BIOINFORMATICS APPLICATIONS NOTE Vol. 19 no. 1 2003
Pages 151–152

MAVG: locating non-overlapping maximum average segments in a given sequence

Yaw-Ling Lin[1], Xiaoqiu Huang[2], Tao Jiang[3] and Kun-Mao Chao[4,*]

[1]Department of Computer Science and Information Management, Providence University, Shalu 433, Taiwan, [2]Department of Computer Science, Iowa State University, Ames, IA 50011, USA, [3]Department of Computer Science, University of California, Riverside, CA 92521, USA and [4]Department of Computer Science and Information Engineering, National Taiwan University, Taipei 106, Taiwan

Received on April 7, 2002; revised on June 13, 2002; accepted on July 4, 2002

ABSTRACT

Summary: MAVG is a software tool for finding k non-overlapping maximum-average segments that are sufficiently long in a given sequence of real numbers, for any $k > 0$. It has applications in several areas of biomolecular sequence analysis including locating GC-rich regions and CpG islands in a genomic sequence, and annotating multiple sequence alignments.

Availability: http://iubio.bio.indiana.edu/soft/molbio/pattern/cpg_islands/

Contact: kmchao@csie.ntu.edu.tw

Given a sequence of real numbers and a lower bound L, Lin *et al.* (2002) recently proposed an empirically linear-time algorithm for finding a segment of length at least L with the maximum average. However, in sequence analysis, we often want to compute more than one segment (or subalignment) that has interesting characteristics. It is therefore desirable to have an algorithm that delivers k non-overlapping maximum-average segments of a given sequence of real numbers, for any fixed $k > 0$. In this work, we have implemented an algorithm for enumerating k maximum-average segments with lengths at least L, where L is given parameter, as a C program called MAVG. A pseudo-code for MAVG is shown in Figure 1. The program can actually be used to produce k non-overlapping segments with the largest or smallest average. It can also be used to report all segments of average at least the given cutoff.

Here, the proposed program utilizes the fact that the size of the search space diminishes as more maximum-average segments have been output. Let the length of the given sequence be n. If we assume that the maximum-average segments are more or less evenly distributed in the sequence, the total length of the sequences searched in the above would be about

$$n + 2(n/2) + 4(n/4) + \cdots + 2^{\log k - 1}(n/2^{\log k - 1}) = n \log k.$$

Therefore, the above program MAVG is expected to run in $O(n \log k)$ time on the average.

MAVG can be used to find those segments with the highest GC ratio (Huang, 1994). Specifically, each of nucleotides C and G is assigned a score of 1, and each of nucleotides A and T is assigned a score of 0. The maximum-average segments of the binary sequence correspond to segments with the highest GC ratio in the DNA sequence.

Stojanovic *et al.* (1999) gave methods for finding conserved regions by assigning a numerical score to each column of a multiple alignment and then looking for runs of columns with high cumulative scores. Since the assigned scores may be all positive (e.g. in the information content case), an alternative is to look for runs of sufficiently many columns in the multiple alignment with the maximum average score, which can be efficiently computed by MAVG.

In the following, we sketch an application of MAVG in locating CpG islands in a genomic sequence. They are typically a few hundred to a few thousand bases long. Though the widely accepted definition of what constitutes a CpG island was proposed by Gardiner-Garden and Frommer (1987), new definitions and methods for a CpG island are still in progress (Takai and Jones, 2002).

A Markov chain model was introduced by Durbin *et al.* (1998) to decide if a short DNA sequence comes from a CpG island or not. The model consists of a dinucleotide table, which, for each of the 16 different dinucleotides, gives the log likelihood ratio of the frequencies (scores) of the dinucleotide in CpG islands and in non-CpG regions. It is known that CpG islands and non-CpG regions can be better discriminated by using average scores than using raw scores. The histogram of the average scores of CpG islands and non-CpG regions in Durbin *et al.* (1998) shows that all non-CpG regions have average scores less than 0.1 and most of the CpG islands have average scores greater than 0.1.

*To whom correspondence should be addressed.

ALGORITHM **MAVG**(A, L, k)

Input: A real sequence $A = \langle a_1, a_2, \ldots, a_n \rangle$ and a lower bound L.
Output: k non-overlapping maximum-average segment of A of length at least L.
Step 1: Insert the sequence $A = \langle a_1, a_2, \ldots, a_n \rangle$ into priority queue Q. Let the output list $P \leftarrow \emptyset$.
Step 2: Repeat Step 3 to Step 5 until the number of outputs in P is k or Q is empty.
Step 3: Among all the sequences in Q, let I be the one with a segment J of length at least L and the largest average. Extract I from Q. Append J to the output list P.
Step 4: Cut J from I, and let the two resulting segments be I_1 and I_2.
Step 5: Insert the sequences I_1 and I_2 into Q if their lengths are not smaller than L.
Step 6: Output P.

Fig. 1. Finding k non-overlapping maximum-average segments in a given real sequence.

The input genomic sequence is converted into a sequence of real numbers using the dinucleotide table mentioned above. The core of a CpG island is defined as a region of length at least 250 bp with the maximum average score. The full extent of a CpG island is a longest region that does not contain any sufficiently long (i.e. 250 bp or longer) subregion with average score below the cutoff.

The parameter k in MAVG should be set sufficiently large so that the k best regions reported by MAVG contain at least one region of average score less than the cutoff. This guarantees that no region with average score above the cutoff is missed. To locate both the core and the full extent of each CpG island in the input genomic sequence, different values are used for the length parameter L. MAVG is run on the input sequence several times, each time with a different value of L. As CpG islands are at least a few hundred bases long, the minimum value for L is set to 250 bp. Then L is increased by multiples of 250 bp until no region of length greater than L and average score above the cutoff can be found.

The method described above was used to locate CpG islands in a human sequence of 222,930 bp (GenBank Accession No. U47924). Regions of scores greater than the cutoff were computed by MAVG with L set to multiples of 250 bp. The regions were partitioned into 21 clusters. For each cluster, a region corresponding to the full extent of the CpG island was determined and used as a representative of the cluster. The 21 representative regions were classified into three groups: regions in 5′ ends of genes, regions in other parts of genes, and regions outside genes. As a comparison, a popular existing program named NEWCPGSEEK (see http://www.hgmp.mrc.ac.uk) for finding CpG islands was also applied to the same human sequence, with default parameter values. A total of 545 regions were reported by NEWCPGSEEK in a second. To compare with the 21 regions produced by MAVG, 21 best regions were selected from the NEWCPGSEEK output with a score cutoff of 105. Table 1 shows the number of regions in each of the three groups. A careful examination suggests that the programs MAVG and NEWCPGSEEK are complementary in the sense that a combination of theirs may provide a more accurate predication of CpG islands.

Table 1. Number of regions in each of the three groups

Program	No. of regions	5′ ends of genes	Other parts of genes	Outside genes
MAVG	21	12	7	2
NEWCPGSEEK	21	10	5	6

ACKNOWLEDGEMENTS

We thank the referees for helpful comments, and Liang Ye for implementing the web interface. Y.-L. Lin was supported in part by grant NSC 89-2218-E-126-006, Taiwan. X. Huang was supported in part by NIH Grants R01 HG01502-05 and R01 HG01676-05 from NHGRI. T. Jiang was supported in part by NSF Grants CCR-9988353 and ITR-0085910. K.-M. Chao was supported in part by grant NSC 90-2213-E-010-003, Taiwan.

REFERENCES

Durbin,R., Eddy,S., Krogh,A. and Mitchison,G. (1998) *Biological Sequence Analysis*. Cambridge University Press.

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

Huang,X. (1994) An algorithm for identifying regions of a DNA sequence that satisfy a content requirement. *CABIOS*, **10**, 219–225.

Lin,Y.-L., Jiang,T. and Chao,K.-M. (2002) Efficient algorithms for locating the length-constrained heaviest segments, with application to biomolecular sequence analysis. *Journal of Computer and System Sciences, in press*.

Stojanovic,N., Florea,L., Riemer,C., Gumucio,D., Slightom,J., Goodman,M., Miller,W. and Hardison,R. (1999) Comparison of five methods for finding conserved sequences in multiple alignments of gene regulatory regions. *Nucleic Acids Res.*, **27**, 3899–3910.

Takai,D. and Jones,P.A. (2002) Comprehensive analysis of CpG islands in human chromosomes 21 and 22. *PNAS*, **99**, 3740–3745.