Analysis of Information Content for Biological Sequences*

Jian Zhang
EURANDOM, Den Dolech 2, 5612 AZ, Eindhoven, and
Department of Computational Molecular Biology
Max-Planck-Institute for Molecular Genetics
Ihnestrasse 73, D-14195 Berlin (Dahlem)

April 27, 2001

Abstract

Decomposing a biological sequence into modular domains is a basic prerequisite to identify functional units in biological molecules. The commonly used segmentation procedures usually have two steps: First, collect and align a set of sequences which are homologous to the target sequence; then parse this multiple alignment into several blocks and identify the functionally important ones by using a semi-automatic method, which combines manual analysis and expert knowledge. In this paper, we present a novel exploratory approach to parsing and analyzing the above multiple alignment. It is based on an analysis-of-variance (ANOVA) type decomposition of the sequence information content. Unlike the traditional change-point method, our approach takes into account not only the composition biases but also the overdispersion effects among the blocks. More generally, our approach provides a better way for judging some important residues in a protein. Our approach tested on the families of ribosomal proteins has a promising performance. Some subsets of residues critical to these proteins are found.

Key words: Information content, multiple alignment, analysis-of-variance, RNA-binding motifs.
Running head: Analysis of Information Content

Address for correspondence: Jian Zhang, EURANDOM, Den Dolech 2, 5612 AZ, Eindhoven, The Netherlands. E-mail: jzhang@euridice.tue.nl; FAX: +3140 247 8190; Phone:+3140 247 8113

1 Introduction

Multiple sequence alignment has now become a standard tool for finding conserved patterns in a set of biological sequences (see, e.g., Durbin et al, 1998; Baxevanis and Ouellette, 1998; Duret and Abdeddaim, 2000). In the case of DNA, such patterns could be binding sites for a protein or cis-regulatory elements (see, e.g., Hughes et al., 2000; Stormo and Fields, 1998). In proteins, these could be DNA- and RNA-binding sites (see, e.g., Casari, Sander, Valencia, 1995; Hofmann et al., 1999). Usually a biological sequence is made up of several segments of the same or different functions. Since not all of these segments evolve at the same rate, a very common situation occurs when some of these segments are well conserved across certain phylogenetic domains, whereas others are very divergent and full of gaps, such that positional homology cannot be precisely determined and multiple substitutions have erased the phylogenetic information. In such a situation, it is important to distinguish between conserved and variable regions of an alignment. This gives rise to the following generalized segmentation problem. Given N (≥ 1) sequences, say, $\mathrm{A}_1, \cdots, \mathrm{A}_N$, which are aligned as follows:

$$\begin{array}{lccccc} \mathrm{A}_1: & a_{11} & a_{12} & \ldots\ldots & a_{1s_0} \\ \mathrm{A}_2: & a_{21} & a_{22} & \ldots\ldots & a_{2s_0} \\ & \vdots & & & \\ \mathrm{A}_N: & a_{N1} & a_{N2} & \ldots\ldots & a_{Ns_0}, \end{array} \tag{1.1}$$

how can the above matrix be partitioned into several blocks in order to identify some specified common patterns in these sequences? Here a_{ij} represents a symbol from the set { A, C, G, T, –} (i.e.,four bases plus the gap) in the DNA case, or a symbol from the set {K, R, H, S, T, N, Q, D, E, A, V, I, L, M, F, Y, W, C, G, P, – } (i.e., twenty amino acids plus the gap) in the protein case. To avoid confusion, we will call these sets the alphabet sets later on. More generally, given a family of unaligned sequences, how can we align them so that we can efficiently partition these sequences? In this paper, focusing on a set of aligned sequences, we develop a new global segmentation method based on an ANOVA type decomposition of the information content (IC) of these sequences. The information content of a set of aligned DNA fragments without gaps is introduced by Berg and von Hippel (1987) and Schneider et al. (1986) as an estimate of average specificity of the DNA-binding protein directly from a collection of regulatory sites. Here we extend this concept to proteins. A main advantage of our procedure over other existing methods is that we can obtain not only an optimal segmentation, but also the composition biases and dispersions within and among these blocks. Then we can identify the important ones through comparing the normalized information contents of these blocks. To show this potential, we apply our procedure to the families of ribosomal proteins.

As our other contribution, we modified the Auger-Lawrence dynamic programming algorithm to solve the computational problem with a polynomial time effort (Auger and Lawrence, 1989). Thus we give an answer to whether the computation of the Jensen-Shannon divergence based segmentation is NP-complete (Román-Roldán et al., 1998; and Clote and Backofen, 2000).

2 Information content approach

2.1 General concept

Here we give a general concept of information content, which is suitable for the multiple alignment in (1.1) and for any subset of the columns $\{1, \cdots, s_0\}$ called $[1, s_0]$. Consider a subset of $[1, s_0]$, say B. If letting $k = 1, 2, \cdots, v_0$ represent the twenty amino acids ($v_0 = 20$) or four DNA bases ($v_0 = 4$), and $k = v_0 + 1$ stands for any gap, the IC of variation (ICV) for block B (i.e., the columns indexed by B) can be defined as

$$\mathrm{ICV}(B, \mathbf{q}(B)) = \sum_{j \in B} \sum_{k=1}^{v_0+1} f(k, j) \log_2 \frac{f(k, j)}{q_B(k)}$$

where $f(k, j)$ is the frequency of k in jth column, modified by the root N type pseudocount:

$$f(k, j) = \frac{Z(k, j) + \sqrt{N} p_0(k)}{N + \sqrt{N}}$$

with $\mathbf{p}_0$ being the vector of background probabilities (Lawrence et al., 1993); and $\mathbf{q}(B) = (q_B(1), \cdots, q_B(v_0 + 1))^T$ is the vector of the average alphabet frequencies in block B (i.e., $\sum_{j \in B} f(\cdot, j)/|B|$ where $|B|$ is the number of columns in B). $\mathrm{ICV}(B, \mathbf{q}(B))$ describes the variation within block B. To show the composition deviation between block B and the whole aligment, we define the following IC of bias (ICB) for B:

$$\mathrm{ICB}(B, \mathbf{q}([1, s_0])) = \sum_{k=1}^{v_0+1} q_B(k) \log_2 \frac{q_B(k)}{q_{[1,s_0]}(k)}.$$

Then we can write the bias-to-variation ratio of block B as

$$S(B) = \frac{|B| \mathrm{ICB}(B, \mathbf{q}([1, s_0]))}{\mathrm{ICV}(B, \mathbf{q}(B))}$$

and the normalized information content of block B as

$$\mathrm{AIC}_T(B) = \mathrm{ICB}(B, \mathbf{q}([1, s_0])) + \mathrm{ICV}(B, \mathbf{q}(B)/|B|.$$

Note that $\mathrm{AIC}_T(B)$, a measure of how important block B is, allows one to compare one block to another.

2.2 ANOVA decomposition

For any partition of $[1, s_0]$, say $[1, s_0] = \cup_t^m B_t, \quad B_t = [l_t + 1, l_{t+1}]$, for some integers l_t ($l_1 = 0, l_{m+1} = s_0$), $1 \le t \le m + 1$, it is directly to prove that the total information content, $\mathrm{ICV}([1, s_0], \mathbf{q}([1, s_0]))$, admits an ANOVA type decomposition:

$$\mathrm{ICV}([1, s_0], \mathbf{q}([1, s_0])) = \mathrm{IC}_W^{(m)} + \mathrm{IC}_B^{(m)} \tag{2.1}$$

where

$$\begin{aligned} \mathrm{IC}_W^{(m)} &= IC_W^{(m)}(\{B_t\}_m) = \sum_{t=1}^{m} \mathrm{ICV}(B_t, \mathbf{q}(B_t)), \\ \mathrm{IC}_B^{(m)} &= IC_B^{(m)}(\{B_t\}_m) = \sum_{t=1}^{m} |B_t| \mathrm{ICB}(B_t, \mathbf{q}([1, s_0])) \end{aligned} \tag{2.2}$$

and $|B_t|$ is the number of columns in B_t.

Like ANOVA, we have a very simple intuitive interpretation for (2.1): $\mathrm{IC}_B^{(m)}$ is the total IC difference among blocks, while $\mathrm{IC}_W^{(m)}$ the total IC spread within blocks. In the literature, $\mathrm{IC}_W^{(m)}$ is often ignored. We find that it may be useful in showing the possible overdispersions among these blocks (see the next section). As a main application of the above interpretation, we find the following segmentation procedure:

When the number of blocks is fixed, the best way to parse the alignment is to choose B_t, $1 \le t \le m$ to maxmize the total IC difference, that is, to maximize $\mathrm{IC}_B^{(m)}$ or equally to minimize $\mathrm{IC}_W^{(m)}$.

This interpretation also gives a motivation of defining the standardized composition difference between two successive blocks, say B_t, $t = s, s+1$, by

$$r(B_s, B_{s+1}) = \frac{\sum_{t=s}^{s+1} |B_t| \mathrm{ICB}(B_t, \mathbf{q}(B_s \cup B_{s+1}))}{\sum_{t=s}^{s+1} \mathrm{ICV}(B_t)/(|B_s| + |B_{s+1}|)}. \tag{2.3}$$

We can use $r(B_s, B_{s+1})$ in (2.3) to test whether the difference between two successive blocks B_s and B_{s+1} is significant.

There are two issues in the implementation of the above method. One is about the computation. The other is about the choice of the number of blocks. We discuss these issues in the following two subsections.

2.3 Algorithm

Obviously for a fixed number of blocks, say m, there are $\frac{s_0!}{m!(s_0-m)!}$ ways of parsing. So it is time-consuming to use the brutal force optimization. In the case of single DNA sequence, Román-Roldán et al. (1998) and Clote and Backofen (2000) even conjectured that this problem may be NP-complete. Fortunately, we find a variation of the Auger-Lawrence dynamic programming with $O(N^3 s_0)$ computational effort. Set $\mathrm{IC}_{B[1,s_0]}^{(m)} = \max_{\{B_t\}_m} \mathrm{IC}_B^{(m)}(\{B_t\}_m)$. Analogously, we define $\mathrm{IC}_{B[1,j]}^{(m)}$ for any $1 \le j \le s_0$. Then the mechanism behind this algorithm is shown by the following recursive forward formulae:

$$\begin{aligned} C_{i,j}^{(0)} &= (j-i+1) * \mathrm{ICB}([i,j], \mathbf{q}([1,s_0])), \quad 1 \le i < j; \\ C_{1,j}^{(k)} &= \max_{1 \le l \le j} \{C_{1,l}^{(k-1)} + C_{l+1,j}^{(0)}\}, \quad 1 \le j \le s_0; \\ \mathrm{IC}_{B[1,s_0]}^{(m)} &= C_{1,s_0}^{(m-1)}. \end{aligned} \tag{2.4}$$

After $\mathrm{IC}_{B[1,s_0]}^{(m)}$ is obtained, the associated partition can be constructed by the standard backforward tracking procedure of the dynamic programming.

Proof of (2.4): It suffices to show by induction on m that for any $1 \le j \le s_0$,

$$\mathrm{IC}_{B[1,j]}^{(m)} = C_{1,j}^{(m-1)}.$$

For $m = 1$, the assertion is obvious. Suppose the assertion is true for $m = n$. Then we show that it holds also for $m = n+1$. To this end, we note that for any partition, say $\{B_t\}_{n+1}$, of $[1, j]$, by the assumption for $m = n$,

$$\begin{aligned} &\sum_{t=1}^{n} |B_t| \mathrm{ICB}(B_t, q([1,s_0])) + |B_{n+1}| \mathrm{ICB}(B_{n+1}, q([1,s_0]))| \\ &\qquad \le \max_l \{C_{1,l}^{(n-1)} + C_{l+1,j}^{(0)}\} = C_{1,j}^{(n)}. \end{aligned}$$

This yields

$$\mathrm{IC}^{(n+1)}_{B[1,j]} \le C^{(n)}_{1,j}.$$

On the other hand, by definition, there are $1 = l^*_1 < \cdots < l^*_{n+2} = j$ such that

$$C^{(n)}_{1,j} = \sum_{t=1}^{n+1} C^{(0)}_{l^*_t, l^*_{t+1}} \le \mathrm{IC}^{(n+1)}_{B[1,j]}.$$

The proof is completed.

2.4 Choice of the number of blocks

The underlying number of blocks is determined by the complexity of the sequences under investigation. There are several ways for determining such a complexity. One is based on the biological knowledge. The others are based on certain loss functions. In the former we first select m using our biological knowledge. Then we make the optimal ANOVA type decompositions followed by identifying the conserved blocks. Finally we predict their roles. Some training samples and structural information seem useful in choosing m. However it is not very reliable because many protein domains are poorly annotated in the current protein sequence data bases (see, e.g. Gracy and Argos, 1998).

For the loss function based methods, we need to tackle the issue of the possible overdispersion among blocks. This is because the alphabet frequencies in some positions of many DNA and protein motifs are highly heterogenous. Such a phenomenon can be partially described by an overdispersion factor (see Lindsey, 1999). The degree of overdispersion in the model is directly related to the number of blocks in a partition and to the complexity of each block.

To highlight the above point, we consider the following change-point testing model (Li, 2001). Under the null hypothesis H_0, the columns in (1.1) have the same alphabet probabilities, say $\mathbf{p} = (p_1, \cdots, p_{v_0+1})^T$; whereas under the alternative hypothesis H_1, we have m blocks, Δ_t, $t = 1, \cdots, m$. The columns in these blocks have some different alphabet probabilities, say $\mathbf{p}^{(t)} = (p^{(t)}_1, \cdots, p^{(t)}_{v_0+1})^T$, $t = 1, \cdots, m$, respectively. Under H_0, the log-likelihood is of the form

$$l(\mathbf{p}|H_0) = \sum_{k=1}^{v_0+1} n_k \log p_k$$

where n_k is the count of k in the whole alignment. Accordingly, under H_1, the log-likelihood becomes

$$l(\mathbf{p}^{(1)}, \cdots, \mathbf{p}^{(m)}|H_1) = \sum_{t=1}^{m} \sum_{k=1}^{v_0+1} n^{(t)}_k \log p^{(t)}_k$$

where $n^{(t)}_k$ is the count of k in the t-th block. Then the corresponding log-likelihood ratio test statistic can be rewritten as

$$\max_{\mathbf{p}^{(1)}, \cdots, \mathbf{p}^{(m)}} l(\mathbf{p}^{(1)}, \cdots, \mathbf{p}^{(m)}|H_1) - \max_{\mathbf{p}} l(\mathbf{p}|H_0) = \mathrm{IC}^{(m)}_B.$$

Thus, compared with the ANOVA decomposition (2.1), the above testing model completely ignores $\mathrm{IC}^{(m)}_W$, the variability of individual positions within blocks. This shows why in some cases the above testing model is too restricted to extract the main features from these blocks. For example, if we are looking for the second

structures for a family of protein sequences, it is obviously unreasonable to assume the homogeneity of the amino acid frequencies across the corresponding region in these sequences.

In what follows, we introduce two loss function based approaches for choosing the number of blocks. Compared with the second approach, the first one has a slightly better performance in terms of $\mathrm{IC}_B^{(m)}$ while performing a little worse in terms of resolution.

Method 1: For several α (e.g., $0.2, 0.25, 0.3, 0.4$), minimize the following modified Bayesian information criterion (BIC)

$$\mathrm{IC}_B^{(m)} - N^\alpha * (m-1)$$

with respect to m, where α reflects the roughness of the optimal partition. For the ribosomal protein families, in most cases using $\alpha = 0.3$ can nearly recover the signatures presented in the data base PROSITE. Note that this constant is slightly larger than that used by Braun, Braun and Müller (2000) in the single DNA sequence case because of the overdispersion effect.

Method 2: Use a resolution criterion (RC) in which we choose B_t, $1 \le t \le m$ so that the differences between the successive blocks are not less than a prespecified constant c_0, i.e., $r(B_t, B_{t+1}) \ge c_0$, $\quad 1 \le t \le m$. Here c_0 shows the roughness of the resulting partition. To adapt the above dynamic programming to this new situation, we slightly modify (2.4) by setting

$$\begin{aligned}
&C_{i,j}^{(0)} = (j-i+1) * \mathrm{ICB}([i,j], \mathbf{q}([1,s_0])), \quad 1 \le i < j;\\
&\Phi_{1j} = \{l : r([1,l],[l+1,j]) \ge c_0\};\\
&C_{1,j}^{(1)}(c_0) = \max_{l \in \Phi_{1j}} \{C_{1,l}^{(0)} + C_{l+1,j}^{(0)}\}, \quad 1 \le j \le s_0.
\end{aligned}$$

For $k \ge 1$, let $1 = l_{1k} < \cdots < l_{(k+1)k} = l$ be the boundaries of the partition induced by $C_{1,l}^{(k-1)}(c_0)$. Then for $k = 2, \cdots, m$, we iteratively define

$$\Phi_{kj} = \{l : r([l_{(k-1)k}+1, l], [l+1, j]) \ge c_0\}$$

and

$$\begin{aligned}
&C_{1,j}^{(k)}(c_0) = \max_{l \in \Phi_{kj}} \{C_{1,l}^{(k-1)}(c_0) + C_{l+1,j}^{(0)}\}, \quad 1 \le j \le s_0;\\
&\mathrm{IC}_{B[1,s_0]}^{(m)} = C_{1,s_0}^{(m)}(c_0)
\end{aligned}$$

where $\mathrm{IC}_{B[1,s_0]}^{(m)}$ is a modified version of that in Subsection 2.3.

How to choose c_0? Observe that without overdispersion, $r(B_i, B_{i+1})$ is approximately χ^2 distributed with a degree of freedom v_0 for the large blocks, which has the 0.005 quantitle 2.00 when $v_0 = 20$. Taking the possible overdispersion into account, we set $c_0 = 2c_1$ for protein sequences where c_1 is used to reflect the overdispersion effect allowed among the blocks. For instance, in the ribosomal protein case, we can choose several values for c_1, say $1.75, 2$, and 2.25. Note that Agalarov et al. (2000) have shown that some ribosomal proteins like S18 may have the multiple functions: RNA-binding and protein-protein interaction. We use $c_1 = 1.65$ to find the composition domains for these functions (see Table 3.6).

In light of the above arguments, we suggest the following strategy in practice: Begin with moderate α and c_0 to find functional domains. Then take low α and c_0 to localize functionally important subsets of residues within these domains.

3 Examples

In this subsection, we evaluate our approaches on the two kinds of ribosomal protein families: small-subunit families and large-subunit families, designated S1, S2, $\cdots$, and L1, L2, $\cdots$, respectively.

Ribosomal proteins, extremely ancient molecules, are windows into the protein evolution. The recent studies showed that there are some strong similarities between the binding-patterns (or structures) of RNA- and DNA-binding proteins (see, e.g., Draper and Reynaldo, 1999). In particular, several binding strategies used by DNA-binding proteins are found in those for ribosomal proteins. The key step in finding the RNA-binding motifs is to form ribosomal protein families with a certain degree of diversity and homology across certain phylogenetic domains. Wong and Zhang (1999) collected all the known ribosomal proteins from the following model organisms. **Archaebacteria:** Archaeoglobus fulgidus, Methanobacterium thermoautotrophicum, Methanococcus jannaschii, Pyrococcus horikosshii. **Eubacteria:** Aquifex aeolicus, Bacillus subtilis, Mycoplasma genitalium, Mycoplasma pneumoniae, Mycobacterium tuberculosis, Borrelia burgdorferi, Treponema pallidum subsp. pallidum, Chlamydia trachomatis, Escherichia coli, Haemophilus influenzae Rd, Helicobacter pylori, Synechocystis PCC6803, Thermus thermophilus. **Eukaryotes:** Saccaromyces cerevisiae, Caenorhabditis elegans, Rat, Drosophila melanogaster. They grouped them into about one hundred families by pairwise alignment. The multiple alignments for these families are available from Jian Zhang. The motifs of these protein families are identified by using the Gibbs motif sampler MACAW (Baxevanis and Quellette, 1998) and the iterative masking. According to the current structural or biochemical information on some sites of these motifs, they found that almost all the most conserved motifs based on MACAW are located in the putative RNA-binding domains. However it is difficult to determine the boundaries of these motifs. Here we first make multiple alignments for all these families by means of CLUSTAL W with default setting (Thompson, Higgins and Gibson, 1994). Without loss of generality we use the uniform background probabilities, namely $\mathbf{p}_0 = (1/21, \cdots, 1/21)$, because it doesn't affect the optimal partition. Then we analyze these alignments by our new approach. Compared with MACAW, our approach gives not only a better boundary resolution but also composition biases and variations for these motifs. As examples, we present these analyses for the L2, S15, and S18 families in Tables 3.1 to 3.6. These families have ICB$([1, s_0], \mathbf{p}_0)$ values of 0.346, 0.374 and 0.250, respectively. Some similar results for the other families are available from the author on request.

In these tables, the i and l_i columns give the indices and right boundaries (i.e., locations of change points) of blocks B_i, $i = 1, \cdots, m$ in the optimal partition. The AIC$_{Ti}$, S_i and r_i columns show the total average information content, bias-to-variation ratio and resolution for each block, respectively. Here we write AIC$_{Ti}$ = AIC$_T(B_i)$, $S_i = S(B_i)$ and $r_i = r(B_i, B_{i+1})$. As pointed out before, AIC$_{Ti}$ allows one to compare one block to another. Our experience confirms that those blocks with a relatively higher S_i are often very divergent and full of gaps or are singletons, whereas those blocks with a moderate $S_i (\geq 1)$ or a relatively higher AIC$_{Ti}$ are often corresponding to some important regions.

We adopt the following procedure for summarizing the information. We first classify the blocks with $S_i \geq 1$ into two groups according to whether they are gap blocks (i.e., more than half of which are gaps) or not. For example, for the L2 family, we classify blocks 2, 13, 17, 27, and 36 in Table 3.2 as gap blocks and assign blocks 6, 7, 10, 19, 23, 29, 30, and 31 group ^{a} of non-gap blocks. The next step is to classify the blocks with $S_i < 1$. For example, we select those blocks, which have a AIC$_{Ti}$ value larger than 25% of the highest among these blocks, to form group ^{b}. We call them potentially important blocks for a further analysis.

For L2, we use both Methods 2 and 1. By using Method 2 and the experimental results of Nakagawa et al. (1999), we find that in Table 3.1 the conserved blocks 3, 4, 6 are located in the N-terminal RNA-binding domain while the conserved block 8 is in the C-terminal RNA-binding domain (Figure 3.1). The functions of two conserved blocks 9 and 10 remain to be determined. Table 3.2 further indicates several potentially important sub-blocks (sites) within these domains: (1) from positions 75 to 84; (2) from positions 101 to 116; (3) from positions 181 to 198; (4) from positions 211 to 213; (5) from positions 241 to 251; (6) from positions 257 to 276. The blocks in regions (1) to (5) are the potentially important residue subsets in the binding domains just mentioned, while the blocks in region (6) are the potentially important residue subsets for block 10 in Table 3.1.

[Figure 3.1 here.]

[Table 3.1 here.]

[Table 3.2 here.]

For S15, from Table 3.3, using Method 2 with $c_0 = 4.5$ we identify three potentially important blocks: blocks 3, 5, and 7. These blocks are located in the RNA-binding domain. Furthermore, using Method 2 with $c_0 = 3.5$, we find five very informative subsets of residues in this domain: blocks 6, 16, 18, 20, and 31 in Table 3.4 (Figure 3.2). Interestingly, they are all functionally important because they serve as the S15-rRNA interfaces (see Agalarov et al., 2000).

[Figure 3.2 here.]

[Table 3.3 here.]

[Table 3.4 here.]

For S18, from Table 3.6, on the basis of the experimental results of Agalarov et al. (2000), we find that blocks 7 and 9, which have several hydrophobic positions, are putative regions for the interactions between S6 and S18, while block 8 is the putative region for the RNA-binding (Figure 3.3). The role of block 6 has not been determined yet although it has the second largest AIC_{Ti} value. It may be a RNA-binding site because it has two very conserved hydrophilic positions.

[Figure 3.3 here.]

[Table 3.5 here.]

[Table 3.6 here.]

4 Discussions and conclusions

4.1 Possible extensions

The concept of information content has been shown to be a very useful objective function for discovering regulatory sites or binding motifs in co-regulated genes (see, e.g., Heumann et al., 1994). But in other settings, we need to exploit some special pattern features. As an example, we modify the above concept for the protein coding region recognition by utilizing its nonuniform codon usage feature: each nucleotide has its own phase because the probability of appearance of a nucleotide is different in each of the three positions of the triplets (see, e.g., Grantham et al., 1981). Since this feature is not present in noncoding regions, it can be used to distinguish coding from noncoding. To this end, for $u = 1, 2, 3$, and for $[1, s_0]$, we set the

phase set

$$[1, s_0]^{(u)} = \{v : v \in [1, s_0], v \equiv u \mod 3\};$$

and let $\mathbf{q}^{(u)} = (q^{(u)}(1), \cdots, q^{(u)}(k))^T$ with

$$q^{(u)}(k) = \frac{1}{|[1, s_0]^{(u)}|} \sum_{j \in [1, s_0]^{(u)}} f(k, j),$$

where $1 \leq k \leq v_0 + 1$ and $|[1, s_0]^{(u)}|$ is the size of $[1, s_0]^{(u)}$. Then the modified information contents are defined as follows:

$$\begin{aligned} \mathrm{ICV}([1, s_0], \mathbf{q}([1, s_0])) &= \sum_{u=1}^{3} \sum_{j \in [1, s_0]^{(u)}} \sum_{k=1}^{5} f(k, j) \log_2 \frac{f(k, j)}{q^{(u)}(k)}, \qquad (4.1) \\ \mathrm{ICB}([1, s_0], \mathbf{p}_0) &= \sum_{u=1}^{3} \sum_{k=1}^{5} \log_2 \frac{q^{(u)}(k)}{p_0(k)} \end{aligned}$$

where $\mathbf{q}([1, s_0]) = (\mathbf{q}^{(1)T}, \mathbf{q}^{(2)T}, \mathbf{q}^{(3)T})^T$.

Similarly, to establish an ANOVA decomposition for the modified IC in (4.1), we consider only those B_t of the forms $[v_l+1, v_r]$ for some integers v_l and v_r. Let the phase set $B_t^{(u)} = \{v : v \in \mathrm{B}_t, v - v_l \equiv u \mod 3\}$, and write

$$q_t^{(u)}(k) = \frac{1}{|B_t^{(u)}|} \sum_{j \in B_t^{(u)}} f(k, j)$$

where $1 \leq k \leq v_0 + 1$, $u = 1, 2, 3$; and modify the summands in (2.2) by letting

$$\begin{aligned} \mathrm{ICV}(B_t, \mathbf{q}(B_t)) &= \sum_{u=1}^{3} \sum_{j \in \Delta_t^{(u)}} \sum_{k=1}^{v_0+1} f(k, j) \log_2 \frac{f(k, j)}{q_t^{(u)}(k)}, \\ \mathrm{ICB}(B_t, \mathbf{q}([1, s_0])) &= \sum_{u=1}^{3} |\Delta_t^{(u)}| \sum_{k=1}^{v_0+1} q_t^{(u)}(k) \log_2 \frac{q_t^{(u)}(k)}{q^{(u)}(k)} \end{aligned}$$

where as (4.1) $\mathbf{q}(B_t)$ and $\mathbf{q}([1, s_0])$ denote the vectors of the average alphabet frequencies in the phase sets $B_t^{(u)}$, $u = 1, 2, 3$ and $[1, s_0]^{(u)}$, $u = 1, 2, 3$, respectively. Then the decomposition formula (2.1) still holds.

The problem we investigated in the previous sections can be viewed as that of clustering a set of ordered objects (i.e., columns) which are characterized by some probabilistic vectors (i .e., the alphabet frequency vectors). We will show elsewhere that our method is even useful in building a phylogenetic tree for a set of sequences.

4.2 Relation to the other methods

For a single DNA sequence, our method is equivalent to a generalization of the Jensen-Shannon divergence based segmentation method (Oliver et al., 1999) except that we take into account the possible overdispersion effect on the choice of the significance level. Here the overdispersion means that in reality there is greater variability among the columns in (1.1) than would be expected from a statistical model (e.g., product multinomial model), because we can not expect each domain to be completely homogeneous. There are many different segmentation methods in literature. See the recent review by Braun and Müller (1998). All these methods, though effective in detecting a long pattern (e.g., a coding region), are not very useful for

detecting a short pattern like a cis-regulatory element. In order to identify these short patterns, we usually need to collect and align a set of similar DNA sequences (or fragments) across the different species (André et al., 2001). Then these patterns can be found by parsing this multiple alignment. In the case of protein, Liu and Lawrence (1999) introduced a Bayesian global segmentation model. However, for other multiple alignment procedures (for instance, Clustal W or profile HMM), there is no general approach to identifying several patterns simultaneously. The traditional methods are based on a moving window with iterative maskings (Hughes et al., 2000). The length of a local window or the boundaries of these patterns are often determined in an adhoc way. For example, we often need to specify the bandwidth of the Gibbs motif sampler (Liu, Neuwald and Lawrence, 1995; Baxevanis and Ouellette,1998) using our experience. Auger and Lawrence (1989) presented a nice discussion on the advantages of global methods over the moving window methods. Finally, we note that our method can be integrated into evolutionary trace analysis for identifying structual features within a protein (Landgraf et al., 1999).

4.3 Conclusions

We have introduced a general concept of information content for a set of aligned sequences. We have presented an ANOVA based information content method for predicting functionally important motifs in both DNA and proteins. A dynamic programming algorithm has been modified for solving the computational problem in parsing a multiple alignment. We have evaluated our method on the ribosomal protein families. Some new motifs related to the interaction between the ribosomal proteins and ribosomal RNA have been recovered.

The major shortcoming of our procedure is the prerequisite of a valid alignment of the input sequences. This could be difficult sometimes, especially, for the genomic sequences. Although it seems possible to develop some procedure for aligning and parsing the input sequences in an iterative way on the basis of certain ANOVA decompositions of the sequence information content, the computational time for such a task turns out prohibitively long.

Acknowlegements. The author has greatly benefited from several discussions with Professor Martin Vingron, Max-Planck Institute for Molecular Genetics, Berlin. The work was partially supported by the research programmes in EURANDOM, Eindhoven and in Max-Planck-Institute for Molecular Genetics, Berlin. It was partially conducted while the author was visiting the Department of Computational Molecular Biology, Max-Planck-Institute for Molecular Genetics. The authors are grateful to Professors M. de Gunst, W. van Zwet and Miss Johanna Holbrook for their critical reading of the manuscript.

References

Agalarov, S.C., Prasad, G.S., and et al. (2000). Structure of the S15, S6, S18-rRNA complex: Assembly of the 30S ribosome central domain. *Science,* **288,** 107-112.

André, C., Vincens, P., and et al. (2001). MOSAIC: segmenting multiple aligned DNA sequences. *Bioinformatics,* **17,** 196-197.

Auger, I.E. and Lawrence, C.E.(1989). Algorithms for the optimal identification of segment neighborhoods.

Bull. Math. Biol., **51,** 39-54.

Baxevanis, A.D. and Ouellette, B.F.F. (1998). *Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins.* John Wiley, New York.

Berg, O.G. and von Hippel, P.H. (1987). Selection of DNA binding sites by regulatory proteins: statistical-mechanical theory and application to operators and promoters. *J. Mol. Biol.*, **193,** 723-750.

Braun, J.V., Braun, R.K., and Müller, H.G. (2000). Multiple changepoint fitting via quasilikelihood, with application to DNA sequence segmentation. *Biometrika,* **87,** 301-314.

Braun, J.V. and Müller, H.G.(1998). Statistical methods for DNA sequence segmentation. *Statist. Sci.,* **13,** 142-162.

Casari, G., Sander, C. and Valencia, A. (1995). A method to predict functional residues in proteins. *Nature Struct. Biol.* **2,** 171-178.

Clote, P. and Backofen, R. (2000). *Computational Molecular Biology –An Introduction.* John Wiley, New York.

Draper, D.E. and Reynaldo, L.P. (1999). RNA binding strategies of ribosomal proteins. *Nucleic Acids Research,* **27,** 381-388.

Durbin, R., Eddy, S., Krogh, A. and Mitchison, G. (1998). *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids.* Cambridge University Press, Cambridge.

Duret, L. and Abdeddaim, S. (2000). Multiple alignments for structural, functional, or phylogenetic analyses of homologous sequences. In *Bioinformatics: Sequence, structure and databanks,* edited by Higgins, D. and Taylor, W., pp.51-74. Oxford University Press, Oxford.

Gracy, J. and Argos, P. (1998). Automated protein sequence database classification. II. Delineation of domain boundaries from sequence similarities. *Bioinformatics,* **14,** 174-187.

Grantham, R., Gautier, C. and et al. (1981). Codon catalog usage is a genome strategy modulated for gene expressivity. *Nucleic Acids Res.,* **9,** r43-r74.

Heumann, J.M., Lapedes, A.S., and Stormo, G.D. (1994). Neural Networks for determining protein specificity and multiple aligment of binding sites. In *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology,* pp. 188-194, AAAI Press.

Hofmann, K., Bucher, P., Falquet, L. and Bairoch, A. (1999). The PROSITE database, its status in 1999. *NucleicAcids Research,* **27,** 215-219.

Hughes, J.D., Estep, P.W., Tavazoie, and Church, G.M. (2000). Computational identification of cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae. *J. Mol. Biol.,* **296,** 1205-1214.

Landgraf, R., Fischer, D., and Eisenberg, D. (1999). Analysis of heregulin symmetry by weighted evolutionary tracing. *Protein Engineering,* **12,** 943-951.

Lawrence, C.E., Altschul, S.F. and et al. (1993). Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science,* **262,** 208-214.

Li, W. (2001). DNA segmentation as a model selection process. In *Proceedings of the Fifth Annual International Conference on Computational Biology,* pp. 204-210.

Lindsey, J. (1999). *Models for Repeated Measurements. 2nd Edition.* Oxford University Press, Oxford.

Liu, J. and Lawrence, C.E. (1999). Bayesian inference on biopolymer models. *Bioinformatics,* **15,** 38-52.

Liu, J., Neuwald, A.F. and Lawrence, C.E. (1995). Bayesian models for multiple local sequence alignment and Gibbs sampling strategies. *J. Am. Stat. Assoc.,* **90,** 1156-1170.

Nakagawa, A., Nakashima, T., and et al. (1999). The three-dimensional structure of the RNA-binding domain of ribosomal protein L2; a protein at the peptidyl transferase center of the ribosome. *The EMBO Journal,* **18,** 1459-1467.

Oliver, J.L., Román-Roldán, R., Pérez, J. and Bernaola-Galván, P. (1999). SEGMENT: identifying compositional domains in DNA sequences. *Bioinformatics,* **15,** 974-979.

Román-Roldán, R., Benaola-Galván, P. and Oliver, J.L. (1998). Sequence compositional complexity of DNA through an entropic segmentation method. *Physical Review Letters,* **80,** 1344-1347.

Schneider, T.D., Stormo, G.D., and et al. (1986). Information content of binding sites on nucleotide sequences. *J. Mol. Biol.,* **188,** 415-431.

Stormo, G.D. and Fields, D.(1998). Specificity, free energy and information content in protein-DNA interactions. *TIBS,* **23,** 109-113.

Thompson, J.D., Higgins, D.G. and Gibson, T.J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice. *Nucleic Acids Res.,* **22,** 4673-4680.

Wong, W.H. and Zhang, J.(1999). Ribosomal proteins: Homology and motifs. *Manuscript.*

```
ecoli MAVVKCKPTSP-GRRHVVKVVNPELHKG---------------KPFAPLLEKNSKSGGRN-NNGRITTRHIGGGH
yeast ------------------------------------------MGRVIRNQRKGAG-S---IFTSHTRLRQGA
arcfu ------------------------------------------MGKRIISQNRGKGTP---TYRAPSHRYKTD

ecoli KQAYRIVDFKR-NKDGIPAVVERLEYDPNRSANIALVLYKD-----GERRYILAPKGLKAGDQIQS---------
yeast AKLRTLDYAER--HGYIRGIVKQIVHDSGRGAPLAKVVFRDPYKYRLREEIFIANEGVHTGQFIYAG--------
arcfu AKLLRFK------DEVVAAKVIDIQHDSARNGPVALVKLPD-----GSETYILAVEGLGTGDVVYAG--------
      aa|aaaaaa----- block 4 --aa-------------|     |----  block 6 ----|
ecoli GVDAAIKPGNTLPMRNIPVGSTVHNVEMKPGKGGQLARSAGTYVQIVARD--GAYVTLRLRSGEMRKVEADCRAT
yeast -KKASLNVGNVLPLGSVPEGTIVSNVEEKPGDRGALARASGNYVIIIGHNPDENKTRVRLPSGAKKVISSDARGV
arcfu -DNVEIASGNITYLKNIPEGTPVCNIEAQPGDGGKFIRASGTFGFVVSREAD--KVLVQMPSGKQKWFHPNCRAM
       |-----------------    block 8 aaaaa ------------------------aaa------------
ecoli LGEVGNAEHMLRVLGKAGAARWRGVR-----PTVRGTAMNPVDHPHGGGEGRNFG--KHPVT---PWGVQTKGKK
yeast IGVIAGGGRVDKPLLKAGRAFHKYRLKRNSWPKTRGVAMNPVDHPHGGGNHQHIGKASTISR-GAVSGQKAGLIA
arcfu IGVVAGGGRTDKPFVKAGKKYHKMKSKAAKWPRVRGVAMNAVDHPFGGGKHQHVGKPKTVSR-NAPPGRKVGSIA
                                          aaaaaaaaaaaaaa
      --------------|               |--  block 10 ----------|
ecoli TRSNKRTDKFIVRRRSK-------
yeast ARRTGLLRGSQKTQD---------
arcfu ARRTGVRR-----------------
```

Figure 3.1. Multiple alignment for three representive sequences in the L2 family and four important blocks defined in Table 3.1. 'a' is used to mark the important 'a' blocks (sites) defined in Table 3.2. Here ecoli: E.coli; yeast: Saccaromyces cere.; arcfu: Arch. fulgidus. The same designations are used in Figure 3.2.

```
ecoli ------------------------------------MSLSTEATAKIVSEFGRDAND----------TG-------
yeast MGRMHSAGKGISSSAIPYSRNAPAWFKLSSESVIEQIVKYARKGLTPSQIGVLLRDAHGVTQARVITG-------
arcfu MARIHARRRGKSGSKRIYRDSPPEWVDMSPEEVEKKVLELYNEGYEPSMIGMILRDRYGIPSVKQVTG-------
                                                                        **
                                          |--- block 3 ------|            ||
ecoli --------------STEVQVALLTAQINHLQGHFAEHKKDHHSRRGLLRMVSQRRKLLDYLKRKDVA----RYTQ
yeast NKIMRILKSNGLAPEIPEDLYYLIKKAVSVRKHLERNRKDKDAKFRLILIESRIHRLARYYRTVAVLPPNWKYES
arcfu KKIQKILKEHGVEIKYPEDLKALIKKALKLRAHLEVHRKDKHNRRGLQLIEAKIWRLSSYYKEKGVLPADWKYNP
                    |------------------**  *  *  block 7 ----------**------|   |---
ecoli LIERLGLRR-
yeast ATASALVN--
arcfu DRLKIEISK-
      block 9 -|
```

Figure 3.2. Multiple alignment for the three representive sequences in the S15 family and the important blocks defined in Table 3.3. * is used to mark the important blocks (sites) defined in Table 3.4.

```
theth --------------------MSTKNAKPKKEAQRRPSRKAKVKATLGEFDLRDYRN-VEVLKRFLSETGKILPRR
hpylo ------------------------------MERKRYSKRYCKYTEAKISFIDYKD-LDMLKHTLSERYKIMPRR
mgen  MMINKEQDLNQLETNQEQSVEQNQTDEKRKPKPNFKRAKKYCRFCAIGQLRIDFIDDLEAIKRFLSPYAKINPRR
                                  |--aaaaaaaaaaa       **********************
theth RTGLSGKEQRILAKTIKRARILGLLPFTEKLVRK--------
hpylo LTGNSKKWQERVEVAIKRARHMALIPYIVDRKKVVDSPFKQH
mgen  ITGNCNMHQRHVANALKRARYLALVPFIKD------------
      *****  block 2       ***** --|
```

Figure 3.3. Multiple alignment for the three representive sequences in the S18 family and the important blocks defined in Table 3.5. * (or a) is used to mark important blocks (sites) defined in Table 3.6. Here theth: Thermus thermophilus; hpylo: Heli. pylori; mgen: M. genitalium.

Table 3.1 Analysis of IC for L2 by Method 2 with $c_0 = 4.5$

i	l_i	AIC_i	S_i	r_i	i	l_i	AIC_{Ti}	S_i	r_i
1	29	1.094	0.20	7.97	**7**	151	1.384	5.53	14.6
2	45	1.366	6.14	19.9	8^b	240	1.780	0.11	6.55
3	77	1.453	0.14	4.60	9	256	1.692	0.24	5.23
4^b	116	1.832	0.15	6.66	10^b	280	2.294	0.25	5.66
5	121	1.382	9.97	7.93	11	317	1.335	0.09	6.47
6^b	141	1.833	0.19	12.73	**12**	324	1.230	4.35	

Gap blocks are in bold type and underlined.

^{b} The blocks have a AIC_{Ti} value larger than 1.720 (25 % of the largest AIC_{Ti} among those with $S_i < 1$).

Table 3.2 Analysis of IC for L2 by Method 1 with $\alpha = 0.2$

i	l_i	AIC_{Ti}	S_i	r_i	i	l_i	AIC_{Ti}	S_i	r_i
1	29	1.094	0.20	7.98	19^a	185	1.832	1.73	4.26
2	45	1.366	6.15	25.0	20^b	198	1.880	0.42	5.11
3	58	1.378	0.34	5.21	21	204	1.073	0.41	5.22
4	63	1.330	0.86	4.96	22^b	210	1.873	0.74	4.12
5	75	1.578	0.45	3.32	23^a	213	2.230	1.23	3.36
6^a	80	1.654	1.26	5.23	24^b	225	1.775	0.36	3.09
7^a	84	1.785	1.67	4.04	25	240	1.631	0.32	4.21
8	91	1.293	0.39	4.24	26^b	251	1.933	0.47	5.82
9^b	100	1.703	0.43	3.75	**27**	256	1.152	2.81	6.54
10^a	102	3.170	3.23	3.59	28^b	262	2.168	0.58	2.96
11^b	108	2.082	0.68	3.50	29^a	267	2.792	1.01	3.54
12^b	116	1.895	0.49	8.32	30^a	271	3.055	1.70	6.19
13	121	1.384	9.81	11.9	31^a	276	2.044	2.66	5.61
14	125	1.594	0.95	4.78	32	287	1.317	0.28	4.18
15^b	130	2.194	0.89	2.90	33^b	292	1.822	0.99	4.96
16^b	141	1.758	0.30	12.6	34	301	1.535	0.34	4.05
17	151	1.382	5.52	14.5	35	317	1.097	0.39	9.74
18^b	180	1.873	0.24	4.42	**36**	324	1.230	4.35	

Gap blocks are in bold type and underlined.

^{a} The non gap blocks with $S_i \geq 1$.

^{b} The blocks have a AIC_{Ti} value larger than 1.643 (25% of the largest AIC_{Ti} among those with $S_i < 1$).

Table 3.3 Analysis of IC for S15 by Method 2 with $c_0 = 4.5$

i	l_i	AIC_{Ti}	S_i	r_i	i	l_i	AIC_{Ti}	S_i	r_i
1	15	0.895	1.04	5.03	**6**	89	0.886	1.09	23.7
2	35	0.756	0.59	14.7	7^b	142	1.886	0.27	7.11
3^b	56	1.533	0.32	10.2	**8**	146	1.147	2.14	4.73
4	66	0.832	0.77	7.10	9^b	160	1.527	0.21	
5^a	68	2.870	2.93	11.4					

Gap blocks are in bold type and underlined.

^{a} The non gap blocks with $S_i \geq 1$.

^{b} The blocks have a AIC_{Ti} value larger than 1.379 (25 % of the largest AIC_{Ti} among those with $S_i < 1$).

Table 3.4 Analysis of IC for S15 by Method 2 with $c_0 = 3.5$

i	l_i	AIC_{Ti}	S_i	r_i	i	l_i	AIC_{Ti}	S_i	r_i
1	15	0.89	1.04	5.03	19	114	2.13	7.32	12.2
2	35	0.76	0.59	15.6	20^*	115	3.49	∞	7.8
3	55	1.44	0.36	3.6	21	117	1.79	5.39	7.3
4	57	2.06	1.84	4.2	22	118	2.30	∞	10.9
5	66	0.85	0.79	6.8	23	120	1.79	6.18	8.2
6^*	68	2.87	2.93	24.2	24	121	2.50	∞	14.0
7	75	1.01	26.1	3.8	25	123	2.03	8.69	5.0
8	89	0.83	0.59	10.5	26	125	1.27	5.07	4.3
9	94	2.03	1.09	4.2	27	127	2.24	2.14	6.7
10	99	1.90	1.21	4.3	28	131	1.84	3.74	11.9
11	101	1.26	3.27	8.6	29	133	2.00	11.9	16.5
12	102	2.43	∞	8.3	30	134	1.13	∞	9.5
13	104	1.25	3.56	7.5	31^*	136	2.67	8.73	11.0
14	105	1.88	∞	5.1	32	139	1.26	3.46	6.6
15	107	1.57	1.87	4.0	33	140	1.67	∞	4.4
16^*	109	2.88	2.86	4.6	34	142	1.45	1.99	7.1
17	111	1.24	3.29	9.6	**35**	146	1.15	2.14	4.7
18^*	112	2.65	∞	10.8	36	160	1.52	0.21	

Gap blocks are in bold type and underlined.

* The blocks have a AIC_{Ti} value larger than 2.62 (25% of the largest AIC_{Ti} among those non gap blocks.

Table 3.5 Analysis of IC for S18 by Method 2 with $c_0 = 4.5$

i	l_i	AIC_{Ti}	S_i	r_i
1	31	0.751	1.08	24.50
2	105	1.828	0.16	15.71
3	117	0.910	3.27	

Table 3.6 Analysis of IC for S18 by Method 2 with $c_0 = 3.3$

i	l_i	AIC_{Ti}	S_i	r_i	i	l_i	AIC_{Ti}	S_i	r_i
1	26	0.790	1.80	13.4	7^*	80	2.069	0.26	3.42
2	34	0.680	0.51	6.90	8	96	1.180	0.43	3.89
3^a	41	1.153	1.98	4.27	9^{a*}	101	2.408	1.29	3.81
4^a	45	1.962	1.57	4.05	10	105	1.607	0.91	15.4
5	52	1.223	0.55	4.22	**11**	117	0.910	3.27	
6^*	57	2.268	0.87	4.12					

Gap blocks are in bold type and underlined.

^{a} The non gap blocks with $S_i \geq 1$.

* The blocks have a AIC_{Ti} value larger than 1.701 (25% of the largest AIC_{Ti} among those non gap blocks.