J. theor. Biol. (2001) **212,** 129–139
doi:10.1006/jtbi.2001.2319, available online at http://www.idealibrary.com on IDEAL®

Automatic Discovery of Sub-molecular Sequence Domains in Multi-aligned Sequences: A Dynamic Programming Algorithm for Multiple Alignment Segmentation

ERIC POE XING*†, DENISE M. WOLF*, INNA DUBCHAK*, SYLVIA SPENGLER*,
MANFRED ZORN*, ILYA MUCHNIK‡ AND CASIMIR KULIKOWSKI‡

**Center for Bioinformatics and Computational Genomics, NERSC, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, U.S.A. and ‡Department of Computer Science and DIMACS, Rutgers University, Piscataway, NJ 08855, U.S.A.*

(*Received* 25 *May* 2000, *Accepted in revised form on* 20 *March* 2001)

Automatic identification of sub-structures in multi-aligned sequences is of great importance for effective and objective structural/functional domain annotation, phylogenetic treeing and other molecular analyses. We present a segmentation algorithm that optimally partitions a given multi-alignment into a set of potentially biologically significant blocks, or segments. This algorithm applies dynamic programming and progressive optimization to the statistical profile of a multi-alignment in order to optimally demarcate relatively homogenous sub-regions. Using this algorithm, a large multi-alignment of eukaryotic 16S rRNA was analyzed. Three types of sequence patterns were identified automatically and efficiently: shared conserved domain; shared variable motif; and rare signature sequence. Results were consistent with the patterns identified through independent phylogenetic and structural approaches. This algorithm facilitates the automation of sequence-based molecular structural and evolutionary analyses through statistical modeling and high performance computation.

1. Introduction

The coding sequences of macromolecules with complex biological functions usually contain alternating invariant and variable regions (Ludwig & Schleifer, 1994). The identification and characterization of these sub-molecular regions is important for many types of sequence-based molecular analyses, such as comparative structural prediction, supervised multiple sequence alignment and phylogenetic tree construction (Felsenstein, 1982; Koonin *et al*., 1998; States & Boguski, 1991).

Pattern extraction in biologically related sequences is traditionally done by manual inspection and curation of a multiple alignment of these sequences, with some empirical expert knowledge or comparison heuristics. Usually, this process is not only time-consuming, but also often lacks strict, consistent, and formal criteria for knowledge discovery. Some computer tools, such as Prettybox (Westerman, 1998) and Genome Channel (Mural *et al*., 1999), have been developed to assist in such a process. However, most of the tools in fact only serve annotation or

† Author to whom correspondence should be addressed. Current address: 593 Soda Hall, Division of Computer Science, UC Berkeley, Berkeley, CA 94706, U.S.A.

0022–5193/01/180129 + 11 $35.00/0

visualization roles rather than doing active and globally optimized pattern recognition based on solid statistical reasoning (Stojanovic *et al.*, 1999). Multiple alignment remains a major technique to unveil hidden structural details in the orthologous gene sequences of different species. In recent years, multiple alignments in publicly available databases [e.g. RDP, release 7.0 (Maidak *et al.*, 1999)] have grown dramatically in size and complexity, which makes empirical pattern extraction from the entire alignment difficult and not even appropriate given the diversity of sequences. More sensitive, consistent and efficient methods, based on formal information retrieval rules and feature definitions, are needed to meet this challenge.

To develop a formal description of sub-molecular regions potentially having a unique and stable property in a gene sequence, we hypothesized the following: a sub-molecular entity with distinguishable structural, functional or evolutionary properties may possess unique statistical features in a multi-alignment. Since a gene usually contains multiple well-preserved domains and is interspersed with less stable or even random sequences, domain-specific statistical features are expected to exhibit discontinuities at the boundaries between different regions and be relatively more uniform within a region. Here, we present a segmentation algorithm, based on dynamic programming and progressive optimization, that identifies such discontinuities and automatically partitions a multi-alignment into a set of segments strictly characterized by the statistical profile of its sequence composition. Based on two simple profile measurements: the degree of homogeneity of character composition at each site, and the gap frequency therein, our algorithm successfully found from a eukaryotic 16S rRNA multi-alignment, a segmentation pattern consistent with the positions of evolutionarily conserved and heterogeneous regions independently determined through other approaches (Gutell, 1993). Quantitative analysis of the resulting segments based on the distribution of hamming distances of each sequence to the consensus, and associated entropies (randomness), supports the assumption underlying our segmentation algorithm of a non-random, near-quantum distribution of statistical features in the multi-alignment. Although still in the prototype stage, we believe our algorithm to be a promising step toward the automation of sequence-based sub-molecular structural and evolutionary analyses.

2. Methods and Algorithms

A multi-alignment can be viewed as a character table that resembles the pixel matrix of a graphical image except that the numerical pixels are replaced by characters from a pre-defined vocabulary set $X = \{\mathrm{A, G, C, T}$ (or U), – (gap)$\}$ (we can easily generalize this setting to protein sequences by replacing the vocabulary set with an amino acid species set). Each column in this table represents a virtual (in case it corresponds to a gap) or an actual nucleotide site within the sequences being aligned. Each row represents a sequence hosted by a particular species. Analogous to the concepts used in image processing (Kittler & Foglein, 1984), we define a segment S_i of the multi-alignment to be an ordered set of consecutive columns within the multi-alignment table. The image-processing-based segmentation technique presented below, described in part in Xing *et al.* (1999), combines column-wise statistical profile information like that used in Gribskov *et al.* (1987) with a dynamic programming approach often employed in alignment and model-fitting algorithms (Auger & Lawrence, 1989; Gorodkin *et al.*, 1997).

2.1. GENERAL DYNAMIC PROGRAMMING PROCEDURE FOR OPTIMAL SEGMENTATION

For a given multi-alignment A and a pre-defined parameter k which specifies the total number of segments to be produced after the segmentation, associate any k-segmentation $S = \langle S_1, \dots ,S_k \rangle$ on A with a segmentation score function:

$$I(S) = \sum_{\alpha=1}^{k} F_\alpha, \tag{1}$$

where F_α is a segment-specific score function of segment α (i.e. proportion of gaps, or other measures of heterogeneity associated with the segment). An optimal segmentation S^* can be obtained by minimizing $I(S)$:

$$S^* = \underset{|S|=k}{\operatorname{argmin}}\ I(S). \qquad (2)$$

Since F_α is dependent on the choice of the segment α and its delimitation, we can rewrite it as $F(S_\alpha)$ or $F(l_a, r_a)$, where S_α is the segment delimited by l_a as its left boundary and r_a as the right one, $\alpha \in \langle 1, \ldots, k\rangle$. For any definition of F_α (on a subinterval indexed by integer 1–N), the minimum of $I(S)$ can be found through a dynamic programming procedure (Bellman, 1957; Mottl & Muchnik, 1998) which progressively (from right to left) establishes the optimum right boundary profiles $j_l^*(i)$ of the segment l for each possible left boundary i, together with their associated partial segmentation score:

$$\Phi_l^i = \min\left(\sum_{\alpha=l}^{k} F_\alpha\right). \qquad (3)$$

This procedure will terminate when the leftmost possible boundary $i = 1$ is reached. Following is the outline of this procedure:

For $l = k - 1$ to 1,

Define $L_l = \langle l, l+1, \ldots, N-(k-l)-1\rangle$ as a set of left boundaries of segment l.

For $\forall i \in L_l, \Rightarrow$

Define $R_l^i = \langle i+1, i+2, N-(k-l)\rangle$ as a set of right boundaries of segment l whose left boundary is i.

For $\forall j \in R_l^i, \Rightarrow$

$$Q_l^i(j) = F_l(i,j) + \Phi_{l+1}^j, \qquad (4)$$

$$\Phi_l^i = \min\left(\sum_{\alpha=l}^{k} F_\alpha\right) = \min_{j\in R_l^i}(Q_l^i(j)), \qquad (5)$$

$$j_l^*(i) = \arg\left(\min_{j\in R_l^i}(Q_l^i)(j)\right). \qquad (6)$$

The prodedure terminates when $I(S^*) = \Phi_1^I$ is obtained. The time complexity of the procedure is $O(kn^2G)$, where G is the cost for the calculation of Q in eqn (4). To further reduce the time cost, one can spend n^2 units of memory to store all pre-calculated $F(i,j)$ values rather than calculating them for each cycle. Once the optimal right boundary profile $j_l^*(i)$ of segment l for each possible left boundary i is produced, it is easy to delimit the multi-aligned sequences such that they form an optimal segmentation. Starting from the leftmost segment, after assigning its left boundary as 1, one can systematically look up in the profile to retrieve the boundaries of all the segments from left to right according to the following functions:

$$l_1 = 1, \qquad r_1 = j_1^*(l_1),$$

$$l_2 = r_1 + 1, \qquad r_2 = j_1^*(l_2),$$

$$\ldots$$

$$l_\alpha = r_{\alpha-1} + 1, \quad r_\alpha = j_\alpha^*(l_\alpha), \quad \alpha \in \langle 1, 2, \cdots, k\rangle$$

The resulting final segmentation is

$$S^* = \langle S_1(1, j_1^*(1)), S_2(j_1^*(1)+1, j_2^*(j_1^*(1)+1)),$$
$$\cdots, S_k(j_{k-1}^*+1, j_k^*(j_{k-1}^*+1))\rangle.$$

2.2. OBJECTIVE FUNCTIONS FOR DYNAMIC OPTIMIZATION

Depending on the desired features to be captured from segmentation, various types of segment-specific score functions F can be chosen [based on the concept of profile analysis (Gribskov *et al.*, 1987)]. We used a set of objective functions that measure the square error of several column-wise alignment features:

$$F_G(l_\alpha, r_\alpha) = \sum_{j=l_\alpha}^{r_\alpha} (n_j^{gap} - \bar{n}_\alpha^{gap})^2, \qquad (7)$$

where $n_j^{gap} =$ frequency of "$-$" at j-th column of the multi-alignment,

$$\bar{n}_\alpha^{gap} = \frac{1}{r_\alpha - l_\alpha + 1}\sum_{j=l_\alpha}^{r_\alpha} n_j^{gap}.$$

$$F_E(l_\alpha, r_\alpha) = \sum_{j=l_\alpha}^{r_\alpha} (e_j - \bar{e}_\alpha)^2, \qquad (8)$$

where $e_j = \sum_{l \in \{-,\mathrm{A},\mathrm{G},\mathrm{C},\mathrm{U}\}} n_{j,l} \log(n_{j,l})$, $n_{j,l} =$ frequency of l at j-th column, $\bar{e}_\alpha = (1/(r_\alpha - l_\alpha + 1)) \sum_{j=l_\alpha}^{r_\alpha} e_j$.

$$F_H(l_\alpha, r_\alpha) = \sum_{j=l_\alpha}^{r_\alpha} (h_j - \bar{h}_\alpha)^2, \qquad (9)$$

where $h_j = \sum_{l \in \{-,\mathrm{A},\mathrm{G},\mathrm{C},\mathrm{U}\}} (n_{j,l})^2$, $n_{j,l} =$ frequency of l at j-th column, $\bar{h}_\alpha = (1/(r_\alpha - l_\alpha + 1)) \sum_{j=l_\alpha}^{r_\alpha} h_j$. Score functions F_G, F_E and F_H measure the level of non-uniformity of (1) the column-wise gap frequency, (2) the column-wise entropy of the character distribution, and (3) the degree of character heterogeneity in each column (as explained in the appendix), respectively, across segment α. Using one of the score functions F as an objective function, the dynamic programming procedure described in Section 2.1 leads to a segmentation of the multi-alignment such that the property of interest (column-wise gap frequency, entropy, etc.) is as uniform as possible within each segment.

3. Hardware, Software and Dataset

The segmentation program was written in C and implemented on Sun Ultra30 workstation. Statistical analyses and plots were done using Splus on PC. The multi-alignment used in this paper was obtained from Ribosomal Database Project (RDP, release 7.0) (Maidak *et al.*, 1999) by choosing a subset of 417 sequences out of the complete multi-alignment of 2055 eukaryotic small subunit 16S ribosomal RNA sequences (in order to facilitate comparison with a smaller earlier release). The "sub-alignment" is 6197 base-pair long. The rRNA multi-alignment provided by RDP is achieved by a joint effort of computer optimization and manual validation/modification.

4. Experiments, Results and Discussions

4.1. SEGMENTATION

As shown in Fig. 1, a multi-alignment of 6197 bp × 417 species, typical of a modern sequence database, is extremely complex and irregular. Even with a plot of the complete profile of a measure of interest, say, the gap frequency at each column, it is still hard to accurately identify structural details therefrom, let alone by directly inspecting an alignment table of this size. We performed a segmentation on this alignment using the objective function F_G (setting $k = 100$), and superimposed the result on the gap profile plot in Fig. 1. Segmentation using F_G minimizes the sum of square errors of column-wise gap frequencies in each segment; in each resulting segment, the frequencies of gap occurrence in the columns therein are relatively uniform. Thus, the gap-rich and gap-rare regions in the multi-alignment are separated in an optimal way for a given pre-specified total number of segments.

However, F_G only captures the distribution of gaps in the multi-alignment. It is often more desirable to also consider the degree of homogeneity of the aligned sequences. An immediate alternative is to replace F_G with F_E, which traces the entropy change of nucleotide occurrence at columns along the multi-alignment. A segment with low entropy across all columns corresponds to a homogeneous fragment, and vice versa. Another choice is to use F_H, which, as briefly explained in the appendix, also reflects the degree

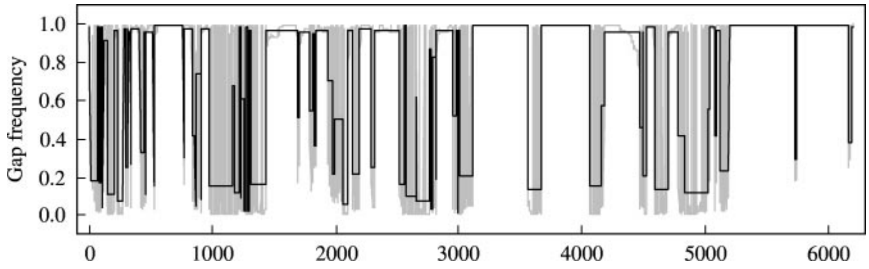

FIG. 1. Segmentation of the rRNA multi-alignment using F_G as objective function. The gray plot at the background is the actual gap frequency profile of the multi-alignment. The black plot represents the $\bar{g}$ (average gap frequency) of each of the resulting 100 segments.

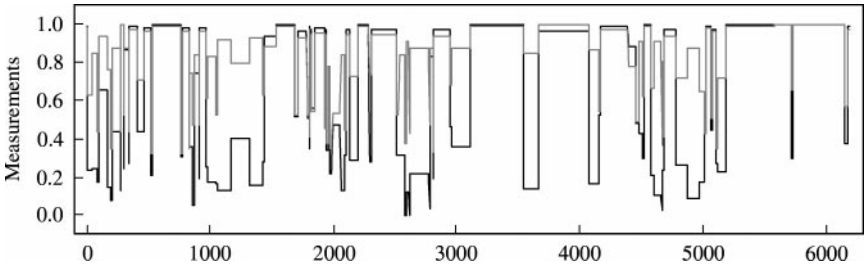

FIG. 2. Segmentation result using F_H as objective function ($k = 100$). The gray line is the $\bar{h}$ (average degree of homogeneity) of consecutive segments along the multi-alignment, the black line is the $\bar{g}$ (average gap frequency) in these segments.

of sequence homogeneity (in here high homogeneity corresponds to high h value), but has a convenient 0–1 value range, and offers a more easily seen connection to the underlining gap or nucleotide frequency [see eqn (A.3)]. We performed segmentation using both F_E and F_H and got consistent results. For brevity, in this paper, we present and discuss only the F_H and F_G results (Fig. 2).

Since segmentation using F_H (H-segmentation) reflects the fluctuation of the level of homogeneity of nucleotides at each site of the aligned sequences, it naturally reveals regional conservation or variation of a particular gene in different organisms. As—for the sake of simplicity—we did not explicitly encode the biological difference between gap and other nucleotide characters in F_H, the resulting segments with high $\bar{h}$ (which implies the existence of a dominant character type across all rows at each column) correspond to either a segment with a predominant, rarely interrupted, sequence pattern, or to one unanimously dominated by gaps in all columns. We combined the results from cost functions F_H and F_G to distinguish these two cases. Thus, the average gap frequency ($\bar{g}$) of each segment resulting from the H-segmentation was calculated (Fig. 2) as an auxiliary measurement in addition to $\bar{h}$.

4.2. CLASSIFICATION OF SEGMENTS

For a character set of size 5 ({A, U, G, C, –}), if all types of characters occur at random in each column, $\bar{g}$ within a segment would be ~0.2, as would $\bar{h}$. We define three types of segments as being of particular interest: (1) highly homogeneous and gap-rare segment ($\bar{h} \geqslant 0.8, \bar{g} \leqslant 0.2$); (2) gap stretches ($\bar{h} \geqslant 0.8, \bar{g} \geqslant 0.8$); and (3) heterogeneous but still gap-rare segment ($\bar{h} \leqslant 0.4, \bar{g} \leqslant 0.2$). Notice that "heterogeneous gap stretch" is not a segment pattern existing in practice. In the 100 segments generated by H-segmentation on the rRNA multiple alignment, 11 belong to type 1, 31 are of type 2 and 8 of type 3 (Table 1).

Most type 1 segments have a length of 50–100 bp. High $\bar{h}$ suggests that different organisms share a similar sequence in the segment [i.e. $\bar{h} = 0.81$ corresponds to a distribution of at least 90% of the sequences in the same pattern, see eqn (A.3)]. Low $\bar{g}$ means that the pattern is not a gap stretch but a continuous nucleotide sequence. Together these are strong indications of a conserved domain shared among multiple organisms. Type 2 segments cover about 60% of the total length of the multi-alignment and range from 5 to over 500 bp long. An overabundance of gaps in some regions of a multi-alignment is usually due to the introduction of stretches of gaps into the sequences of the majority species devoid of some rare patterns possessed by a few co-aligned species in the corresponding region. Therefore, such segments may harbor an uncommon sequence pattern (i.e. signature pattern of some species) or sequences that are "shared" in a highly interrupted fashion among species represented in the alignment. Type 3 segments are generally very short, and their biological meaning is unclear. They may merely be the result of suboptimal alignment, but may also represent a novel class of sequence motifs whose exact contexts vary from species to species and reside at specific locations in the gene of all species. It is possible that these short and heterogeneous motifs may encode some special structural or

TABLE 1

*Summary of three types of segments resulting from H-segmentation**

Segment number (α)	Boundary	Average homogeneity ($\bar{h}_\alpha$)	Average gap frequency ($\bar{g}_\alpha$)	Adjusted consensus length† (L)	Peak Hamming distance	H/L ratio‡	Normalized entropy	Reference i.d.
Type 1								
2	39..83	0.843	0.246	36	0	0.000	0.583	1.1
27	988..1053	0.829	0.17	55	3	0.055	0.666	1.2
29	1060..1167	0.918	0.125	95	2	0.021	0.561	1.3
31	1326..1430	0.925	0.154	90	1	0.011	0.500	1.4
49	2063..2101	0.836	0.131	34	1	0.029	0.604	1.5
59	2596..2620	0.909	0.12	22	0	0.000	0.515	1.6
61	2627..2782	0.868	0.216	122	5	0.041	0.641	1.7
69	3557..3668	0.85	0.136	99	2	0.020	0.690	1.8
71	4081..4158	0.863	0.163	66	3	0.045	0.596	1.9
80	4609..4672	0.911	0.105	58	0	0.000	0.538	1.10
86	4887..4983	0.87	0.087	90	2	0.022	0.597	1.11
Type 2								
17	531..767	0.983	0.991	237				2.1
33–	1444..1542	0.879	0.936	245				2.2§
–34	1543..1688	0.988	0.994					
55	2317..2523	0.944	0.971	207				2.3
68	3115..3556	0.987	0.994	442				2.4
70	3669..4080	0.994	0.959	412				2.5
73	4180..4399	0.964	0.981	220				2.6
94–	5193..5593	0.987	0.993	538				2.7§
–95	5594..5730	0.992	0.996					
97	5739..6157	0.995	0.998	419				2.8
Other 21 segments				5–128				
Type 3								
6	199..207	0.35	0.08	9	7	0.778	0.664	3.1
16	522..530	0.325	0.206	8	7	0.875	0.324	3.2
47	1976..1993	0.345	0.215	17	14	0.824	0.531	3.3
58	2591..2595	0.376	0.002	5	2 & 3	0.500	0.883	3.4
60	2621..2626	0.426	0.001	6	4	0.667	0.902	3.5
64	2816..2823	0.33	0.19	7	5	0.714	0.734	3.7
81	4673..4678	0.361	0.022	6	5	0.833	0.802	3.8
92	5119..5129	0.35	0.27	8	7	0.875	0.521	3.9

*Shaded row marks the marginal segments, those that are close to the respective thresholds of $\bar{h}_\alpha$ and $\bar{g}_\alpha$.

†For type 1 and 3 segments, the consensus excludes the gaps and thus has a shorter length compared to the segment originally from the multi-alignment. This is to avoid including gap counts in the calculation of hamming distance from each sequence to the consensus. L of type 2 segments is the original length.

‡The ratio between the peak hamming distance and L.

§The two adjacent gap segments (with slightly different statistics) are fused together.

functional entities present in different organisms, but have a lesser degree of conservation at the sequence level (probably due to alternative implementations of a common function in different organisms).

Altogether, 50 of the 100 segments fall into these three types, and they cover 75.9% of the total length of the multi-alignment. These are the regions that are unambiguously aligned in the multi-alignment, and do not tend to contain a mixture of gaps and nucleotides across different species. The remaining 50 segments have intermediate $\bar{h}$ and $\bar{g}$ values, and only cover a small portion of the multi-alignment. These are the regions where gaps are mixed with broken sequences and somehow have a uniform degree of randomness across columns. They are likely to be the heterogeneous regions, only present in some species, and in different forms, which makes them difficult to match across species in both context and position.

4.3. EFFECT OF GRANULARITY OF SEGMENTATION

The granularity of the segmentation can be changed by choosing different values of k in eqn (1–3). We segmented with $k = 25$, 50, 75 and 100. The CPU time increased linearly with k as expected, at a modest rate ($t \cong 3.35k + 173\,\mathrm{s}$). In general, changes in granularity did not perturb the overall pattern of segmentation on a significant scale. Type 2 segments are especially stable. Some rearrangements, such as split or boundary adjustment, did occur in a few segments as the granularity increased. These segments tended to have high $\bar{h}$ but intermediate $\bar{g}$ values in the coarse-grain segmentation. We found that long homogenous sequence stretches interleaved with some short heterogeneous fragments can be further dissected under finer granularity. The successive unfolding of finer structures of multi-alignment with increasing segmentation granularity suggests that finer-grain segmentation produces a higher resolution of the details of the sequences and is preferred if the linear increase of time-cost and memory demand (to store internal states in the loop) is tolerable. Nevertheless, once identified, a good portion of the types 1 and 2 segments were well preserved with changing granularity, and nearly no type 3 segments changed their boundaries during further fine-grained segmentation. Therefore, with a reasonable choice of k, our segmentation can identify segments with potentially biologically meaningful properties with a high degree of robustness and consistency.

4.4. SEGMENTATION OF A DIFFERENT VERSION OF MULTI-ALIGNMENT OF THE SAME SET OF SEQUENCE

Our segmentation software has undergone several upgrades after its initial development, and so has the multi-alignment we analysed. In addition to any changes of alignment technique implemented and applied to a given multi-alignment, the continuous addition of new sequences into the database also results in frequent updates of the multi-alignment of the same set of sequences over time. The trend is to put all available sequences of a gene into a single huge alignment (although the validity of such a practice is arguable).

When we first applied our software to analyse rRNA sequences, the entire collection in RDP of eukaryotic 16S rRNA contained 437 sequences in a multi-alignment 4036 bp long (release 6.0). The release 7.0 used in this paper contained 417 of the 437 sequences (others are missing for unknown reasons) plus a few thousands more (which we did not include), in a new multi-alignment of 6197 bp for this subset (chunked from the originally ~ 8000 bp-long multiple alignment of the entire sequence set, and with columns consisting entirely of gaps removed). We compared the segmentation patterns of these two different versions of multi-alignment in Fig. 3 ($k = 70$ for release 6 and $k = 100$ for release 7 to ensure comparable granularity). For direct comparison, segments were mapped onto the original rRNA sequence of the *Cryptococcus neoformans* (1805 bp). The position and length of type 1 segments were consistent in both multi-alignments, except for two of the marginal type 1 segments (2nd and 7th) in version 6, which were either unrecognized or split into smaller strict type 1 segments in the later version. A few new type 1 segments showed up in the later version as well. This suggests that the conserved sequence domains are stably captured through alignment upgrades. Although most of the type 2 segments in release 6 remain in release 7, the later version has significantly more/longer type 2 segments, meaning that the new multi-alignment contains more gap stretches. This is consistent with our previous speculation that type 2 segments are created to accommodate rare sequence patterns, more likely in the new release containing a much greater total number of sequences. We observed fewer type 3 segments in version 7, which seems to imply that some of them were alignment artifacts in the older version, eliminated in the later (presumably improved) version of the multi-alignment. But this does not exclude the possibility that some of them may still be special unconserved motifs, as will be discussed later.

4.5. A CLOSE LOOK AT DEGREE OF SEQUENCE HOMOLOGY WITHIN A SEGMENT

To verify that types 1 and 3 segments represent homogeneous and heterogeneous sequence segments, respectively, we studied the distributions of the sequence patterns within segments. To avoid the undesirable over-determination (and thus lack of statistical abstraction) often encountered in brute force classification of a limited number of samples using a high-dimensional descriptor (i.e. sequence context), we performed a simple classification of sequences within each segment according to their hamming distances to the consensus sequence. The physical meaning of this distance regarding the difference between two sequences is as following: for any pair of sequences having distance $d1$ and $d2$, respectively, to the consensus, the number of nucleotide sites (D) they could differ satisfies the following inequality:

$$|d1 - d2| \leqslant D \leqslant |d1 + d2|. \qquad (10)$$

Therefore, all sequences with hamming distance d to the consensus can differ at most by $\min(2d, L)$ nucleotides, where L is the length of the sequences. To quantitatively measure the impurity of sequence patterns in terms of this distance, the entropy associated with the partition of sequences incurred by the distance is calculated, and normalized with the maximal possible entropy of the segment, $\log_2 L$, for easy comparison of different length segments.

The distributions of d of the 417 sequences aligned in a type 1 and a type 3 segment are shown in Fig. 4. For the type 1 segment, the distance distribution peaks at a small d (compared to the length of the segment), and as a result of peaked distribution, has a relatively small normalized entropy (Table 1). This suggests that a majority of the multi-aligned sequences differ very little within the segment, consistent with the prediction based on the $\bar{h}$ value. On the other hand, for a type 3 segment, either the distribution is scattered (resulting in larger normalized entropy) or/and the peak shifts toward L, the maximal possible hamming distance for a sequence within the segment. This distribution revealed that most of the sequences are grossly different from one another in a type 3 segment, agreeing with our inference that they cover either unconserved or poorly aligned regions.

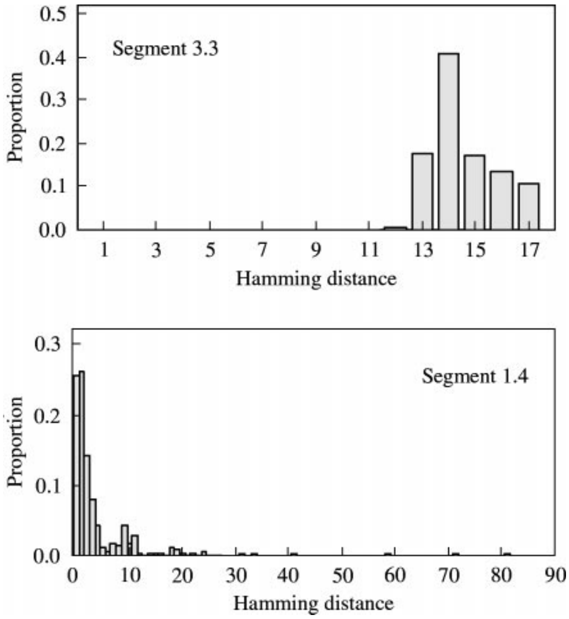

FIG. 4. The distribution of the hamming distances of each sequence to the consensus in two types of segments. Upper: distance distribution for a type 3 segment; lower: distance distribution for a type 1 segment.

4.6. MAPPING OF SEGMENTS ON SECONDARY STRUCTURE

To further explore the biological implication of the three types of segments, we mapped them onto the secondary structure of the *Cryptococcus neoformans* small subunit rRNA (Gutell, 1993) [Fig. 5(a)]. Ten of the 11 type 1 segments corresponded well to phylogenetically and structurally conserved regions independently identified using comparative analysis for higher-order structures conserved among species (Gutell, 1994; Gutell *et al.*, 1994), many of which are core domains forming the backbone of the molecule or involved in important secondary and tertiary structure interactions [Fig. 5(b)]. Many of the most conserved nucleotide sites labeled by Gutell *et al.* were covered in the type 1 segments. However, some regions, such as the 5′ and 3′ ends of the molecule, although also labeled with many conserved sites, did not match type 1 segments. A close inspection of the original multi-alignment showed that the 5′ end region contains frequently alternating (short) runs of gaps and sequences, which suggest that as the sequence collection grows bigger, more variations were revealed. Interestingly, one of the type 1 segments (i.d. 1.5), corresponds to a region labeled highly variable [Fig. 5(b)]. It is possible that this is a "new" conserved domain that will be increasingly apparent as more sequence entries are considered for comparison. In contrast to type 1 segments, type 3 segments all correspond to short sequence patches residing at the periphery or within the variable regions. However, some of these segments are involved in the formation of the most stable thermodynamic foldings [Fig. 5(a), indicated by thick tick marks] (Konings & Gutell, 1995), suggesting that they may be indeed functionally essential to the RNA molecule although contextually heterogeneous. Type 2 segments (gap stretches) mostly fall into the most variable regions, except at both ends of the molecule. The two runs of gaps near the 3′ end of the molecule are about 540 and 400 bp long (notice that entire length of the molecule is 1805 bp), suggesting that some species may contain unique signature motifs near this location that could not be aligned against each other (and thus are juxtaposed together to cause the long gap runs).

In summary, although no manual mapping/cross-validation, secondary structure comparison and expert knowledge of phylogenetic property was involved, the information obtained through a pure statistical segmentation approach about the domain location and degree of conservation, was remarkably consistent with that obtained by human analysis.

5. Biological Applications

The statistical-profile-based segmentation technique presented in this paper can serve as a robust, general-purpose automatic knowledge discovery tool to analyse the structure of large, unwieldy multi-alignments containing a large number of sequences. Such alignments are difficult, if possible, to inspect manually.

Unlike a simple alignment display tool such as PrettyBox (Westerman, 1998), which marks out the "conserved box" simply by highlighting the nucleotides in the aligned sequences that agree with the consensus, this method *infers* all the conserved segments along with other segments using statistical properties of character composition and distribution based on global optimization. This process involves little artificial modeling and arbitrary parameterization and is extremely efficient. As updates of multi-alignments of various genes are becoming more frequently available and ever bigger, our method provides an important alternative to the manual approach as a fast and reliable domain identifier.

One of the most important applications of the segmentation algorithm presented herein is to identify different types of sequence motifs (i.e. orthologous functional domains, signature motifs and non-orthologous functional motifs) from aligned gene sequences. Such an application is useful for functional annotation and the design of organism-specific gene amplifiers. Furthermore, the results of segmentation of multi-aligned sequences can be fed back to the aligner for auto-readjustment of the alignment. At present, multi-alignment is best done using a hybrid approach involving both machine calculation and manual local readjustment (Schuler *et al.*, 1991; States & Boguski, 1991). Algorithms can be designed to mimic such a process by iteratively incorporating segmentation knowledge to readjust and optimize local alignment (i.e. locally realign all sequences in the gap-rare segments to improve homogeneity, or selectively adjust poorly aligned sequences in such segments using adjacent gap-rich segments as relaxation buffer). Another potential application of segmentation is in phylogenetic treeing. Sequence regions with different degrees of variability reflect evolutionary history at different scales and stages (Ludwig & Schleifer, 1994). It would be informative to distinguish different regions through segmentation, and use them during different stages of tree construction, or constitute a proper weighting scheme for the distance measurement (indeed, one of the main pitfalls of current treeing techniques is that the selection of qualified alignment sections and the removal of ambiguous or noisy segments are routinely done manually via eye inspection). It might be useful to construct multiple trees based on sequences in individual segments, and then to aggregate these trees, derived from different parts of the molecule.

Before proceeding to the conclusion, we address some of the pitfalls in our algorithm. (1) Some segments may slightly suffer a "boundary effect" (i.e. a long type 1 segment may contain a short patch of heterogeneous sequence at the boundary), which arises due to the buffering effect of the long segment with a uniform target statistical measure that can absorb the perturbation of small variations at the boundary. (2) Some special alignment patterns, such as a juxtaposition of very short and alternating gap-rich (gap-rare) segments with comparable gap (nucleotide) frequencies, may confuse the segmentor. This pattern could be falsely determined as a single long segment because in terms of character homogeneity, it is "uniform" (gap is taken as one of the characters). As a result, some small motifs may be missed. However, these problems did not seem to seriously affect the performance of the segmentor and can be cured by cross-validation between results from different objective functions and by using greater granularity to improve resolution.

6. Conclusion

We described a segmentation algorithm that can efficiently partition a multiple alignment into a set of biologically sensible segments based on its statistical profile using dynamic programming and progressive optimization. Using this algorithm, a multiple alignment can be segmented into sub-regions each with a uniform level of statistical measurement (i.e. gap frequency or character homogeneity). In the performance test on a large eukaryotic 16S rRNA multiple alignment, our algorithm enabled automatic discovery of the following structures from the aligned sequences with good accuracy: (1) Highly conserved motifs with a shared context among a large number of species. (2) Unique signature motif present only in the sequences of a small number of species. (3) Motifs adapted by a large number of species in the same region of the molecule but displaying variable sequence context among species. This algorithm potentially leads to an efficient and fully automated way of extracting structural details from large datasets, thus facilitating faster and better signature discovery, domain annotation, multiple alignment optimization and high-resolution phylogenetic treeing.

REFERENCES

Auger, I. E. & Lawrence, E. L. (1989). Algorithms for the optimal identification of segment neighborhoods. *Bull. Math. Bio.* **51,** 39–54.

Bellman, R. (1957). *Dynamic Programming*. Princeton: Princeton University Press.

Felsenstein, J. (1982). Numerical methods for inferring evolutionary trees. *Q. Rev. Biol.* **57,** 379–404.

Gorodkin, J., Heyer, L. J. & Stormo, G. D. (1997). Finding the most significant common sequence and structure motifs in a set of RNA sequences. *Nucleic Acids Res.* **25,** 3724–3732.

Gribskov, M., McLachlan, A. D. & Eisenberg, D. (1987). Profile analysis: detection of distantly related proteins. *Proc. Natl Acad. Sci. U.S.A.* **84,** 4355–4358.

Gutell, R. R. (1993). Comparative studies of RNA: inferring higher-order structure for patterns of sequence variation. *Curr. Opin. Struct. Biol.* **3,** 313–322.

Gutell, R. R. (1994). Collection of small subunit (16S- and 16S-like) ribosomal RNA structures. *Nucleic Acid Res.* **22,** 3502–3507.

Gutell, R. R., Larsen, N. & R., W. C. (1994). Lessons from an Evolving rRNA: 16S and 23S rRNA Structures from a comparative perspective. *Microbiol. Rev.* **58,** 10–26.

Kittler, J. & Foglein, J. (1984). Contextual classification of multispectral pixel data. *Image & Vision Comput.* **2,** 13–29.

Konings, D. A. & Gutell, R. R. (1995). A comparison of thermodynamic foldings with comparatively derived structures of 16S and 16S-like rRNAs. *Rna* **1,** 559–574.

Koonin, E. V., Tatusov, R. L. & Galperin, M. Y. (1998). Beyond complete genomes: from sequence to structure and function. *Curr. Opin. Struct. Biol.* **8,** 355–363.

Ludwig, W. & Schleifer, K. H. (1994). Bacterial phylogeny based on 16S and 23S rRNA sequence analysis. *FEMS Microbiol. Rev.* **15,** 155–173.

Maidak, B. L., Cole, J. R., Parker Jr, C. T., Garrity, G. M., Larsen, N., Li, B., Lilburn, T. G., McCaughey, M. J., Olsen, G. J., Overbeek, R., Pramanik, S., Schmidt, T. M., Tiedje, J. M. & Woese, C. R. (1999). A new version of the RDP (Ribosomal Database Project). *Nucleic Acids Res.* **27,** 171–173.

Mottl, V. V. & Muchnik, I. (1998). Bellman functions on trees for segmentation, generalized smoothing. matching multi-alignment in massive data sets. *DIMACS*, *Technical Report*, Vol. 17, pp. 1–54.

Mural, R. J., Parang, M., Shah, M., Snoddy, J. & Uberbacher, E. C. (1999). The Genome Channel: a browser to a uniform first-pass annotation of genomic DNA. *Trends Genet.* **15,** 38–39.

Schuler, G. D., Altschul, S. F. & Lipman, D. J. (1991). A workbench for multiple alignment construction and analysis. *Proteins* **9,** 180–190.

States, D. J. & Boguski, M. S. (1991). Similarity and homology. In: Sequence Analysis Primer (Gribskov, M. & Devereux, J., eds), pp. 90–158. New York: Stockton Press.

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

WESTERMAN, R. (1998). PrettyBox. Madison, Wisc.: Genetics Computer Group (GCG).

XING, P., KULIKOWSKI, C., MUCHNIK, I., DUBCHAK, I., WOLf, D., SPENGLER, S. & ZORN, M. (1999). Analysis of ribosomal RNA sequences by combinatorial clustering. *Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology*. (Lengauer, T., Schneider, R., Bark, P., Brutlag, D., Glasgow, J., Mewes, H. W., & Zimmer, R., eds), pp. 287–296. Menlo Park, CA: AAAI/MIT Press.

APPENDIX

Homogeneity Measurement

We used h as a homogeneity measurement of each column in eqn (9). Here is an empirical explanation through a simple geometric approach:

Regarding h, we have the following equalities:

$$h = \sum_{l \in X} n_l^2, \quad \text{where } X = \{-, \mathrm{A}, \mathrm{G}, \mathrm{C}, \mathrm{U}\}, \qquad (\mathrm{A.1})$$

$$\sum_{l \in X} n_l = 1, \quad \text{where } n_l \geqslant 0 \text{ for } \forall l. \qquad (\mathrm{A.2})$$

Suppose $C = |X|$, eqn (A.2) defines a convex polygon in C-dimensional Euclidean space, and h corresponds to the distance from the origin to any point in it (Fig. A1). Obviously, the distance to the geometrical center $\{n_l = 1/C, \forall l\}$ of the polygon gives h_{min}. In terms of character distribution in a column of multiple alignment, this means that each type of character contributes equally, causing maximal heterogeneity. As the point moves away from the center to any of the axis, h increases monotonically until it reaches an extreme point $\{n_i = 1, n_l = 0, \forall l \neq i\}$ of the convex polygon, where h is maximized. This situation corresponds to the minimal possible heterogeneity of characters in a column: all characters belong to the same type "i". In reality, if there exists a predominant character "i" in a column, h and n_i has the following relationship:

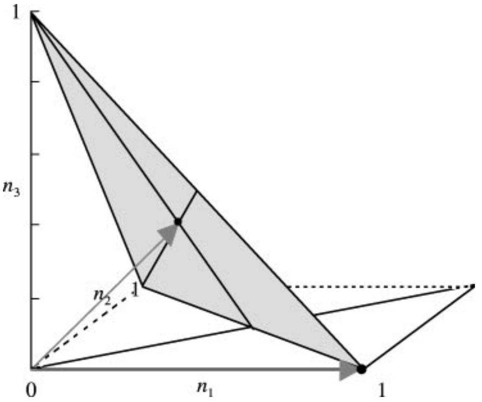

FIG. A1. Geometric illustration of the Character homogeneity function (A.1) in a three-dimensional Euclidean space, in which each dimension represents a character type "n_l". The shaded area corresponds to the convex polygon defined by function (A.2). The lengths of the red arrows correspond to specific values of h defined in eqn (A.1).

$$\begin{aligned} h = \sum_{l \in ?} n_l^2 &= n_i^2 + \sum_{l \in ? - i} n_l^2 \\ &\leqslant n_i^2 + \left(\sum_{l \in ? - i} n_l \right)^2 \qquad (\mathrm{A.3}) \\ &= n_i^2 + (1 - n_i)^2 = 2n_i^2 - 2n_i + 1. \end{aligned}$$

As an intuitive exemplification, this means, an h score of 0.8 roughly corresponds to the case in which a predominant character type "i" occurs in at least 90% of the rows for a column.

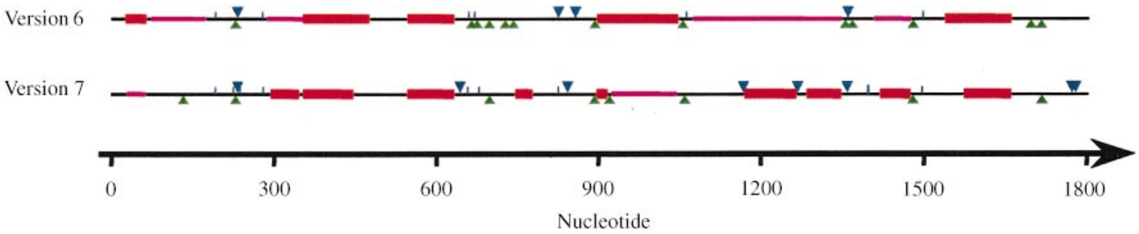

FIG. 3. Comparison of *H*-segmentation results for two different versions of multi-alignments of a same set of rRNA sequences. Three types of segments, type 1 (red bar for strict and magenta bar for marginal segment), type 2 (blue down triangle for major gaps segments (> 200 bp) and blue for short ones (< 150 bp)), and type 3 (green up triangle, were marked on the sequence of *Cryptococcus neoformans* small subunit rRNA (1806 bp, a member of the aligned sequence set) at the original locations where they reside.

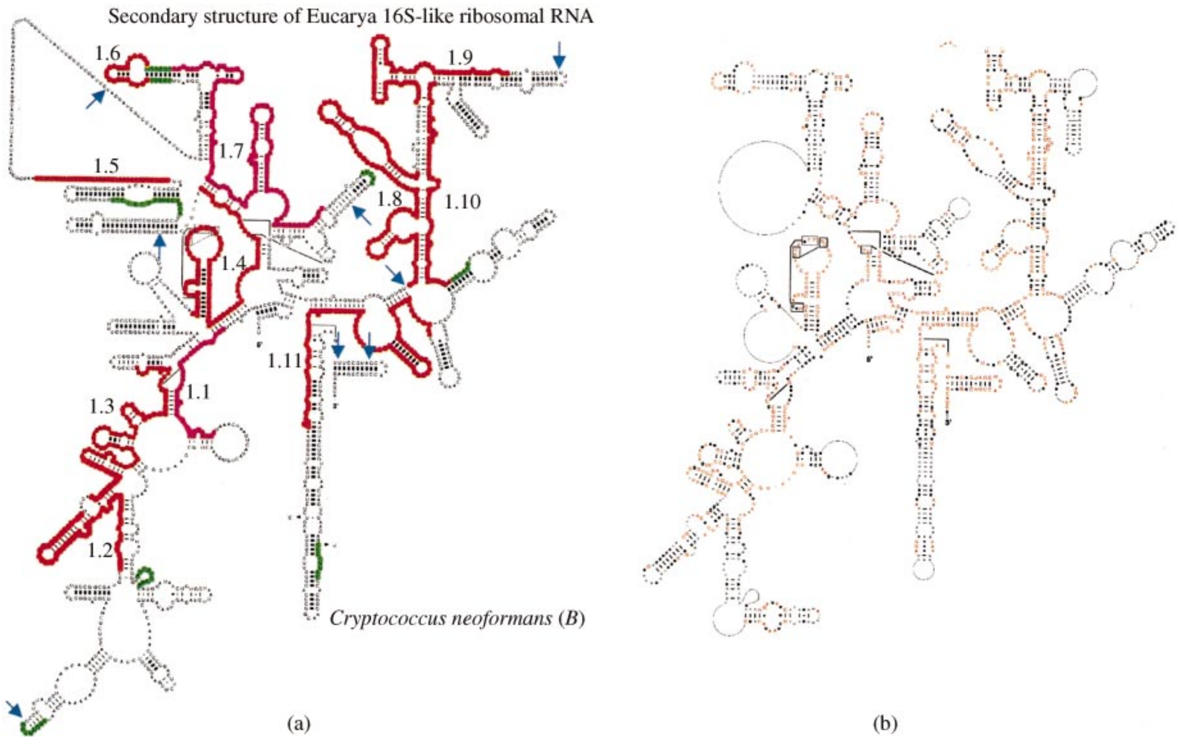
Secondary structure of Eucarya 16S-like ribosomal RNA
1.6
1.9
1.7
1.5
1.8
1.10
1.4
1.11
1.1
1.3
1.2
Cryptococcus neoformans (B)
(a)
(b)

FIG. 5. The mapping of the type 1 (red and magenta shade), type 2 (blue arrows, only for major segments) and type 3 (green shade) segments identified by *H*-segmentation to the secondary structure of the *Cryptococcus neoformans* small subunit rRNA. (a) Full structure, with the most stable thermodynamic foldings indicated by thick tick marks. (b) A structure diagram with phylogenetically conserved and variable structure labeled out. Both structure diagrams were originally from Gutell (1993).

doi.10.1006.jtbi.20012319