

Bayesian Approach to DNA Segmentation into Regions with Different Average Nucleotide Composition

Vsevolod Makeev

Engel'hard Institute of Molecular Biology, Moscow

October 7, 1999

Summary by Mireille Régnier

1. Biological Motivation

Local nucleotide composition, that is, the distribution of nucleotides A, C, G, T along a chromosome, is important for many biological issues. Moreover, local nucleotide composition is accounted for in many algorithms developed to search for different patterns in DNA sequences. We present a method of segmentation of nucleotide sequences into regions with different average composition. The sequence is modelled as a series of segments; within each segment the sequence is considered as a random Bernoulli process. The partition algorithm proceeds in two stages. In the first stage the optimal partition is found, which maximizes the overall product of marginal likelihoods computed for each segment and prevents segmentation into short segments. In the next stage, optimal boundaries are filtered, and segments with close compositions are merged. This allows us to study segments with the chosen length-scale.

2. Optimal Segmentation

2.1. Probabilistic formulation. A symbolic sequence over an alphabet Ω of V letters is considered as a series of segments. Each segment is modelled as a Bernoulli random sequence. Bernoulli probabilities are estimated from the vector $\mathbf{n} = (n_1, \dots, n_V)$ where n_j denotes the number of occurrences of the j th symbol in the segment. In the Bayesian approach [1] estimated parameters are random variables. The probability distribution of these random variables is estimated from the data by a bootstrapping approach. First, one assumes an initial probability distribution—the so-called prior distribution—that may be chosen rather arbitrarily. These probability distributions are re-estimated from the data using the Bayes formula. The results of Bayesian estimation are always some probability distributions of the estimated quantity. Bayesian and classical statistics, however, agree for large samples because Bayesian distributions converge to the maximal likelihood estimation for any reasonable prior distribution. Denote the set of letter probabilities (the segment composition) as $\sigma = (\theta_1, \dots, \theta_V)$ with $\sum_{k=1}^V \theta_k = 1$. The likelihood of the individual sequence is $L(\sigma) = \prod_{k=1}^V \theta_k^{n_k}$. Given a composition $\sigma = (\theta_1, \dots, \theta_V)$, one writes the probability density function $p(\sigma)$, with normalisation condition $\int p(\sigma) d\sigma = 1$.

One starts from some prior distribution $p(\sigma)$, say the uniform distribution on $\sum_k \theta_k = 1$. The composition σ of the Bernoulli random process is picked up according to this prior distribution, $p(\sigma)$. The estimated probability density function $p(\sigma/\mathbf{n})$ satisfies Bayes's theorem:

$$p(\sigma/\mathbf{n}) = \frac{L(\mathbf{n}/\sigma)p(\sigma)}{P(\mathbf{n})}$$

where $P(\mathbf{n}) = \int L(\mathbf{n}/\sigma)p(\sigma) d\sigma$. The normalisation constant $P(\mathbf{n})$ is called marginal likelihood [3]. It reflects the overall probability of the given sequence in the two stage random process. For a uniform prior distribution, one has:

$$P(\mathbf{n}) = \frac{(V-1)!}{(N+V-1)!} n_1! \dots n_V!$$

Surprisingly, this quantity is also obtained in a conceptually similar but different probabilistic model (G. Shaeffer, 1999). For a sequence of length N , the probability of this sequence in the shuffling procedure is computed. Numbers (n_1, \dots, n_V) are picked up according to uniform distribution. With the assumption that segments are independent, the complete likelihood of the sequence segmentation into k segments with known boundary location is:

$$P = \prod_k P_k(n_k).$$

This quantity is optimized over the set of all possible boundary configurations yielding the optimal segmentation.

2.2. Dynamic programming. The maximization algorithm is as follows. Consider a sequence $S = s_1 s_2 s_3 \dots s_N$ of length N , where $s_i \in \Omega$. For every segment $S(a, b) = s_a \dots s_b$, one introduces a weight $W(a, b)$: for example, $W(a, b)$ can be $\ln P(S(a, b))$. A segmentation R in m blocks is determined as a set of boundaries $R = \{k_0 = 0, k_1, \dots, k_{m-1}, k_m = N\}$, where k_i separates s_k and s_{k+1} . Its weight is:

$$F(R) = \sum_{j=1}^m W(k_{j-1} + 1, k_j).$$

For functions determined on the segmentations, one also uses another set of variables, the indicators of the boundary positions q_k , $1 \leq k \leq N$. By definition, $q_k = 1$ if there exists a segment boundary after the k th letter, otherwise it is 0. Below, we use the notations $F(R)$ and $F(q_1, \dots, q_k)$ indifferently. The segmentation R^* with maximal weight is computed in a recursive manner. Denote by $R^*(k)$ the optimal segmentation of the fragment $S(1, k)$, $1 \leq k \leq N$. $R^*(1)$ is trivial. When optimal segmentations $R^*(1), \dots, R^*(k-1)$ are known, the optimal segmentation $R^*(k)$ is found using the following recurrence expression:

$$(1) \quad F(R^*(k)) = \max_{0 \leq i \leq k-1} [F(R^*(i)) + W(i+1, k)],$$

with $F(R^*(0)) = 0$. This equation yields the algorithm. Since the segmentation $R^*(k)$ is built in time $O(k)$, the total time can be estimated as $O(N^2)$.

2.3. Fluctuations in local composition. It appears that segments in optimal segmentation are usually very short. Even a random uniform Bernoulli sequence is divided into many segments. More generally, when the sequence consists of several random homogeneous domains, the optimal segmentation may include many borders located within the domains. This phenomenon is due to statistical fluctuations of the local nucleotide composition in random sequences. Thus it is advantageous to extract boundaries, which separate long regions with different compositions from those that reflect statistical fluctuations. This can be done by penalizing those segmentations that contain more boundaries. The correct penalty choice was initially chosen from computer simulations.

3. Filtration of Boundaries

3.1. Partition function. To study the relative significance of a boundary, one can calculate a score, that reflects how the addition of this particular boundary influences weights of segmentations. Given the probability $\Pi(\mathbf{q})$ of each segmentation $\mathbf{q} = (q_1, \dots, q_N)$, one defines the partition function of the segmentations in a standard way [2] by summing the probabilities of all possible partitions:

$$(2) \quad Z(N) = \sum_{q_1, \dots, q_{N-1}} \Pi(q_1, \dots, q_{N-1})$$

With the partition function at hand, one can compute the probability of a boundary to be located after a particular letter k . One computes two partition functions for the regions to the left and to the right of this border, Z_L and Z_R respectively:

$$(3) \quad \Pi(k) = \frac{Z_L(k)Z_R(N-k)}{Z(N)}.$$

3.2. Dynamic programming. The partition function in (2) rewrites as follows [2]:

$$(4) \quad Z(N) = \sum_{q_1, \dots, q_{N-1}} e^{F(q_1, \dots, q_{N-1})}.$$

To compute the probability of a boundary after the letter k , we also need the partition functions of the segments to the left and to the right of this boundary, and recursive formulæ to compute $Z_L(k)$ and $Z_R(k)$ are analogous to (1). They are obtained through the formal substitution of operations. Summation is used instead of taking the maximum, and multiplication is used instead of summation [2]. Equation (1) becomes:

$$Z_L(k) = \sum_{j=0}^{k-1} e^{W(j+1, k-1)} Z_L(j),$$

$$Z_R(k) = \sum_{j=k}^N e^{W(k, j)} Z_R(j),$$

with boundary conditions $Z_L(0) = Z_R(N+1) = 1$ and $W(k-1, k) = W(N, N+1) = 0$. An obvious modification of dynamic programming calculates the partition function in the case when only the given set of boundaries is allowed.

3.3. Filtration strategy. For the best result one should combine calculation of optimal segmentation with filtration. At the first stage, an optimal segmentation is found. Then a cut-off value is chosen and all the boundaries with probabilities (3) lower than that cut-off value are removed. The resulting set of boundaries usually is not optimal in the sense that some boundaries can also be removed, yielding a configuration with a higher probability P . So an additional round of optimisation is performed, removing some boundaries. Iterations converge rapidly to the stable set of boundaries all of which have the partition function probabilities greater than the cut-off value.

Bibliography

- [1] Durbin (R.), Eddy (S.), Krogh (A.), and Mitchinson (G.). – *Biological sequences analysis: probabilistic models of protein and nucleic acids*. – Cambridge University Press, 1998.
- [2] Finkelstein (A. V.) and Roytberg (M. A.). – Computation of biopolymers: A general approach to different problems. *Biosystems*, vol. 30, 1993, pp. 1–19.
- [3] Liu (S. L.) and Lawrence (C. E.). – Bayesian inference of biopolymer models. *Bioinformatics*, vol. 15, 1999, pp. 38–52.