Analysis of ribosomal RNA sequences by combinatorial clustering

Poe Xing[1], Casimir Kulikowski[1], Ilya Muchnik[1], Inna Dubchak[2], Denise M Wolf[2], Sylvia Spengler[2], and Manfred Zorn[2]

[1]DIMACS and CS Department, Rutgers University, Piscataway, NJ; 08855-1179.
xingpoe@cs.rutgers.edu , kulikows@cs.rutgers.edu , muchnik@dimacs.rutgers.edu.

[2]National Energy Research Scientific Computing Center, Lawrence Berkeley National Laboratory, MS 84-171, Berkeley, CA., 94720, USA
ildubchak@lbl.gov , dmwolf@lbl.gov, sjspengler@lbl.gov, mdzorn@lbl.gov

Abstract

We present an analysis of multi-aligned eukaryotic and procaryotic small subunit rRNA sequences using a novel segmentation and clustering procedure capable of extracting subsets of sequences that share common sequence features. This procedure consists of: i) segmentation of aligned sequences using a dynamic programming procedure, and subsequent identification of likely conserved segments; ii) for each putative conserved segment, extraction of a locall homogeneous cluster using a novel polynomial procedure; and iii) intersection of clusters associated with each conserved segment. Aside from their utilit in processing large gap-filled multi-alignments, these algorithms can be applied to a broad spectrum of rRNA analysis functions such as subalignment, phylogenetic subtree extraction and construction, and organism tree-placement, and can serve as a framework to organize sequence data in an efficient and easily searchable manner. The sequence classification we obtained using the method presented here shows a remarkable consistency with the independently constructed eukaryotic phylogenetic tree.

1. Introduction

Ribosomes are complex pieces of subcellula machinery that catalyze protein synthesis. They ar composed of one small subunit, which is responsible for mRNA and tRNA binding, and one large subunit that catalyzes peptide bond formation. More than half of the weight of a ribosome is RNA, and increasing evidence shows that ribosomal RNA (rRNA) molecules play a central role in protein synthesis (Moore 1988, Spirin 1986). Due to their conserved nature, rRNA sequences are ideal objects fo phylogenetic analysis of genetic evolution. In addition, since rRNA is also a skeleton where man ribosome proteins are attached and exercise their catalytic or regulatory functions, detailed analysis of rRNA structure is also crucial for understanding the nature of RNA-protein interaction during ribosome assemblage and the regulation of protein synthesis.

Large alignments of ribosomal RNA are maintained at various sites, including the Ribosomal Database Project (RDP) database (Maidak et al. 1999). Since rRNA sequences vary greatly in their length and composition, the resulting multi-alignment is a large, complex, sparse matrix significantl interrupted by gaps. This fact makes an analysis of the entire rRNA alignment difficult and not even appropriate, given the diversity of sequences.

In the present study, multi-aligned eukaryotic and procaryotic small subunit rRNA sequences taken from RDP were analyzed using a novel segmentation and clustering procedure in an attempt to extract subsets of sequences that share common sequence features. This procedure allowed us to cluster the rRNA sequences using fragments potentially corresponding to essential functional units of rRNA molecules in different organisms. As they stand, the algorithms in this paper serve as an efficient way to take a large, unwieldy, gap filled multi-alignment and (1) optimally partition it into homogeneous segments, some of which may be large stretches of unambiguous alignment with high sequence identity and functional importance, (2) fo each segment of interest (not dominated by gaps and relatively conserved) determine which sequences belong to the cluster associated with the consensus sequence of the segment, and (3) further classif sequences based on the number and combination of segment-based clusters they belong to. Our algorithm enables application of a polynomial clustering procedure of $O(n^2)$ by using the special properties of the objective function defined on th conserved segments.

2. Clustering procedure for aligned ribosomal RNA sequences

2.1 Segmentation of multi-aligned sequences (an analog of procedures for image processing of Kittler and Foglein 1984).

We define a segment S_i of the multi-alignment A to be an ordered set of consecutive columns within the multi-alignment. The purpose of the segmentation step in the algorithm is to partition the multi-alignment into a number (k) of maximally homogeneous segments. This image processing based technique for segmenting a multi-alignment, presented below, combines column-wise statistical profile information like that used in (Gribskov et al 1997) with a dynamic programming approach often employed in alignment and model-fitting sequence segmentation algorithms (Auger and Lawrence 1989, Gorodkin et al 1997).

For a given multiple sequence alignment A and a predefined parameter k which represents the total number of segments to be produced after the segmentation, associate an k-segmentation $S = < S_1, \ldots, S_k >$ on A with a segmentation score function:

$$I(S) = \sum_{\alpha=1}^{k} F_\alpha \qquad (2.1.1)$$

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 b minimizing $I(S)$:

$$S^* = \arg\min_{|S|=k} I(S) \qquad (2.1.2)$$

Since F_α is a function 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 one of the segments delimited b l_a as its left boundary and r_a as its right boundary, $\alpha \in <1, \ldots, k>$. For any definition of F_α, the minimum of $I(S)$ can be found through a dynamic programming procedure (Bellman 1957, Mottl and Muchnik 1998) which progressively (from right to left) establishes optimum right boundary profiles $j_l^*(i)$ of the segment l for each possible left boundary i, together with their associated partial segmentation score:

$$\Phi_l^i = \min\left(\sum_{\alpha=l}^{k} F_\alpha\right) \qquad (2.1.3)$$

This procedure will terminate when the leftmost possible boundary i=1 is reached:

For l = k–1 to 1,

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

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

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

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

$$Q_l^i(j) = F_l(i,j) + \Phi_{l+1}^j \qquad (2.1.4)$$

$$\Phi_l^i = \min\left(\sum_{\alpha=l}^{k} F_\alpha\right) = \min_{j \in R_l^i}\left(Q_l^i(j)\right) \qquad (2.1.5)$$

$$j_l^*(i) = \arg\left(\min_{j \in R_l^i}\left(Q_l^i(j)\right)\right) \qquad (2.1.6)$$

The prodcdure terminates when $I(S^*) = \Phi_1^1$ is obtained. The time complexity of the procedure i $O(kn^2 G)$, where G is the cost for the calculation of Q in equation (2.1.4). 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 optimized 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 for the

boundaries of all the segments from left to right according to the following functions:

$$l_1 = 1, r_1 = j_1^*(l_1) \quad \ldots.$$
$$l_\alpha = r_{\alpha-1} + 1, r_\alpha = j_\alpha^*(l_\alpha), \quad \text{where,}$$
$$\alpha \in < 1,2,\ldots,k >$$

The final segmentation that results is:

$$S^* = < S_1(1, j_1^*(1)), S_2(j_1^*(1)+1, j_2^*(j_1^*(1)+1)), \ldots, S_k(j_{k-1}^*+1, j_k^*(j_1^*(1)+1)) >$$

Depending on the requirements for the features of the segmentation, various different types of segment-specific score function F_α can be chosen (based on concept of profile analysis (Gribskov, McLachlan and Eisenberg 1987)):

$$F_1(l_\alpha, r_\alpha) = \frac{1}{N \times |S_\alpha|} \sum_{i=1}^{N} \sum_{j=l_\alpha}^{r_\alpha} 1(g_{ij}), \qquad \text{where } 1(g_{ij}) = \begin{cases} 1, & \text{if gap is at the } j\text{th position of } i\text{th sequence} \\ 0, & \text{otherwise} \end{cases}$$

$$F_2(l_\alpha, r_\alpha) = \sum_{i=1}^{N-1} \sum_{i'>i} \sum_{j=l_\alpha}^{r_\alpha} R_{ii'}^j, \qquad \text{where } R_{ii'}^j = \begin{cases} 1, & \text{if the nucleotides at the jth position} \\ & \text{differ in sequences } i \text{ and } i' \\ 0, & \text{otherwise} \end{cases}$$

$$F_3(l_\alpha, r_\alpha) = \sum_{j=l_\alpha}^{r_\alpha} \left(n_j - \bar{n}_\alpha \right)^2 = \sum_{j=l_\alpha}^{r_\alpha} \sum_{l \in \{g,A,U,G,C\}} \left(n_j^l - \bar{n}_\alpha^l \right)^2, \quad \text{where vector } n_j = \left\{ n_j^g, n_j^A, n_j^U, n_j^G, n_j^C \right\}$$

represents the distribution of a gap and the four nucleotides at the j th position of the multialignment, and $\bar{n}_\alpha^l = \frac{1}{r_\alpha - l_\alpha + 1} \sum_{j=l_\alpha}^{r_\alpha} n_j^l$

F_1 measures the proportion of within-segment gaps for all sequences. F_2 represents the dissimilarit between all sequences as seen within the segment. F_3 is the variance of the frequencies of gaps and the four possible nucleotides at all the positions of the multi-aligned sequences within a segment. Depending on the choice among these as the objective function F_α in equations 2.1.1-2.1.5, the dynamic programming procedure described above will lead to a segmentation of the multi-alignment with alternative specific features. For example, when using F_1, one will obtain segmentation such that the sum of the average number of gaps at each position within each segment is minimized. F_2 leads to segmentation such that the dissimilarity of all the sequences within a segment is minimized. In the experimental work described below, we use F_3 as the objective function which leads to a segmentation in which the frequencies of gaps and the four nucleotides at each position within a segment are as uniform as possible.

2.2 Extraction of a locally homogenous group of sequences based on a single segment.

Once the multi-alignment has been segmented, the next step is to perform clustering on the segments of interest.

Methods of clustering can be classified into two general strategies (Arabie, Huber and De Soete 1996). On the one hand, one can analyze pair-wise relationships between objects, and group those close to each other into clusters according to some criterion. This method is particularly suitable for clustering data that are highly decentralized. On the other hand, one can start by finding the mean or center of the objects, and include in the "core" cluster all objects that are close to the mean according to some measurement and leave all the remaining ones as a "tail" cluster. This method is suitable for data with compact single central clusters. However, its performance is highly dependent on the choice of a proper threshold for characterizing the closeness of objects to the mean. A too liberal threshold will place most objects in the core cluster without sufficient guarantee of their closeness, whereas a too tight threshold will place most objects in the "tail" cluster. We have used a particular optimization procedure of the second strategy in which the threshold is defined

automatically (see a general description in (Kempner, Mirkin and Muchnik 1997)).

Generally, for the power set (all nonempty proper subsets) of $W:=\{1, \ldots, N\}$, 2^W, given an internal-linkage function $\pi(i, H)$, $H\in 2^W$, $i\in H$, which measures the similarity between an entit i and set H, a set function F_π can be defined on 2^W as follows:

$$F_\pi(H) := \min_{i\in H} \pi(i,H) \qquad (2.2.1)$$

$F_\pi(H)$ is referred to as the *minimum split function* for the internal-linkage function π. The *minimum split functio* measures the compactness of $H\in 2^W$, whose optimization is generally exponentially hard. However, it can be proved that when $\pi(i, H)$ is a monotonic linkage function, $F_\pi(H)$ has a special feature called *quasi-concavity* which enables a serial procedure that optimizes $F_\pi(H)$, $H\in 2^W-\phi$, in polynomial time (Kempner, Mirkin and Muchnik 1997). The "core cluster" mentioned above is defined as the subset H^* which gives a global maximum of the function $F_\pi(H)$:

$$H^* = \arg\max_{H\in 2^W} F(H) \qquad (2.2.2)$$

To cluster the multi-aligned sequences within a segment of choice in this study, we performed a three-step preprocessing of the sequence alignment: First, a consensus sequence of the alignment was composed by taking the most frequently occurred nucleotide at each position. Then, all sequences were transformed into binary codes by labeling a nucleotide position in the sequence as '1' if it is same as the consensus, and '0' otherwise. Finally, the cardinality (the number of '1's) of each sequence code was calculated. Based on these, we defined an internal-linkage function as following:

$$\pi(i,H) = \left|y_i\right| - a\left|Y^H\right|, \text{ where } Y^H = \cap_{i\in H} y_i,\ y_i \text{ is the binary code of sequence } i \text{ and } a \text{ is a constant.} \qquad (2.2.3)$$

The function $\pi(i, H)$ measures the degree of matching of sequence i to the consensus within a segment excluding the shared features of all sequences in the cluster that i belongs to. Therefore, $F_\pi(H)$ measures the minimum of pattern matches within H. Since $\pi(i, H)$ is a monotonically increasing function, $F_\pi(H)$ is therefore *quasi-concave*, so the polynomial serial procedure applies, as illustrated in Figure 1 below as three steps:

In step 1, all the objects y_i ($i\in<1, \ldots, N>$) in the whole set W are sorted by their cardinality $|y_i|$ and afterwards reordered in ascending order of $|y_i|$. In step 2, starting from the whole set $H_1 = W$, by each time leaving the first object y_{k-1} of the ordered list $y_{k-1}, \ldots, y_N$ (the one with the smallest cardinality) out and make the remaining objects a new subset $H_k=\{y_k, \ldots, y_N\}$, an ordered list of subsets $H_k \subseteq W$ ($k\in<1, \ldots, N>$) will be obtained and the value of the minimum split function of (2.2.1) corresponding to set H_k can be directly calculated

$$\text{as } F_\pi(H_k) = \pi(y_k, H_k) = \left|y_k\right| - a\left|Y^{H_k}\right|$$

because y_k is the object with smallest cardinality in set H_k. It can be shown that the global maximum of $F_\pi(H)$ exists amongst the list of $F_\pi(H_k)$ ($k\in<1, \ldots, N>$) obtained in step 2 (Kempner, Mirkin and Muchnik 1997). Therefore in step 3, the maximum was taken from the list of $F_\pi(H_k)$ and the corresponding set H_k identified thereby is the core cluster H^* we are looking for. The time complexit of this procedure is $O(N^2g)$, where N is the total number of sequences and g is the average time required to calculate $\pi(y_k, H_k)$.

2.3 Construction of a classification by combinatorial comparisons of homogeneou groups

The clustering described above is performed on each segment of interest (for example those with relatively few gaps) individually and independently. To stud the relationship between the clusters associated with different segments, we examined their intersection. Given L selected segments, we assigned each of the N sequences being analyzed with an L-bit occurrence label $b_1b_2\ldots b_L$. For each sequence label, b_i ($i\in<1, \ldots, L>$) was set to 1 if this sequence is in the homogeneous group determined from segment i, and 0 otherwise. The set of all patterns defined by the occurrence label therefore gives a sequence partition. If all sequence occurrence contributions are equall likely, there will be 2^L possible types of patterns. Thi is the basis for further clustering based on the degree of homogeneity revealed by the number of segments for which a sequence is considered a member of the locally homogeneous cluster.

All sequences were classified again according to the cardinality of the occurrence label associated with it, and fall into $L+1$ different clusters. A rank was assigned to each cluster according to the cardinality of the including sequences.

Dataset

RDP is a curated database that offers ribosomal RNA nucleotide sequence data in aligned and unaligned forms, analysis services, and associated computer programs (Maidak et al. 1999). The ribosomal RNA sequences in the RDP alignments are drawn from major sequence repositories (GenBank and EBI) and

Step I Sort Sequence by Cardinality

Step II Sequential Linkage Function Calculation

$\pi(i_1, H_1)$, where $H_1 = W$

$\pi(i_2, H_2)$, where $H_2 = W - i_1$

...

$\pi(i_N, H_N)$, where $H_N = W - \{i_1, i_2, \ldots, i_{N-1}\}$

where $\pi(i, H) = |y_i| - a| \cap y_i|$

Step III Maximization

$$F(H^*) = \max\{\pi(i_1, H_1), \pi(i_2, H_2), \ldots, \pi(i_N, H_N)\}$$
$$H^* = \{i_{h+1}, i_{h+2}, \ldots, i_N\}$$

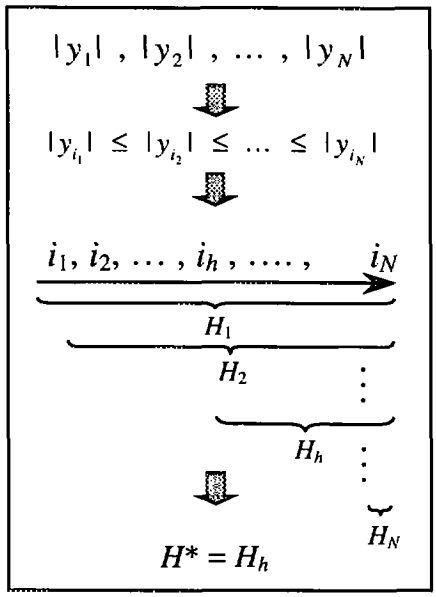


Figure 1. A serial procedure for optimizing the minimum split function $F_\pi(H)$.

from direct submissions. Currently, the multiple rRNA sequence alignment provided by the RDP database is achieved by a joint effort of computer optimization and manual validation and modification. For our study we used two multi-aligned sets of small subunit rRNA sequences, SSU_Euk and SSU_Proc containing 405 and 6205 sequences respectively (RDP data as of January 1998).

Results

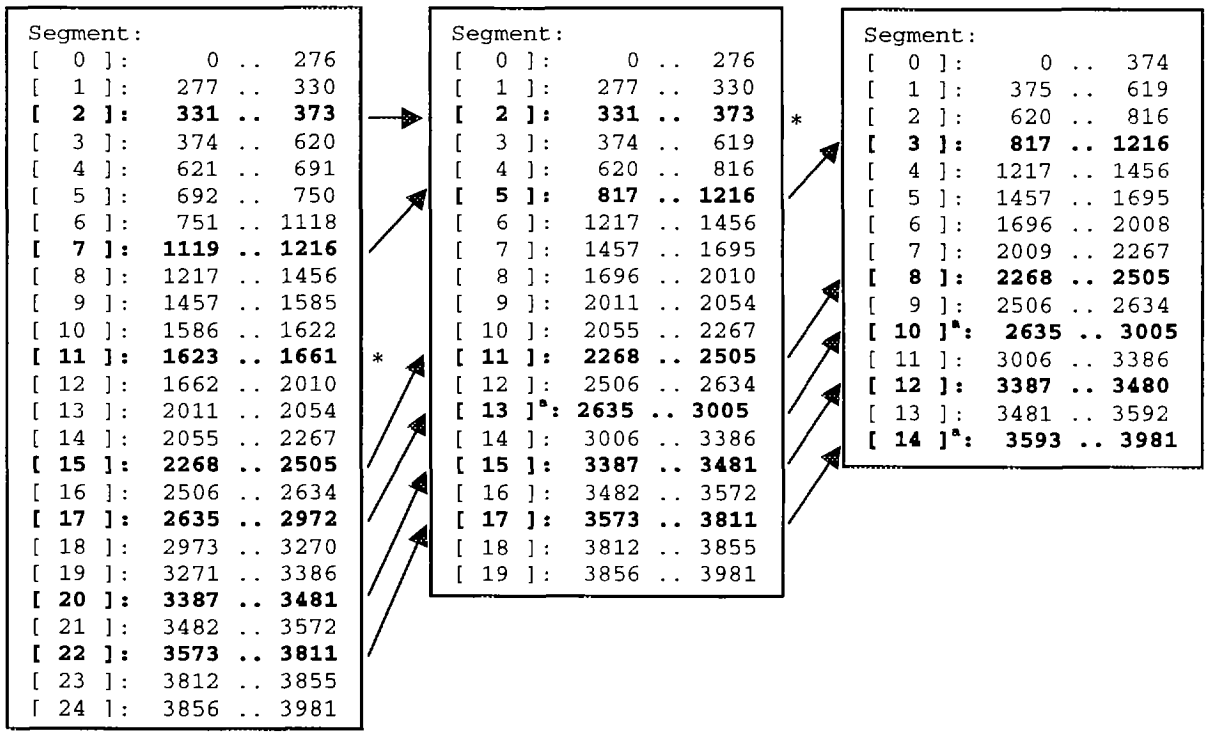
3. Analysis of eukaryotic ribosomal RNA sequences.

3.1 Segmentation of the multi-aligned sequences. Sequence segmentation was performed on the 3982 nucleotide-long multi-alignment covering 409 small subunit rRNA sequences, with the total number of segments prespecified as 25. We used F_3 described in section 2.1 in order to partition the multi-alignment into a number (k) of maximally homogenous segments. This segmentation was used to discriminate sequence segments mostly composed of gaps from those less frequently interrupted by gaps. As shown in Figure 2, the length of the resulting 25 segments varied from 34 to 367 nucleotides. Figure 3 shows the average frequency vector for each segment.

The average frequency of gaps within eac segment varied from 0.086 to 0.986. For th particular multi-alignment in RDP, a high frequency of gaps in a segment indicates that it is poorl conserved among different organisms. On the othe hand, a low content of gaps in a segment implies that for most of the organisms the sequences are uninterrupted or rarely interrupted which suggests it as a conserved unit. Our software has a free parameter "Gap_threshold" which allows the user to decide the stringency for accepting a segment as having a low gap occurrence. In our experiment, we considered a segment as having low gap occurrence if the gap frequency of the segment is lower than the average of the frequency of all the other four types of nucleotides (Gap_threshold = 0.2). Altogether, 7 out of the 25 segments met our criterion (as marked **bold** in Figure 2), and were therefore accepted as conserved segments. Their gap frequencies ranged between 0.086 to 0.165 (Figure 3). To test the robustness of the segmentation procedure, we performed the segmentation of the same multi-alignment for different levels of granularity by changing the k value, which leads to different numbers of segments being produced. Under three different k values (k=25, 20 and 15 respectively), the recognition of the conserved segments in the multi-alignment is relatively stable with respect to change in the granularity of segmentation.

Under a more coarse-grain segmentation (k=20 or 15), among the seven conserved segment identified from the 25 total segments, with the exception of two very short segments (segments 2 [43nt] and 11[39nt]), all other conserved segments

were recognized with little or no change in th boundaries. Even when k was reduced from 25 to 15, conserved segments with sufficient length were still well preserved. This result suggests that with a reasonable choice of the total number of segments, our segmentation procedure is able to identif relatively conserved segments from the multi-alignment with a high degree of robustness.



Segment:			Segment:			Segment:		
[0]:	0 ..	276	[0]:	0 ..	276	[0]:	0 ..	374
[1]:	277 ..	330	[1]:	277 ..	330	[1]:	375 ..	619
[2]:	**331 ..**	**373**	**[2]:**	**331 ..**	**373**	[2]:	620 ..	816
[3]:	374 ..	620	[3]:	374 ..	619	**[3]:**	**817 ..**	**1216**
[4]:	621 ..	691	[4]:	620 ..	816	[4]:	1217 ..	1456
[5]:	692 ..	750	**[5]:**	**817 ..**	**1216**	[5]:	1457 ..	1695
[6]:	751 ..	1118	[6]:	1217 ..	1456	[6]:	1696 ..	2008
[7]:	**1119 ..**	**1216**	[7]:	1457 ..	1695	[7]:	2009 ..	2267
[8]:	1217 ..	1456	[8]:	1696 ..	2010	**[8]:**	**2268 ..**	**2505**
[9]:	1457 ..	1585	[9]:	2011 ..	2054	[9]:	2506 ..	2634
[10]:	1586 ..	1622	[10]:	2055 ..	2267	**[10]^a:**	**2635 ..**	**3005**
[11]:	**1623 ..**	**1661** *	**[11]:**	**2268 ..**	**2505**	[11]:	3006 ..	3386
[12]:	1662 ..	2010	[12]:	2506 ..	2634	**[12]:**	**3387 ..**	**3480**
[13]:	2011 ..	2054	**[13]^a:**	**2635 ..**	**3005**	[13]:	3481 ..	3592
[14]:	2055 ..	2267	[14]:	3006 ..	3386	**[14]^a:**	**3593 ..**	**3981**
[15]:	**2268 ..**	**2505**	**[15]:**	**3387 ..**	**3481**			
[16]:	2506 ..	2634	[16]:	3482 ..	3572			
[17]:	**2635 ..**	**2972**	**[17]:**	**3573 ..**	**3811**			
[18]:	2973 ..	3270	[18]:	3812 ..	3855			
[19]:	3271 ..	3386	[19]:	3856 ..	3981			
[20]:	**3387 ..**	**3481**						
[21]:	3482 ..	3572						
[22]:	**3573 ..**	**3811**						
[23]:	3812 ..	3855						
[24]:	3856 ..	3981						

Figure 2. Segmentation of the multi-alignment fo k=25, 20 and 15 respectively. Conserved segments are marked **Bold**. Arrows indicate the correspondence of conserved segments identified for each segmentation, conserved segments that are missed in the lower-granularity segmentation are marked by a '*'. 'Gap_threshold' for all conserved segments is set to 0.2 except for that of segments marked b ^{a}, which is set to 0.3.

	[0]					[5]				
gap	0.346	0.939	**0.128**	0.984	0.377	0.986	0.202	**0.161**	0.890	0.251
A	0.191	0.012	**0.279**	0.003	0.144	0.003	0.223	**0.211**	0.015	0.169
U	0.202	0.015	**0.220**	0.004	0.183	0.004	0.170	**0.259**	0.031	0.216
C	0.128	0.019	**0.224**	0.005	0.148	0.003	0.230	**0.214**	0.033	0.202
G	0.133	0.015	**0.150**	0.005	0.148	0.004	0.175	**0.155**	0.031	0.162

[10]					[15]				
0.965	**0.106**	0.964	0.324	0.976	**0.141**	0.954	**0.165**	0.916	0.611
0.010	**0.315**	0.008	0.126	0.003	**0.261**	0.010	**0.209**	0.019	0.077
0.010	**0.254**	0.009	0.274	0.007	**0.222**	0.015	**0.217**	0.023	0.115
0.007	**0.186**	0.013	0.171	0.007	**0.228**	0.009	**0.244**	0.022	0.110
0.008	**0.139**	0.007	0.104	0.007	**0.148**	0.011	**0.164**	0.019	0.088

[20]					[25]
0.086	0.949	**0.149**	0.965	0.331	
0.227	0.012	**0.183**	0.005	0.207	
0.205	0.023	**0.230**	0.006	0.154	
0.266	0.009	**0.225**	0.011	0.184	
0.216	0.007	**0.213**	0.014	0.125	

Figure 3. The average frequency vector $\bar{n} = \{\bar{n}^g, \bar{n}^A, \bar{n}^U, \bar{n}^G, \bar{n}^C\}$ of each segment. The conserved segments are marked bold.

3.2 Building clusters associated with conserved segments.

Since only the most homogenous segments of the multi-aligned sequences were used for sequence clustering, we centered our clusters on consensus sequences. Using the clustering procedure described in 2.2, we analyzed the seven conserved segments of the multi-alignment obtained from 3.1. From each segment we obtained one "core" cluster and one complementary "tail" cluster. In the core cluster, all sequences are close to each other and also similar to the consensus sequence of the corresponding segment. For this reason, we call the core cluster a 'homogeneous group', and the tail cluster a 'heterogeneous group'. The sizes (the number of sequences) of the homogeneous groups derived from each segment are 284, 344, 361, 343, 366, 335, 317, respectively.

From this result, we can see that: 1) rRNA sequences are indeed highly conserved in eukaryotic organisms, which is consistent with the biological observation of a high degree of functional and structural conservation of this molecule among species. Among 409 sequences analyzed, the majority of the sequences belong to the homogenous groups, and 2) Segment 5 is probably the most conserved unit in the rRNA sequence because it defines the largest homogeneous group among all segments (366). In our procedure, it is possible to change the criterion for sequence clustering to control clustering stringency. For instance, the results listed above are obtained by using a=1, for the internal linkage function 2.2.3. It results in a largest possible homogeneous group that can be extracted from the entire sequence collection. However, with decreasing values of a, the stringency of clustering increases and tends to result in a smaller homogeneous sequence group, a subset of the homogeneous group corresponding to the larger a. By gradually reducing a from 1 to 0, it is possible to obtain a chain of noncontinuously shrinking homogenous groups, each corresponding to a certain range of the value of a where $a \in (0,1)$.

Although the clustering was carried out independently on 7 different segments of the multi-aligned sequences, the resulting partitioning of the sequences shared substantial similarities. This result supports the contention that conserved segments of sequences can produce consistent classifications of the sequence. It also indicates that the serial procedure we developed for the combinatorial clustering is stable for different data inputs.

3.3 Combination of clusters from individual conserved segments and ranking of the resulting sequence groups.

We performed the intersection of all clustering result on the 7 segments by labeling each sequence with an occurrence label as described in 2.4. Although there are 2^7, or 128 types of occurrence patterns possible logically, under a random assumption, only 33 patterns were observed among the 409 analyzed sequences, which indicates a significant deviation from a random sequence classification. To integrate clustering information from all conserved segments, we ranked each sequence according to its occurrence label, and aggregated them based on different ranks. We found that 249 of the 409 rRNA sequences fell into the group with the highest rank 7, which means they are homogeneous as determined by clustering of all the seven conserved sequence segments. There are 55 and 31 sequences in the clusters of rank 6 and 5 respectively, which also suggests a substantial homogeneity among these sequences. Thus using only conserved sequence segments during clustering greatly reduces the effect of random information from non-conserved or nonessential sequence fragments on the evaluation of relationships between sequences.

3.4 Comparison of clustering results with phylogenetic classification.

Our clustering of the 409 rRNA sequences is based on no prior knowledge of any biological classification of the sequences. Phylogenetic analysis of these sequences showed that they fall into 24 phyla of the eukaryotic organisms (Table 1). Interestingly, comparison of the phylogenetic classification of th rRNA sequences with our clustering results showed that each phylum usually corresponds to one or two major clusters that are adjacently ranked in ou analysis. When we referred to the evolutionary tree, we observed that highly homogeneous clusters (i.e. with rank 7, 6, and 5) correspond to organisms of more developed phyla, while highly heterogeneous clusters (lower ranked up to 4) correspond to organisms of more primitive phyla. Furthermore, the organisms in the most primitive phyla have a relatively simple cluster composition, whereas in those from more advanced phyla, the cluster composition is mixed across the ranks. This can be interpreted by a speculation that: 1) Different types of primitive organisms have more heterogeneous genetic background, 2) Higher organisms have mor shared genetic material, 3) Higher organisms (i.e. the metazoan phylum) develop functional diversity of the genes based on this shared genetic background.

Analysis of prokaryotic ribosomal RNA sequences (in progress).

All three steps of the analysis procedure, consisting of i) sequence segmentation and identification of likely conserved segments, ii) clustering of sequences, based on each conserved segment, and iii) intersection of clustering results from all the conserved segments, wer also performed on 3320 nucleotide-long multi-alignment covering 6205 small subunit prokaryotic rRNA sequences. The sequence of rRNA is not highly conserved in a majority of prokaryotes, which is consistent with the biological observation of a high degree of functional and structural diversity among species. Although the level of sequence conservation in procaryotic rRNA is significantly lower than in eukaryotes, our procedure still allowed us to reveal conserved segments and to analyze them. The segmentation procedure was performed on the data three times at three different levels of granularity and the position of the conserved segments in the multi-alignment was relatively stable. Values of average gap content in obtained optimal segments were not lower than 0.167 (16.7% gaps in the segment of aligned sequences) and reached 0.999 in some cases (these segments consist entirely of gaps). Only for nine segments was the gap frequency lower than 0.4. We accepted this value as a parameter defining relatively conserved segments (Gap_threshold = 0.4) and used these segments in the clustering section of our procedure. The size of the homogeneous groups derived from each segment were 3838, 3343, 2378, 2447, 4312, 2641, 1491, 837, and 3179. In spite of the low conservation level, clusters derived from two of the nine segments, joined together 70% and 53.8% of 6205 sequences.

It is clear that the clusters resulting from 9 different conserved segments are not very consistent. We performed the intersection of all clustering results on the 9 segments as described in 2.4. Although there are 2^9, or 512 occurrence patterns logically possible, only 320 patterns were observed, and among those only 6 patterns were shared by more than 3% of sequences (380, 163, 235, 285, 250, 243 respectively). We found that 59 sequences fell into the group with the highest rank: 9, which means they are homogeneous as determined by clustering of all the nine conserved sequence segments. There are 415, 705 and 940 sequences in the clusters of rank 8, 7 and 6 respectively.

Discussion and Future Directions

In this paper we presented a segmentation and clustering algorithm for sequence multi-alignments and its application to eukaryotic and prokaryotic small subunit rRNA analysis. This algorithm consists of i) identification of likel conserved segments using segmentation of aligned sequences; ii) for each putative conserved segment, extraction of a locally homogeneous cluster; and iii) intersection of clusters associated with each conserved segment. This novel polynomial-time procedure was shown to be capable of efficientl extracting subsets of sequences that share common sequence features and to map informatively onto a phylogenetic tree.

Large alignments of ribosomal RNA sequences are maintained at various sites (Van de Peer et al. 1998, Maidak et al. 1999). New sequences are added to these alignments using a combination of manual and automatic methods. These alignments ar large and complex and, beside the difficulty of comparative analysis of individual rRNAs, it is extremely hard to add new sequences to existing alignments (O'Brien, Notredame and Higgins 1998). The RDP (Maidak et al. 1999) is the main source of rRNA data and tools for its analysis. It contains at present 9700 aligned small subunit rRNA sequences and 22,000 unaligned sequences. The following RDP tools are considered to be necessary, and are widel used: subalignment, extraction of a portion of a phylogenetic tree, sequence probe checking, extraction of a most similar sequence, sequence alignment, and placement of a organism on a subtree.

The procedures described in this paper can play an important role in the development of the next generation tools fo rRNA processing, database searching and biological interpretation. They can be applied to the above functional areas in the following manner: 1) After choosing a segment of a specific sequence as a pattern, one can obtain a cluster composed of sequences closely related to a given sequence within the chosen segment, i.e. this procedure provides a rational approach to the extraction of significant subalignment . 2) By unifying or intersecting different sequence clusters in the database, one can create all the nodes and edges to construct a tree corresponding to the hierarchy of the clusters and, accordingly, sequences. 3) Th development of an objective and biologicall supported approach for the multi-alignment of sequences is a major challenge (Thompson et al 1994, O'Brien et al 1998 and references therein). We will explore new methods of partial alignments based on specific sequence patterns, which are extracted from a general multi-alignment, as well as structural motifs. The clustering procedure described in this paper can potentially be used for such a purpose as well as for classifying sequences in general. 4) As mentioned above, our clustering procedure facilitates phylogenetic tree construction by providing a

Table 1. Comparison between the inter-cluster homogeneity ranking of eukaryotic 18S rRNA sequences from combinatorial clustering and the taxonomical classification of the species *

Taxonomical classification	EUMYCOTA	METAZOA AND RELATIVES	CHLORARACHNIOPHYCEAE	CHLOROPHYTA-EMBRYOPHYTA GROUP (PLANTS AND GREEN ALGAE)	HAPTOPHYCEAE	RHODOPHYTA (RED ALGAE)	STRAMENOPILES	ALVEOLATA	CRYPTOPHYTA	HARTMANNELLIDAE	ACANTHAMOEBIDAE	HAPLOSPORIDA	DICTYOSTELIIDA (CELLULAR SLIME MOLDS)	EUGLENOZOA	PHREATAMOEBIDS	ENTAMOEBIDAE	HETEROLOBOSEA	MYXOMYCETE (ACELLULAR SLIME MOLDS)	PARABASALIDEA	DIPLOMONADIDA	MICROSPORIDIA
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
Homogeneity Rank 7	96%	64%		89%	100%		65%	34%	33%	100%	50%										
6		8%	100%	5%		22%	35%	34%	67%		50%										
5	3%	8%				22%		24%													
4	1%	5%		1%		56%		7%				100%			100%						
3		9%		1%									100%								
2		4%		3%																	
1														6%		100%	100%	100%			
0		1%												94%					100%	100%	100%
totai	91	75	4	75	1	9	23	82	3	1	2	1	1	16	1	3	4	1	1	5	10

*Evolutionary tree is adapted from (Alberts et al. 1995)

clustering hierarchy for all the sequences. It can also interpret a given phylogenetic tree by correlating each node with a particular sequence pattern and its associated homogeneous cluster.

Aside from being useful in a practical sense for sequence clustering, sub-alignment and tree construction, we are interested in seeing whether the identified `conserved segments' have any functional or structural significance. To this end we plan to scour the literature and various databases fo functional correlations, and to thread the consensus sequences of each segment through known and predicted ribosomal structures.

The segmentation and clustering algorithms described in this paper can be applied to testing and developing evolutionary hypotheses, both on th micro and macro level, and to investigating the role of modularity in the evolution of rRNA. Once the conserved segments have been calculated, along with their associated clusters, we can explore evolutionary hypotheses on the intra- segment level. Applying Maximum Likelihood techniques allows for th determination of a `best' evolutionary hypothesis to explain the data in each segment cluster. We plan to see how (and if) `best' hypotheses vary from segment to segment, and to compare these results to the combinatorial clustering results and the segment-level evolutionary scale calculations described below.

Proposed techniques can be used to explore th nature and (if quasi-discrete) number and distribution of evolutionary scales within a single molecule. On possible experiment is to calculate an estimate of the mutational rates of each relatively conserved segment along the molecule, along with those of the interspersed, less conserved regions. We can then quantify this distribution of mutational rates (assuming a non-random one), and look for a set of quasi-discrete evolutionary modes.

References

Alberts, B., Bray, D., Lewis, J., Raff, M., Roberts, K., and Watson, J. 1994. *Molecular Biology of The Cell*, 3rd Edition, New York: Garland Publishing Inc.

Auger, I. E., and Lawrence, E.L. 1989. Algorithm for the optimal identification of segment neighborhoods. *Bulletin of Math. Bio.*,Vol 51, No 1: 39-54.

Arabie, P., Hubert, L., and De Soete, G. (eds.).1996. *Classification and Clustering*. River Edge, NJ: World Scientific Publishiers.

Bellma , R. 1957. *Dynamic Programming*. Princeton: Princeton Univ. Pres.

O'Brien, E., Notredame, C., Higgins, D.G. 1998. Optimization of ribosomal RNA profile alignments. *Bioinformatics*, 14:332-341.

Gorodkin, J., Heyer, L.J., and 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., and Eisenberg, D. 1987. Profile analysis: detection of distantly related proteins. *Proc Natl Acad Sci U S A*. 84:4355-4358.

Kempner, Y., Mirkin B, and Muchnik I. 1997. Monotone linkage clustering and quasi-concave set functions. *Applied Mathematics Letters*. 10:19-24.

Kittler, J. and Foglein, J. 1984. *Image & Vision Computing*. 2: 13-29.

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., and Woese, C. R. 1999. A new version of the RDP (Ribosomal Database Project). *Nucleic Acids Res*. 27:171-173. (http://www.cme.msu.edu/RDP/)

Moore, P. B. 1988. The ribosome returns *Nature*. 331:223-7.

Mottl, V. V. and. Muchnik, I. 1998. Bellman functions on trees for segmentation, generalized smoothing. matching multi-alignment in massive data sets *DIMACS, Technical Report*. 17-98:1-54.

Spirin, A. S. 1986. *Ribosome Structure and Protein Synthesis*. Menlo Park, CA: Benjamin-Cummings.

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

Van de Peer, Y., Caers, A., De Rijk, P., and De Wachter, R. 1998. Database on the structure of small subunit ribosomal RNA. *Nucleic Acids Res.* 26:179-82. (http://www-rrna.uia.ac.be/).