BASIO: A SOFTWARE SYSTEM FOR SEGMENTATION OF BIOLOGICAL SEQUENCES INTO DOMAINS WITH HOMOGENOUS COMPOSITION

[1]*Ramensky V.E., [1]Makeev V.Ju., [2]Roytberg M.A., [1]Tumanyan V.G.

[1]Engelhardt Institute of Molecular Biology, Moscow, Russia
[2]Institute of Mathematical Problems of Biology, Puschino, Russia
e-mail: ramensky@imb.ac.ru
*Corresponding author

Keywords: DNA, sequence, composition, domains

Resume

Motivation:

Assessing compositional organisation is an important step in DNA sequence analysis. Functionally important sequence regions are often biased in their local nucleotide composition from the average composition of the whole sequence, or separate regions with a more uniform composition. Besides that, many search tools use local sequence composition as a reference point in statistical tests; thus preliminary segmentation to the regions with uniform composition can improve performance of such tools.

Results:

We have developed the BASIO (Bayesian Approach to Sequence segmentatIOn) system, which allows one to segment the sequence into regions with homogenous nucleotide composition at different length scales. We consider a sequence as a series of independent random Bernoulli segments. A Bayesian estimator is used to calculate likelihood of a partition and of a boundary between segments. A parameter is set to control the length scale of resulting segmentation.

Availability:

The BASIO package is available free of charge as a set of executables for Windows 9x,NT, Linux and SGI Irix from http://www.imb.ac.ru/combio/basio. The source code can be obtained from authors on request.

Introduction

The heterogeneous composition of DNA sequences is a long-discussed issue. The organization of genomes is arranged over a wide range of length scales, and different functional regions are believed to be associated with the domains of a specific composition. Among other examples one can call middle range clusters in eukaryotic sequences, particularly GpC islands, open reading frames in budding yeast chromosomes, long G+C-richest regions in human genomes, which contain the major part of coding sequences. Heterogeneity of DNA sequences at least partially responsible for correlation found at many length scales. Assessing such correlation depends significantly on the statistical model of DNA sequence. Different statistical models of DNA were compared by for human and *E. coli* and the results in the two cases were remarkably similar. In this study it was demonstrated that a given local base composition tends to persist over a scale of at least kilobases or tens of kilobases. Thus a multidomain model, in which different parts of the genome are modelled by different stochastic processes, provides a reasonable first approximation.

Algorithms and implementation

A symbolic sequence over an alphabet Ω of L letters is considered as a series of segments. Each segment is characterised with counts n = $(n_1,...,n_L)$ and is modelled as a random series of the Bernoulli type. For each configuration tested for the optimum, the Bernoulli probabilities are estimated from the counts. The measure of the statistical homogeneity of the a block is its marginal likelihood, which reflects the overall probability of obtaining the given sequence in the two stage random process. First, the composition σ is picked up according to the prior distribution, and then the sequence is generated in the Bernoulli random process with the letter probabilities σ. If $p(\sigma)$ is the uniform distribution on the surface of the simplex S, then

$$P(\mathbf{n}) = \frac{(L-1)!}{(N+L-1)!} n_1!...n_L! \qquad (1)$$

Since we consider the segments as independent, the complete likelihood of the sequence segmentation into *k* segments with known boundary location writes

$$P = \prod_k P_k(\mathbf{n}_k) \,. \qquad (2)$$

This quantity is optimised over the set of all possible boundary configurations yielding the optimal segmentation. Since the total value of the optimization functional is the product of the values for the blocks, the dynamic programming can be used for the efficient optimization.

The optimal segmentation usually yields too short blocks. If one tries to study segmentation of a longer scale, one need to remove boundaries that separates segments with close composition. This can be done in a two ways. First, the problem of segmenting the sequence into the fixed number of segments can be studied via introducing an additional multiplier β^k in (2), which corresponds to assigning penalty to each boundary added. The other way is to consider all possible partitions that retain the particular boundary and to calculate the probability of this boundary using the partition function approach. The results on boundary filtration obtained via these two techniques agree for the majority of samples.

Acknowledhements

This study has been partially supported by Russian Human Genome Program 18/48 (V.E.R., V.Ju.M., V.G.T), and 99-0153-F(018) (M.A.R.), INTAS grant 99-1476, and MGRI project 244.