential of entropy and bayesian tools for promoter localization in human DNA sequences. Promoters are regulatory regions (at least one for each gene, located near the first exon) that governs the expression of genes, and their prediction is reputed difficult, so that this issue is still open.

To process DNA sequences it is useful to convert them using numerical representation that preserve their statistical properties. We choose the Chaos Game representation (CGR) [Jeffrey1990] of DNA sequences which has interesting properties: the source sequence can be recovered uniquely from the CGR transcription and the distance between CGR position measures similarity between corresponding sequences. This representation is applied to sequences of “words” of variable length (number of elementary bases). Typically we used words from 1 to 6 nucleotides. Using this CGR we have put in evidence the non stationarity 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 – *Transcription Start Site*, and the window length is less than 200 bases. Such a classifier has already be useful for classifying species as in [Sandberg2001].

Local probabilities were used to evaluate (i) the local entropy of the sequence, (ii) the Kullback divergence to the background (with respect to the hypothesis on the nature – genomic or promoter, of the background). Again, our experiments showed that these indicators clearly reveal the core-promoter and TSS positions in many cases. However, we also noticed, as was already pointed in the litterature [Hannenhalli2001,Zhang2003], that the set of promoters can be divided in (at least) two classes, the first one (with high CG ratio) being relatively easy to predict, while the second (that may in fact be divided in more subclasses) gives more mitigated results.

An interesting point is that a promoter prediction tool can assess or infirm the bioinformatic prediction of a gene. Such examples will be presented at the conference.

References

[Jeffrey1990] H. J. Jeffrey, *Chaos Game Representation of gene structure.* Nucleic Acids Research. 18:2163-2170, 1990.

[Sandberg2001] 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, Issue 8, 1404-1409, August 2001.

[Zhang2003] M.Q. Zhang, *Prediction, Annotation and Analysis of Human Promoters* - CSHL Quantitative Biology Symposium 68. pp 217-225 2003.

[Hannenhalli2001] S. Hannenhalli and S. Levy, *Promoter prediction in the human genome,* Bioinformatics, vol. 17, suppl. 1, pp S90-S96, 2001.