CpG island characterization

M. Ongenaert & W. Van Criekinge

Lab. Bioinformatics and Computational Genomics
Department of Molecular Biotechnology
Faculty of Bioscience Engineering
Ghent University
Belgium

Coupure Links 653
9000 Gent
Belgium
Tel: +32 9 264 59 69
Fax: +32 9 264 62 19
http://biobix.ugent.be
Wim.VanCriekinge@UGent.be

Introduction

During the development of cancer, the promoter regions of some genes are heavily methylated, blocking their transcription. This methylation happens almost always on cytosine bases in so called CpG islands. These are regions with a very high amount of CG dinucleotides, which can thus be densely methylated. In this study, we try to find properties of these CpG islands, which are specific for those regions that become methylated during the development of cancer. If this is possible, there is evidence that the sequence of the promoter region determines the difference between methylated in the development of cancer or not.

Sequences

We make use of two datasets: one with about 125 sequences of promoter regions that are described in various literature references to be methylated during the development of cancer (the positive list). The other list consists of promoter sequences of randomly chosen genes (the negative list). Sequences of the promoters are extracted form DBTSS (Ref. 1), and are from 9000 bp upstream to 3000 bp downstream from the transcriptional start site (TSS).

Finding CpG islands

For finding CpG islands in these regions, we use the following parameters in CpGIE (CpG Island Explorer, Ref. 2):

- minimal length=200 bp
- minimal C+G-content=55 %
- minimal observed number of CG dinucleotides over the expected number of CG (O/E ratio)= 0.60

Properties of CpG islands depend on their position

We check whether there is a difference in the properties of CpG islands (length, CG-content, O/E ratio) between CpG islands closer or further away from the TSS. We use CpG islands of the complete sequence (-9000 to +3000 from the TSS), as well as CpG islands that overlap (partially) 6 kb (-3000 to +3000 around TSS) and 1 kb (-500 to +500 around TSS) regions.

The closer to the transcription start, the higher the GC-content of the CpG islands is, and the longer the CpG islands are. The O/E-ratio stays about the same. CpG islands very close to the transcriptional start site can so be much more heavily en more densely methylated than CpG islands furher away from the TSS. So when the region very close to he TSS is methylated, this could have a great influence on the accessibility for transcription. Therefore, CpG islands 500 bp upstream to 500 bp downstream of the TSS are the most important CpG islands, they are the longest and most densely methylated and have the greatest influence on the accessibility of the gene for transcription.

In this study, no statistical significant difference of the properties has been observed between the positive and negative datasets: the properties of the CpG islands have no predictive value whether the promoter can be methylated during the development of cancer or not.

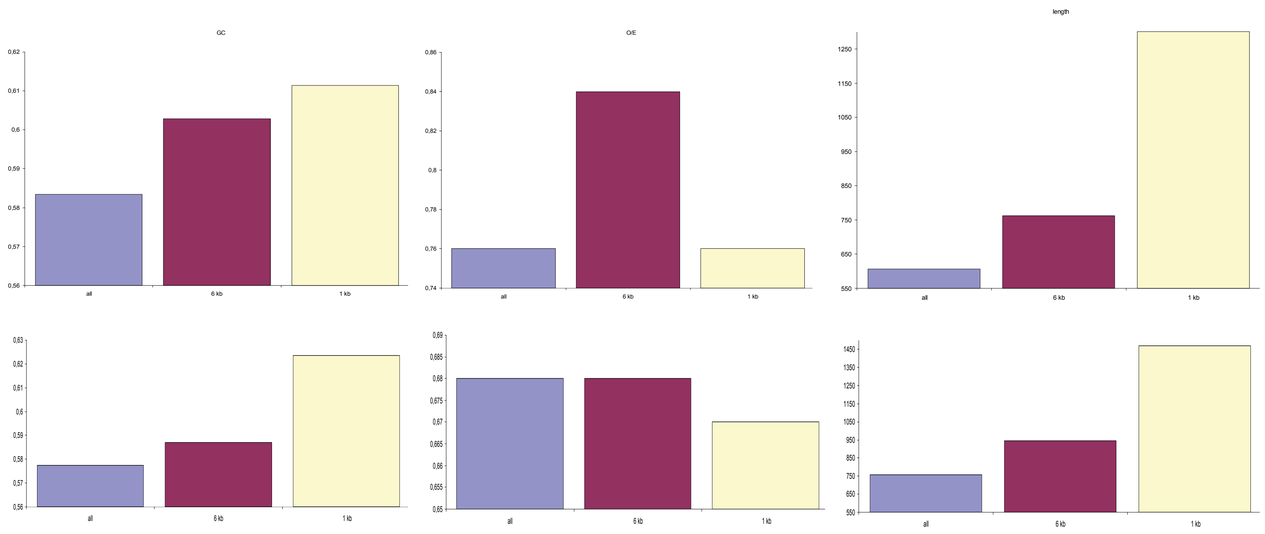

Figure 1: above: negative dataset; below: positive dataset. From left to right: GC-content, O/E ratio and length of the CpG islands. Blue: CpG islands in the region -9000 to +3000 from TSS; red: (partially) in -3000 to +3000; beige: (partially) in -500 to +500.

Binding sites for transcription factors

We try to distinguish both datasets, based on the difference of the number of bindingssites per kb for transcription factors in the CpG islands found in the region from -500 to +500 around the TSS. For finding these binding site for trancription factors, we use the Match™ algorithm (Ref. 3) that predicts binding sites for all transcription factors, listed in Transfac® (Ref. 3).

Classification shows that for about five transcription factors, the number of binding sites per kb has a certain predictive value in differentiating between the positive and negative datasets.

Disadvantage is the large variability of the number of binding sites of transcription factors. This is due to the fact that these binding sites are defined as a PSSM matrix. Also, Match™ makes use of the 'core match': 4 or 5 nucleotides that are the most important for binding of the transcription factor. This relative short core sequence can appear in the sequences just by chance.

Patterns

In stead of searching for binding sites of transcription factors, it is also possible to search for DNA patterns we generate ourselves. Patterns are generated using the Teiresias algorithm (Ref. 4), they have 7 fixed nucleotides and up to two wildcards (that stand for every nucleotide). Patterns must be present in at least 20 % of the sequences of either the positive or the negative dataset. This way, 7683 patterns are generated, for which we calculate the number of appearance per kb.

Classification only uses 10 patterns for predicting the class of a promoter region. The advantage of the use of patterns is that the number of appearance per kb is less variable than the number of binding sites for transcription factors. Also, the error made during the classification is smaller. The total error can be reduced up to 28 %.

Algorithm	Precision positive (%)	Error positive (%)	Precision negative (%)	Error negative(%)	Total error (%)
Background Zero R	0	100	52,7	0	47,3
Rules NNGE	68,0	34,6	69,9	27,7	31,0
Tree ADTree	67,2	27,1	73,6	31,9	29,6
Tree LMT	68,0	22,4	76,9	32,8	27,9
Meta Bagging	68,4	25,2	75,2	31,1	28,3
Lazy IB1	66,7	34,6	69,4	29,4	31,9
Functions Logistic	66,9	22,4	76,5	34,5	28,8
Functions SMO	72,6	43,0	67,6	19,3	30,5
Bayes BayesNet	50,0	46,3	55,9	52,1	47,3

Conclusion

Elements in the CpG islands within promoter sequences may determine whether the promoter can be silenced by methylation during the development of cancer or not. The most promising property of a CpG island to make this difference seems to be the number of some DNA patterns per kb. Disadvantage is that no biological function is known for these generated patterns, validation is needed.

References

1. Suzuki, Y, Yamashita, R, Sugano, S, Nakai, K. (2004). DBTSS, DataBase of Transcriptional Start Sites: progress report 2004. Nucleic Acids Research, 32, D78-81.
2. Wang, Y., Leung, F.C.C. (2004). An evaluation of new criteria for CpG islands in the human genome as gene markers. Bioinformatics, 7, 1170-1177.
3. http://www.gene-regulation.com
4. Rigoutsos, I., Floratos, A. (1998). Combinatorial Pattern Discovery In Biological Sequences: The TEIRESIAS Algorithm. Bioinformatics, 14, 229.