APPLICATION NOTE Genomics & Informatics Vol. 8(1) 58-61, March 2010

CpG Islands Detector: a Window-based CpG Island Search Tool

Ki-Bong Kim*

Department of Biomedical Technology, Sangmyung University, Cheonan 330-720, Korea

Abstract

CpG is the pair of nucleotides C and G, appearing successively, in this order, along one DNA strand. It is known that due to biochemical considerations CpG is relatively rare in most DNA sequences. However, in particular subsequences, which are a few hundred to a few thousand nucleotides long, the couple CpG is more frequent. These subsequences, called CpG islands, are known to appear in biologically more significant parts of the genome. The ability to identify CpG islands along a chromosome will therefore help us spot its more significant regions of interest, such as the promoters or 'start' regions of many genes. In this respect, I developed the CpG islands search tool, CpG Islands Detector, which was implemented in JAVA to be run on any platform. The window-based graphical user interface of CpG Islands Detector may facilitate the end user to employ this tool to pinpoint CpG islands in a genomic DNA sequence. In addition, this tool can be used to highlight potential genes in genomic sequences since CpG islands are very often found in the 5' regions of vertebrate genes.

Availability: The tool is available free of charge for non-commercial use only. Contact the corresponding author

Keywords: CpG island, promoter, CpG islands detector, JAVA, graphical user interface

Introduction

CpG dinucleotide is remarkably underrepresented in vertebrate genomes. In the human genome, for example, its frequency is five times less than statistically expected frequency on the basis of the nucleotide composition. The depletion of CpG dinucleotides is caused by the spontaneous deamination of methylated cytosines to yield thymine and generate a T:G mismatch that will be fixed as TpG (or CpA on the complementary strand) if the thymine is not repaired by cytosine before the next round of DNA replication (Lander *et al*., 2001; Sved and Bird, 1990; Venter *et al*., 2001). However, there are genomic regions rich in CpG dinucleotides, termed CpG islands, where the level of methylation is significantly lower than the overall genome. In these regions, the occurrence of CpGs is significantly higher, close to the expected frequency. The existing CpG islands are traditionally thought to be unmethylated. However, many CpG islands were subsequently found to be hypermethylated in the imprinted genes (Jones and Baylin, 2002). It is now known that some CpG islands in non-imprinted regions are also methylated in normal cells and this is believed to be related to tissue-restricted gene expression patterns (Grunau *et al*., 2000; Song *et al*., 2005). A large number of methylated CpG islands are also found in tumor cells (Jones and Baylin, 2002).

Most CpG islands are found at significant regions of interest, such as the promoters or 'start' regions of many genes. The CpG islands that are located upstream of the transcription start site are critical in gene expression regulation and cell differentiation (Bird, 2002). They are usually an important signature of the 5' region of many mammalian genes, often overlapping with, or within, a thousand bases downstream of the promoter. The identification of promoters by CpG islands with a resolution of 2 KB may be most useful for large-scale sequence annotation. Visual inspection of CpG islands is often used for gene identification by many molecular biologists. In a word, searching CpG islands are very important from various aspects. In this respect, I developed the CpG islands search tool, CpG Islands Detector, which can be used for CpG islands determination. This tool is a window-based Java application implemented with JBuilder 9.0 which is a Java IDE (Integrated Development Environment). There are several computer programs, including CpG Island Searcher (Takai and Jones, 2002) and CpGIF (Ye *et al*., 2008), which search genomes for CpG islands and are available on the Web. Their simple and limited interfaces mean that users are unable to capitalize on the programs by using them to find out the best parameters - parameters which would allow users to locate all of the CpG islands and none of the junk.

*Corresponding author: E-mail kbkim@smu.ac.kr
Tel +82-41-550-5377, Fax +82-41-550-5184
Accepted 5 February 2010

Overview and features of CpG Islands Detector

Search criteria and algorithm

The first large-scale computational analysis of CpG islands using vertebrate gene sequences in GenBank was performed by Gardiner-Garden and Frommer in 1987 (Gardiner-Garden and Frommmer, 1987). As defined by Gardiner-Garden and Frommer (Gardiner-Garden and Frommer, 1987), CpG islands are greater than 200 bp in length, have more than 50 percent of G+C content, and have a ratio of CpG frequency to the product of the C and G frequencies above 0.6. Over time, the criteria for a CpG island have evolved. Currently, the generally accepted criteria have become more stringent, requiring a minimum DNA sequence length of 500 bp. The importance of these criteria lies in that they are able to exclude most Alu repeats, which were identified as CpG islands by the old criteria. Takai and Jones proposed more stringent criteria, with G+C content and observed CpG/expected CpG ratio (Obs_{CpG}/Exp_{CpG}) increased to 55% and 0.65 respectively, which would be more effective in excluding *Alu* repeats (Takai and Jones, 2002). The criteria used in this work basically comply with those defined not only by Gardiner-Garden and Frommer but also by Takai and Jones.

In this work, a sliding 200 base pair window algorithm was fundamentally applied to the CpG Islands Detector. Even though the algorithm was designed according to the criteria of CpG islands described by Gardiner-Garden and Frommer, two more parameters - N_{CpG} and V_{gap} were introduced in this work. N_{CpG} and V_{gap} mean the number of CpGs in the 200 bp and the value of the gap between successively adjacent CpG islands respectively. N_{CpG} was used to exclude "mathematical CpG islands". A 200 bp sequence containing one G, 150?Cs, and only one CpG, which would meet the criteria of a CpG island, can become an example of "mathematical CpG islands". That is, there must be at least seven CpGs in 200 bp. This number was selected on the basis that there would be 200/16 (i.e. 12.5) CpGs in a random DNA fragment containing no suppression of CpG. Because Gardiner-Garden and Frommer's criterion of Obs_{CpG}/Exp_{CpG} of 0.6 would accommodate (0.6×12.5) CpGs (i.e., 7.5), seven CpGs were regarded as being a reasonable cutoff. V_{gap} was adopted to extend single stretch meeting the criteria of CpG island by means of merging two immediately adjacent CpG islands. The CpG search algorithm can be summarized as following (Fig. 1):

(A) Set a sliding window of 200 bp in the beginning of a Contig.
(B) Shift the window 1 bp after evaluation until the window meets the criteria (G+C %, Obs_{CpG}/Exp_{CpG}, and, N_{CpG}) of a CpG island.
(C) If the window meets the criteria, shift the window 200 bp, then evaluate again and repeat these 200 bp shifts until the window does not meet the criteria.
(D) If the last window in the step C does not meet the criteria, shift the last window 1 bp toward the 5' end until it meets the criteria.
(E) If the right end of the last window reaches the end of the Contig, go to the step *F*. Otherwise, set a 200-base window in the position where 3? end of the last window was located and repeat the step *B* through *D*.
(F) Two immediately adjacent CpG islands are connected if they are separated by less than V_{gap} bp and the total stretch to be connected meets the criteria of G+C % and Obs_{CpG}/Exp_{CpG}.

G+C % plotting module and graphical user interface

The nuclear genomes of vertebrates are mosaics of isochores, very long stretches (>300 kb) of DNA that are homogeneous in nucleotide composition and are compositionally correlated with the coding sequences that they embed. Isochores can be partitioned in a small number of families that cover a range of G+C levels, which is narrow in cold-blooded vertebrates, but broad in warm-blooded vertebrates. The G+C % plotting can give a clue to discriminating the promoter and coding

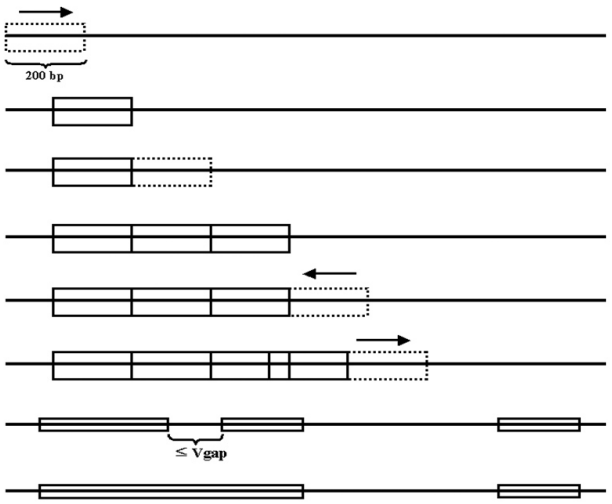

Fig. 1. Schematics for the algorithm of the CpG Islands Detector (refer to the text).

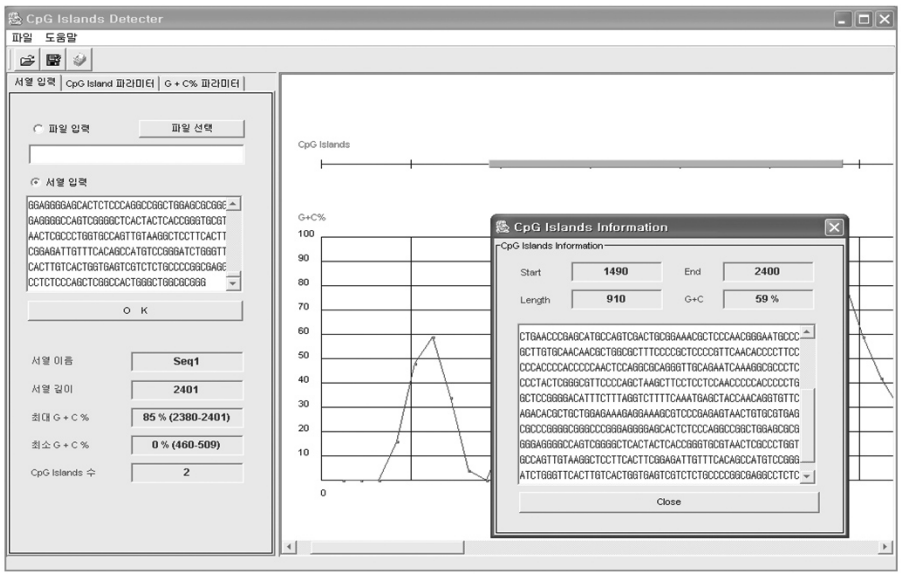

Fig. 2. Graphical user interface and analysis output screen shot of CpG Islands Detector.

regions from intron sequences. In this respect, the module for G+C % plotting was brought in CpG Islands Detector in order to facilitate the end user to discriminate promoter-associated CpG island from non-promoter associated CpG island. The module plots G+C content within a sliding window that steps along the sequences at a specified shift. Like GpG islands Detector, it also allows user-defined parameters (i.e., window size and step size) so that the user-defined window is moved down the sequence by specified step size. The G+C content is calculated and plotted at each central position that the window is moved to (Fig. 2).

Window-based graphical user interface enables users to change the preset default parameter values into the ones tailored to their analysis intent (i.e. variable limit of G+C %, Obs_{CpG}/Exp_{CpG}, *length*, N_{CpG}, and V_{gap}) (Fig. 2). The user can get the summary of analysis result on the left panel of graphical user interface which includes three main spreadsheets - sequence input, CpG island parameters and G+C % parameters. In addition, the user can get the detailed information (i.e., length, coordinate (start and end), G+C %) on a CpG island detected by CpG Islands Detector through pop-up window (front window in the Fig. 2) by double-clicking on the corresponding CpG island. The input sequence can be directly cut and pasted into the edit box or be uploaded from a file. Both of FASTA format and plain text format are allowed for the input sequence. The analysis output looks like the one in Fig. 2. The upper part of the analysis output displays the detected CpG islands and the lower part displays the G+C % plot. An example sequence (ID: EPI_7031) derived from EPD (Eukaryotic Promoter Database) was used to be analyzed in Fig. 2.

Discussion

Any definition of a CpG island is somewhat arbitrary but it is a recent trend that more stringent criteria are used for determination of CpG islands. While CpG islands meeting the stringent criteria are more likely to be associated with the 5' regions of genes and the criteria can exclude most *Alu*-repetitive elements, using the stringent criteria may deteriorate the sensitivity of detection. Considering all of these, search algorithm in this work was allowed to user defined lower limits of parameters for the initial scanning of a submitted sequence to avoid missing any CpG islands. Using a larger window size potentially allows the extraction of CpG islands which could not be extracted with a smaller window size. Since our first priority for this algorithm is not to miss any sequences meeting the criteria, the initial result might differ from the perception of the user. The user can pinpoint the CpG island region using a larger G+C %, Obs_{CpG}/Exp_{CpG} and smaller length after the initial search. Ultimately, using the parameters defined by the user, CpG Islands Detectors searches for all potential CpG islands and maps them on the input sequence graphically. Furthermore, users can get the related information such as parameter values and subsequence on a CpG island by double-clicking on it. In conclusion, this tool can facilitate users to pinpoint CpG islands in a genomic DNA sequence and also to highlight potential genes by means of promoter-associated CpG islands. However it will be needed additional studies on parameters optimization that can maximize sensitivity and specificity of this application.

References

Bird, A. (2002). DNA methylation patterns and epigenetic momory. *Genes Dev.* 16, 6-21.

Gardiner-Garden, M., and Frommer, M. (1987). CpG islands in vertebrate genomes. *J. Mol. Biol.* 196, 261-282.

Gruenbaum, Y., Stein, R., Cedar, H., and Razin, A. (1981). Methylation of CpG sequences in eukaryotic DNA. *FEBS lett.* 124, 67-71.

Jones, P.A., and Baylin, S.B. (2002). The fundamental role of epigenetic events in cancer. *Nat. Rev. Genet.* 3, 415-428.

Song, F., Smith, J.F., Kimura, M.T., Morrow, A.D., Matsuyama, T., Nagase, H., and Held, W.A. (2005). Association of tissue-specific differentially methylated regions (TDMs) with differential gene expression. *Proc. Natl. Acad. Sci. USA* 102, 3336-3341.

Sved, J., and Bird, A. (1990). The expected equilibrium of the CpG dinucleotide in vertebrate genomes under a mutation model. *Proc. Natl. Acad. Sci. USA* 87, 4692-4696.

Takai, D., and Jones, P.A. (2002). Comprehensive analysis of CpG islands in human chromosomes 21 and 22. *Proc. Natl. Acad. Sci. USA* 99, 3740-3745.

Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., Gocayne, J.D., Amanatides, P., Ballew, R.M., Huson, D.H., Wortman, J.R., Zhang, Q., Kodira, C.D., Zheng, X.H., Chen, L., Skupski, M., Subramanian, G., Thomas, P.D., Zhang, J., Gabor Miklos, G.L., Nelson, C., Broder, S., Clark, A.G., Nadeau, J., McKusick, V.A., Zinder, N., Levine, A.J., Roberts, R.J., Simon, M., Slayman, C., Hunkapiller, M., Bolanos, R., Delcher, A., Dew, I., Fasulo, D., Flanigan, M., Florea, L., Halpern, A., Hannenhalli, S., Kravitz, S., Levy, S., Mobarry, C., Reinert, K., Remington, K., Abu-Threideh, J., Beasley, E., Biddick, K., Bonazzi, V., Brandon, R., Cargill, M., Chandramouliswaran, I., Charlab, R., Chaturvedi, K., Deng, Z., Di Francesco, V., Dunn, P., Eilbeck, K., Evangelista, C., Gabrielian, A.E., Gan, W., Ge, W., Gong, F., Gu, Z., Guan, P., Heiman, T.J., Higgins, M.E., Ji, R.R., Ke, Z., Ketchum, K.A., Lai, Z., Lei, Y., Li, Z., Li, J., Liang, Y., Lin, X., Lu, F., Merkulov, G.V., Milshina, N., Moore, H.M., Naik, A.K., Narayan, V.A., Neelam, B., Nusskern, D., Rusch, D.B., Salzberg, S., Shao, W., Shue, B., Sun, J., Wang, Z., Wang, A., Wang, X., Wang, J., Wei, M., Wides, R., Xiao, C., Yan, C., Yao, A., Ye, J., Zhan, M., Zhang, W., Zhang, H., Zhao, Q., Zheng, L., Zhong, F., Zhong, W., Zhu, S., Zhao, S., Gilbert, D., Baumhueter, S., Spier, G., Carter, C., Cravchik, A., Woodage, T., Ali, F., An, H., Awe, A., Baldwin, D., Baden, H., Barnstead, M., Barrow, I., Beeson, K., Busam, D., Carver, A., Center, A., Cheng, M.L., Curry, L., Danaher, S., Davenport, L., Desilets, R., Dietz, S., Dodson, K., Doup, L., Ferriera, S., Garg, N., Gluecksmann, A., Hart, B., Haynes, J., Haynes, C., Heiner, C., Hladun, S., Hostin, D., Houck, J., Howland, T., Ibegwam, C., Johnson, J., Kalush, F., Kline, L., Koduru, S., Love, A., Mann, F., May, D., McCawley, S., McIntosh, T., McMullen, I., Moy, M., Moy, L., Murphy, B., Nelson, K., Pfannkoch, C., Pratts, E., Puri, V., Qureshi, H., Reardon, M., Rodriguez, R., Rogers, Y.H., Romblad, D., Ruhfel, B., Scott, R., Sitter, C., Smallwood, M., Stewart, E., Strong, R., Suh, E., Thomas, R., Tint, N.N., Tse, S., Vech, C., Wang, G., Wetter, J., Williams, S., Williams, M., Windsor, S., Winn-Deen, E., Wolfe, K., Zaveri, J., Zaveri, K., Abril, J.F., Guigó, R., Campbell, M.J., Sjolander, K.V., Karlak, B., Kejariwal, A., Mi, H., Lazareva, B., Hatton, T., Narechania, A., Diemer, K., Muruganujan, A., Guo, N., Sato, S., Bafna, V., Istrail, S., Lippert, R., Schwartz, R., Walenz, B., Yooseph, S., Allen, D., Basu, A., Baxendale, J., Blick, L., Caminha, M., Carnes-Stine, J., Caulk, P., Chiang, Y.H., Coyne, M., Dahlke, C., Mays, A., Dombroski, M., Donnelly, M., Ely, D., Esparham, S., Fosler, C., Gire, H., Glanowski, S., Glasser, K., Glodek, A., Gorokhov, M., Graham, K., Gropman, B., Harris, M., Heil, J., Henderson, S., Hoover, J., Jennings, D., Jordan, C., Jordan, J., Kasha, J., Kagan, L., Kraft, C., Levitsky, A., Lewis, M., Liu, X., Lopez, J., Ma, D., Majoros, W., McDaniel, J., Murphy, S., Newman, M., Nguyen, T., Nguyen, N., Nodell, M., Pan, S., Peck, J., Peterson, M., Rowe, W., Sanders, R., Scott, J., Simpson, M., Smith, T., Sprague, A., Stockwell, T., Turner, R., Venter, E., Wang, M., Wen, M., Wu, D., Wu, M., Xia, A., Zandieh, A., and Zhu, X. (2001). The sequence of the human genome. *Science* 291, 1304-1351.

Ye, S., Asai, A., and Yunkai, L. (2008). CpGIF: an algorithm for the identification of CpG islands. *Bioinformation* 2, 335-338.