BIOINFORMATICS APPLICATIONS NOTE Vol. 17 no. 11 2001
Pages 1065–1066

Segmentation of long genomic sequences into domains with homogeneous composition with BASIO *software*

Vasily E. Ramensky[1,*], *Vsevolod Ju. Makeev*[1], *Mikhail A. Roytberg*[2] *and Vladimir G. Tumanyan*[1]

[1]*Engelhardt Institute of Molecular Biology, Vavilova, 32 Moscow 117984, Russia and* [2]*Institute of Mathematical Problems of Biology, Institutskaya, 4 Puschino, Moscow region 142292, Russia*

Received on December 13, 2000; revised on May 29, 2001; accepted on June 4, 2001

ABSTRACT

Summary: We present a software system *BASIO* that allows one to segment a sequence into regions with homogeneous nucleotide composition at a desired length scale. The system can work with arbitrary alphabet and therefore can be applied to various (e.g. protein) sequences. Several sequences of complete genomes of eukaryotes are used to demonstrate the efficiency of the software.

Availability: The *BASIO* suite is available for non-commercial users free of charge as a set of executables and accompanying segmentation scenarios from http://www.imb.ac.ru/compbio/basio. To obtain the source code, contact the authors.

Contact: basio@imb.ac.ru; ramensky@imb.ac.ru; makeev@imb.ac.ru

The compositional organization of DNA sequences is a long-discussed issue (Li, 1997). There have been a variety of methods developed, window algorithms (Fickett *et al.*, 1992), tree-splitting algorithms (Oliver *et al.*, 1999), Hidden Markov Models (Churchill, 1992; Peshkin and Gelfand, 1999), and Bayesian techniques (Liu and Lawrence, 1999) to assess sequence heterogeneity; however none of the existing implementations of segmentation algorithms is free from limitations. We have developed an approach, which combines the strong sides of all of the techniques mentioned above.

The sequence is considered as a series of statistically independent and compositionally homogeneous blocks of arbitrary length. The direct Bayesian marginal likelihood is calculated for each segment, and the overall likelihood is calculated as a product of segment marginal likelihoods:

$$L(\beta) = \prod_{i=1}^{k} L^{(i)}(\mathbf{n}^{(i)})\beta^{k-1}. \qquad (1)$$

Here, $\mathbf{n}^{(i)}$ is the vector of letter counts for the ith segment, $L^{(i)}$ is the marginal likelihood of ith segment (Liu and Lawrence, 1999), k is the number of segments for some configuration of boundaries, and β is the 'Border Insertion Penalty' (BIP) parameter. In our programs we use parameter $B = -\ln\beta \geqslant 0$, so that the greater B values correspond to greater segment lengths. We maximize for the maximal likelihood over all possible configurations of borders.

Thus in *BASIO* we do not require *a priori* information on the composition of the segments, the number of the compositional states and the lengths of possible segments. The single parameter the program operates with is the border insertion penalty (Makeev *et al.*, 2001). It is used both to prevent overfitting and to define the overall length scale of the segmentation. By manipulating this parameter one can study the domains-within-domains phenomenon (Fickett *et al.*, 1992).

For the given B, the configuration for which marginal likelihood (1) attains its global maximal value can be found via Vitterbi-like dynamic programming (Finkelstein and Roytberg, 1993; Durbin *et al.*, 1998; Ramensky *et al.*, 2000), with the time required proportional to the square of the sequence length. For a desktop computer the maximal practical length is about 50 000 bp. For a longer sequence, for instance a sequence of a complete genome, we developed a split-and-merge procedure for the approximate search of the optimal segmentation.

Boundaries in the segmentation can also be removed according to their weights, which are usually calculated during the segmentation procedure. For each boundary one can calculate such a weight using the partition function approach (Ramensky *et al.*, 2000) also known as a forward–backward algorithm (Durbin *et al.*, 1998).

This algorithm described in detail in Ramensky *et al.* (2000) and Makeev *et al.* (2001) has been implemented as a complex of six program modules, *BASIO*, written in

*To whom correspondence should be addressed.

portable ANSI C language. The core module, called *basio*, implements the computational algorithm itself, whereas the others (*split*, *control*, *filter*, *merge*, *report*) allow a user to manage the processing of the given sequence, evaluate the results, and present them in a convenient form. The *BASIO* system may process a sequence in an arbitrary alphabet. The modular structure of the software complex enables one to construct segmentation scenarios in which *BASIO* modules are invoked from Perl or another scripting language. The scenarios may provide the iterative processing of the sequence, which results in segmentations with different characteristic segment length scales.

Using *BASIO*, we have segmented several examples of sequences of entire eukaryotic chromosomes including *Saccharomyces cerevisae* chromosomes I, III and IV and *Plasmodium falciparum* chromosomes II and III. All sequences were processed in the four-letter alphabet in a uniform manner: a sequence was split into parts of 10 000 bp each with the 5000 bp overlap; the optimal segmentation with border insertion penalty $B = 3$ was found for each part and then borders with probabilities lower than 0.9999 were filtered out; the parts were merged and the optimal segmentation of merged sequence was performed with $B = 3$; again, filtration with the threshold 0.9999 was done. Average time required for the whole procedure is about several hours on a dual P-II 350 MHz computer.

It turns out that the longest segments in *P.falciparum* are the low complexity regions such as repeats (e.g. the 22 000 bp subtelomeric repeat in chromosome II). Genome of *Plasmodium* is rich with A and T (up to 80%), therefore many AT-rich strands are also shown up. We have found that the majority of segments longer than 500 bp with G + C content greater than 0.2 overlap significantly with long exons.

In yeast chromosomes the correlation of long segments with exons is weaker than in *L.major* and *P.falciparum*. However, we observed some correlation of long segments with genetic distance charts. This observation definitely deserves more thorough investigation.

REFERENCES

Churchill,G. (1992) Hidden Markov chains and the analysis of genome structure. *Comput. Chem.*, **16**, 107–115.

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

Fickett,J.W., Torney,D.C. and Wolf,D.R. (1992) Base compositional structure of genomes. *Genomics*, **13**, 1056–1064.

Finkelstein,A.V. and Roytberg,M.A. (1993) Computation of biopolymers: a general approach to different problems. *BioSystems*, **30**, 1–19.

Li,W. (1997) The study of correlation structures of DNA sequences: a critical review. *Comput. Chem.*, **21**, 257–271.

Liu,S.L. and Lawrence,C.E. (1999) Bayesian inference of biopolymer models. *Bioinformatics*, **15**, 38–52.

Makeev,V.Ju., Ramensky,V.E., Gelfand,M.S., Roytberg,M.A. and Tumanyan,V.G. (2001) Bayesian approach to DNA segmentation into regions with different average nucleotide composition. In Gascuel,O. (ed.), *Lecture Notes in Computer Science 2066*. pp. 54–73.

Oliver,J.L., Roman-Roldan,R., Perez,J. and Bernaola-Galvan,P. (1999) SEGMENT: identifying compositional domains in DNA sequences. *Bioinformatics*, **15**, 974–979.

Peshkin,L. and Gelfand,M.S. (1999) Segmentation of yeast DNA using hidden Markov models. *Bioinformatics*, **15**, 980–986.

Ramensky,V., Makeev,V., Roytberg,M. and Tumanyan,V. (2000) DNA segmentation through the Bayesian approach. *J. Comput. Biol.*, **7**, 215–231.