Visualization of Microorganism Classification on Power Spectrums of Complete DNA Sequences Using Clustering Analysis and Multidimensional Scaling

Cheng-Chang Jeng¹, I-Ching Yang¹, Ting-Jia Lain¹, Kun-Lin Hsieh¹, and Chun-Nan Lin²
¹Systematic and Theoretical Science Research Group
²Department of Management Information System
¹National Taitung University
²National Chung Cheng University
684 Chunghua Rd., Sec. 1, Taitung, Taiwan
168, University Rd., Min-Hsiung Chia-Yi, Taiwan

Abstract: - The coding structure of nucleotide sequences can be investigated and visually observe via power spectrums obtained by applying Fourier transforms to complete DNA sequences. The proposed method also provides dimension reduction for complete DNA sequences. To construct tree-like diagram, which looks like a taxonomy tree, hierarchical cluster analysis (HCA) was applied to transformed DNA sequences. Also, multidimensional scaling (MDS) was utilized to help visually investigate relationships among microorganisms in a low-dimensional space. The combination of the analysis of power spectrums on complete DNA sequences with HCA and MDS is suggested to examine coding structure and clustering of 15 microorganisms, which consists of some members of Escherichia coli, Pyrococcus, and Bacillus genera. The results of clustering and perceptual map from the combined mathematical model correspond with the recent findings found by others. Especially note that, with the proposed method, Pyrococcus furiosus is not within the lineage of Pyrococcus abyssi and Pyrococcus horikoshii, and also Bacillus halodurans C-125 is not within the lineage of Bacillus cereus genus that consists of Bacillus anthracis and Bacillus cereus family. In conclusion, the results constructed by HCA and MDS on the power spectrums of complete DNA sequences provide biologists mathematical tools to verify the taxonomy of microorganism and speed up analytic process.

Key-Words: - Microorganism classification, hierarchical clustering analysis, multidimensional scaling

1 Introduction
Transformation of nucleotides in DNA sequences with Fourier transform have been widely adapted to investigate differences among microorganisms [1-3]. With the transformed DNA sequences, researchers also indicated that some certain periodicities, such as in 3 base-pairs (bp) and in 10-11 bp, may reveal coding structure of protein and DNA folding [4-6]. There are some research that focus on the periodicities 10-11 bp in the power spectrums of nucleotides to perform bacteria classification [7-8]. Those results roughly show bacteria classification can be done by comparing the differences with periodicities in the power spectrums of nucleotide sequences. However, those processes of bacteria classification are rather heuristic than systematic; it is highly depended on human judgments rather than some ways of auto-recognition.

With implementations of bacteria classification, the results may lead to the understanding of the origin of living microorganisms [9]. Since microorganisms may differ on the power spectrums of nucleotide sequences that reflect the topological state of genomic DNA, there is an immediate interest to investigate a systematic procedure for classifying microorganisms, and a systematic procedure for classifying microorganisms will be viewed as a meaningful issue to be studied. In this study, two multivariate statistical methods, hierarchical cluster analysis (HCA) and multidimensional scaling (MDS), were suggested to be carried out on power spectrums of nucleotides in complete DNA sequences of 15 microorganisms in order to construct a tree-like taxonomy diagram of the 15 microorganisms and visualized distances among the microorganisms. The primary reason for employing HCA is intended to obtain the information about the similarity or dissimilarity among microorganisms and construct a tree-like diagram. Also, applying
MDS to the classification of the microorganisms may be complementary to the result of HCA. It is able to geometrically interpret the classification of the microorganisms. To demonstrate the feasibility and rationality of the proposed procedure, in the next section on Materials and Methods, the complete DNA sequences of microorganisms used in this study and the proposed systematic procedure will be described in detail.

2 Materials and Methods

For verifying the rationality and feasibility for our proposed procedure, the three different families, which consist of 4 Escherichia bacteria, 3 Pyrococcus archaea and 8 Bacillus bacteria, of complete DNA sequences with accession numbers obtained from www.ncbi.nlm.nih.gov are listed in Table 1.

2.1 Transformation of DNA Sequences

A DNA sequence can be regarded as a combination of 4 nucleotides: A (adenine), C (cytosine), G (guanine), and T (thymine). Therefore, a DNA sequence with N nucleotides can be initially represented as a vector:

\[ D = [w_k] ; k = 0...N-1 \]

where \( w_k \) denotes the symbol of the kth nucleotide. Furthermore, D can be translated to numerical form \( E_d \) by

\[ E_d = [x_d] ; x_d = 1 \text{if} \ w_k = d, \text{or} x_d = 0 ; k = 0...N-1 \text{and} d \in \{A,C,G,T\} \]

Based on eq. (2), it will be easy to obtain \( E_A, E_C, E_G \) and \( E_T \) for nucleotides A, C, G and T, respectively [10]. According to eq. (1), for example, the vector D for an organism can be represented as


Then the vector D for the organism can be further transformed to a binary form \( E_A \) by applying eq. (2). So, \( E_A = [1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0] \) can be obtained.

The power spectrum for the numerical representation of nucleotides \( E_d \) is defined as

\[ P_d(f_j) = |V_d(f_j)|^2 = V_d(f_j)V_d^*(f_j), \]

where

\[ V_d(f_j) = \frac{1}{N} \sum_{k=0}^{N-1} x_d e^{-2\pi ikf_j} \]

is the Fourier transform of \( E_d \), and the frequency \( f_j \) is calculated by

\[ f_j = \frac{j}{N}, j = 0...N-1, \]

and therefore the corresponding periodicity \( g_j \) for \( f_j \) is

\[ g_j = \frac{1}{f_j}. \]

For a detailed analysis of a complete genomic sequence, it is necessary to consider not only the coding sequences, but also the noncoding sequences, which are extracted from the noncoding sequences on the forward and backward strands. For a coding sequence on one of the DNA strands, the sequences on the complementary strand which have common intersecting regions with the coding sequence are necessarily coded. Thus, a complementary noncoding sequence contains a coded region with periodic structure of the coding sequence. Here, the noncoding sequences are defined as the common intersecting regions of the noncoding sequences on both strands. To analyze noncoding sequences, \( N = 2048 \) is taken in to the computation if sequences have a length of 2048 bp or larger. Then, \( N=2048 \) is used to slice a complete DNA sequence into small pieces. Let us denotes the average power spectrum of nucleotide \( d \) for the \( n \)th organism in Table 1. The matrix of average power spectrum for the 15 organisms can be obtained by the eq. (7)

\[ Q_d = [P_{dn}'^T]. \]

Because the power spectrums converted from Fourier transform is symmetric, which means that it is unnecessary to use the all 2048 power spectrums, the matrix \( Q_d \) shall be a \( 1024 \times 15 \) matrix with the consideration of half number of power spectrums. After \( Q_d \) is obtained, the visual representation of \( Q_d \) can be obtained by directly map \( Q_d \) into two-dimensional space.

2.2 Applying HCA and MDS to Classification

For a better understanding of distances and structural differences among the 15 microorganisms, the procedure of hierarchical clustering analysis (HCA) is carried out on all microorganisms, which denotes the average power spectrum of nucleotide \( d \) for the \( n \)th organism in Table 1. HCA may be helpful here, because it is capable to partition data into arbitrary groups without any a-priori knowledge and evaluate the distances among data and groups to investigate the similarity or dissimilarity [11-12]. The general procedure for performing HCA can be found in Anderberg’s publication [11].
With HCA on the transformed DNA sequences, a tree-like diagram, which looks like a taxonomy tree, can be constructed. However, it is not easy to identify the distances among the microorganisms in a quick glance with the tree-like diagram. To geometrically illustrate the classification, another multivariate statistical method multidimensional scaling is applied to construct a perceptual map. A perceptual map is used to determine the perceived relative graph of a set of objects. In this study, objects can be transformed DNA sequences. Suppose that microorganisms A and B are transformed by the method proposed in this study, and they are the most similar compared with all other possible pairs of microorganisms. In a perceptual map, microorganisms A and B would like to be positioned so that the distance between them in multidimensional space is the smallest among the distance between any other two pairs of microorganisms. After microorganisms A and B are positioned, with the exception of A and B, the other most similar pair of microorganisms are selected to be positioned in the perceptual map until all the microorganisms are relatively positioned. Although DNA sequences are multidimensional, with the concept of perceptual map relationships among microorganisms can be visually investigated in a low-dimensional space.

In this study, a multidimensional scaling (MDS) method, the PROXSCAL algorithm, from SPSS v.13 was used to construct perceptual maps [13]. In particular, standardized residual sum of squares (STRESS) are used to determine the measures of closeness to the original distances among objects. The goal of PROXSCAL algorithm is to minimize the measure of STRESS called normalized raw STRESS, also known as badness of fit index, i.e. the lower the value the better the fit. The value of normalized raw STRESS is from 0 to 1. In general, if the value of normalized raw STRESS below 0.05, it is considered to be good.

In the next section, the results of HCA and MDS on the transformed DNA sequences, average power spectrums, of the 15 microorganisms will be shown and discussed.

3 Results

Table 2 represents the numerical summary of HCA. The distance between two closest clusters in each merging stage is also listed in Table 2. As shown in Table 2, there is a sudden jump of dissimilarity measure between stage 7 and stage 8. It clearly shows that Bacillus subtilis (B1) and Bacillus licheniformis (B4) are not combined together until stage 8 is performed. On the other hand, Pyrococcus furiosus (P2) is grouped with other 2 archaea right after Pyrococcus abyssi (P1) and Pyrococcus horikoshii (P3) is combined together as late as in the stage 9. And, finally, Bacillus halodurans C-125 (B8) is not grouped with others until stage 11.

It is possible to illustrate Table 2 graphically. The visual representation of Table 2 is depicted in Fig. 1. The tree-like diagram is called a dendrogram of HCA solution. As shown in Table 2 and Fig. 1, Fig. 1 is a homogeneous and graphical version of Table 2. However, the dendrogram shown in Fig. 1 can visually suggests the clustering of the 15 microorganisms. Observing Fig. 1 from right to left, before stage 13, which splits the microorganisms into 2 big clusters, the second level of joinings before stage 11 separates the microorganisms into 4 clusters, which are the family of E. coli (E1-E4), family of Pyrococcus archaea (P1-P3), group formed by Bacillus subtilis (B1) and Bacillus licheniformis (B4), and group of other Bacillus members (B2, B3, and B5-B8).

Notice that there are some big gaps among the joinings. A “big gap” usually suggests two clusters are being forced to be grouped into a single cluster, i.e. not a good clustering suggestion. The first gap can be observed between stage 6 and stage 12. Therefore, the gap between stage 6 and stage 12 splits the family of E. coli (E1-E4) as well as the group formed by Bacillus subtilis (B1) and Bacillus licheniformis (B4) into 2 clusters. The second gap can be found between stage 7 and stage 11, which suggests Bacillus halodurans C-125 (B8) is an outlier from the group formed by the two families: the family of Bacillus anthracis (B2, B3, and B5) and the family of Bacillus cereus (B6 and B7). If the joinings of the two big gaps are removed, the 15 microorganisms will be categorized into 5 clusters. Notice that all of the archaea are grouped together at this time. If the third gap, where can be found at the joining of Pyrococcus archaea between Pyrococcus furiosus (P2) and the group formed by Pyrococcus abyssi (P1) and Pyrococcus horikoshii (P3), is removed, the clustering schema will suggest 6 clusters for the 15 microorganisms. Note that P2, which is Pyrococcus furiosus, will be the single member of its cluster. This result suggests that Pyrococcus furiosus (P2) is quite far away from Pyrococcus abyssi (P1) and Pyrococcus horikoshii (P3).

After HCA is performed, Fig. 2 shows the two-dimensional perceptual map constructed by MDS.
Proceedings of the 10th WSEAS International Conference on COMPUTERS, Vouliagmeni, Athens, Greece, July 13-15, 2006 (pp1308-1315)

with PROXSCAL algorithm. The index normalized raw STRESS is 0.009, which indicates that the perceptual map well represents the distances among the 15 microorganisms. As shown in Fig. 2, comparing the distances of the other three Bacillus bacteria B1, B4, and B8, bacteria B2, B3, B5, B6 and B7 are relatively close. There is another big group located on the right side of the perceptual map. The group is formed by the bacteria of E. Coli genus. Then, the archaea of Pyrococcus genus are dispersed widely in the upper middle portion of the perceptual map. Nonetheless, the results constructed by MDS is quite corresponding to the dendrogram constructed by HCA.

3 Discussion

Based on the discussion and figures constructed in the section on Results, above, the recommendations of clustering for the 15 microorganisms can be categorized into several clusters. They are:

(i) Cluster 1: The family of E. coli, which are Escherichia coli CFT073 (E1), Escherichia coli K12 (E2), Escherichia coli O157:H7 (E3), and Escherichia coli O157:H7 EDL933 (E4).

(ii) Cluster 2: The cluster formed by 6 bacteria, which are the 3 members of Bacillus anthracis family (B2, B3, and B5), and 2 members of Bacillus cereus family (B6 and B7).

(iii) Cluster 3: The cluster formed by Bacillus subtilis (B1) and Bacillus licheniformis (B4) and Bacillus halodurans C-125 (B8). Although in Fig. 1 the dendrogram suggests that B8 is grouped with Cluster 2, the perceptual map, Fig. 2, constructed by MDS shows that B8 shouldn’t be considered as a member of Cluster 2. Note that the members of this cluster, Cluster 3, are rather far away from each other.

(iv) Cluster 4: The family of Pyrococcus, which are Pyrococcus abyssi GE5 (P1), Pyrococcus furiosus DSM 3638 (P2), and Pyrococcus horikoshii OT3 (P3).

The above clustering recommendation gives the effect of being reasonable, but Cluster 3 formed by Bacillus subtilis (B1), Bacillus licheniformis (B4), and Bacillus halodurans C-125 (B8) seems unusual. Moreover, B1 is close to B4 and far away from B8, if the clustering suggestions provided by HCA and MDS are considered. In the corresponding research results stated by Rey and his colleague [14], with the analysis of rRNA and tRNA on Bacillus subtilis and Bacillus licheniformis, the two bacteria have extensive organizational similarity. Rey and his colleague’s findings, which validate the effectiveness of the methods proposed in this study, demonstrate that clustering Bacillus subtilis and Bacillus licheniformis together is reliable, i.e. HCA and MDS with the power spectral method can be used to verify structural differences among microorganisms.

Observing the dendrogram and perceptual map, especially comparing the distance among P2 and the other two Pyrococcus archaea, the two figures suggest that P1 and P3 is relatively close to each other. Although Pyrococcus furiosus (P2) is a member of Pyrococcus genus, however, according to Lecompte and his colleague’s research findings [15], they state that based on the analysis of the amino acid substitution rate there is no difference between Pyrococcus abyssi (P1) and Pyrococcus horikoshii (P3) for most of their shared proteins, even for fast-evolving ones. But, the substitution rates of amino acid observed in Pyrococcus furiosus (P2) are rather different from the other two Pyrococcus archaea. They strongly assume that Pyrococcus furiosus (P2) is not within the lineage of Pyrococcus abyssi (P1) and Pyrococcus horikoshii (P3). Our clustering results are comparable to Lecompte and his colleague’s findings that there is a big gap between Pyrococcus furiosus and the other two archaea.

On the other hand, Cluster 2 consists of the 3 members of Bacillus anthracis family (B2, B3, and B5), and 2 members of Bacillus cereus family (B6 and B7), which 5 microorganisms are members of the Bacillus cereus family, and have been shown that Bacillus anthracis should be considered as a lineage of Bacillus cereus by multilocus enzyme electrophoresis and by sequence analysis [16]. Note that Bacillus halodurans C-125 (B8) is not within the lineage of Bacillus cereus family from the observation of figures constructed by HCA and MDS. The results obtained from our clustering method are corresponding to the results found by different methods of biological research.

4 Conclusions

In sum, from the above discussion and related literature survey, it is found that the rationality and feasibility of the proposed methods well discriminate the 15 microorganisms. The methods and results discussed in this study state some significance that is described as follows:

(1) The technique of power spectrum analysis effectively shrinks the dimensions of complete bacteria DNA sequences for being the
input data of HCA. For example, in Table 1, the no. 1 organism Escherichia coli CFT073 (E1) gets 5231428 bps, that there are 5231428 nucleotides in E1’s DNA sequence. After transformed only with one of the 4 nucleotides, which is ‘A’ (adenine) in this study, by the technique introduced previously, the dimension of E1 is as small as 1024.

(2) As stated in the previous section, compared with the recent biological findings, categorizing the power spectrums of bacteria DNA sequences using HCA and MDS is reliable and effective. In other words, the clustering results of our mathematical model correspond with the recent findings found by biologists.

(3) The results constructed by applying HCA and MDS to the power spectrums of complete DNA sequences provide supplements to verify the taxonomy and DNA distances of microorganisms.

Briefly, in this study an illustrative example is provided to demonstrate the complete procedure which combining the dimension reduction, data transformation, and clustering analysis on microorganisms. Researchers are able to adopt the proposed procedure to their future studies. However, due to the number of bacteria is rather small in this study, investigations on more microorganisms with complete nucleotides should be conducted in future research.

Acknowledgments
This research was supported by the National Science Council of Taiwan under contract number NSC 94-2416-H-143-001 and NSC 94-2112-M-143-001.

References:

Table 1. The three different groups of 15 microorganisms

<table>
<thead>
<tr>
<th>No.</th>
<th>Organism Label</th>
<th>Accession No.</th>
<th>Bases(bps)</th>
<th>Taxo. ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>E. coli CFT073</td>
<td>AE014075</td>
<td>5231428</td>
<td>199310</td>
</tr>
<tr>
<td>2</td>
<td>E. coli K12</td>
<td>U00096</td>
<td>4639675</td>
<td>83333</td>
</tr>
<tr>
<td>3</td>
<td>E. coli O157:H7</td>
<td>BA000007</td>
<td>5498450</td>
<td>83334</td>
</tr>
<tr>
<td>4</td>
<td>E. coli O157:H7</td>
<td>AE005174</td>
<td>5528445</td>
<td>155864</td>
</tr>
<tr>
<td>5</td>
<td>Pyrococcus abyssi GE5</td>
<td>AL096836</td>
<td>1765118</td>
<td>272844</td>
</tr>
<tr>
<td>6</td>
<td>Pyrococcus furiosus DSM3638</td>
<td>AE009950</td>
<td>1908256</td>
<td>186497</td>
</tr>
<tr>
<td>7</td>
<td>Pyrococcus horikoshii OT3</td>
<td>BA000001</td>
<td>1738505</td>
<td>70601</td>
</tr>
<tr>
<td>8</td>
<td>B. subtilis subsp. Subtilis</td>
<td>AL009126</td>
<td>4214630</td>
<td>224308</td>
</tr>
<tr>
<td>9</td>
<td>B. anthracis A2012</td>
<td>AAC01000001</td>
<td>5093554</td>
<td>191218</td>
</tr>
<tr>
<td>10</td>
<td>B. anthracis str. 'Ames Ancestor'</td>
<td>AE017334</td>
<td>5227419</td>
<td>261594</td>
</tr>
<tr>
<td>11</td>
<td>B. licheniformis ATCC14580</td>
<td>CP000002</td>
<td>4222334</td>
<td>279010</td>
</tr>
<tr>
<td>12</td>
<td>B. anthracis str. Ames</td>
<td>AE016879</td>
<td>5227293</td>
<td>198094</td>
</tr>
<tr>
<td>13</td>
<td>B. cereus ATCC10987</td>
<td>AE017194</td>
<td>5224283</td>
<td>222523</td>
</tr>
<tr>
<td>14</td>
<td>B. cereus E33L</td>
<td>CP000001</td>
<td>5300915</td>
<td>288681</td>
</tr>
<tr>
<td>15</td>
<td>B. halodurans C-125</td>
<td>BA000004</td>
<td>4202352</td>
<td>272558</td>
</tr>
</tbody>
</table>

Table 2. Linkage Stage

<table>
<thead>
<tr>
<th>Stage</th>
<th>Clusters Combined</th>
<th>Dissimilarity</th>
<th>Stage Cluster First Appears</th>
<th>Next Stage</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Cluster 1</td>
<td>Cluster 2</td>
<td></td>
<td>Cluster 1</td>
</tr>
<tr>
<td>1</td>
<td>10</td>
<td>12</td>
<td>2.737</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>4</td>
<td>4.699</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>9</td>
<td>10</td>
<td>5.897</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>3</td>
<td>6.726</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>9</td>
<td>14</td>
<td>6.971</td>
<td>3</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>2</td>
<td>7.018</td>
<td>4</td>
</tr>
<tr>
<td>7</td>
<td>9</td>
<td>13</td>
<td>7.517</td>
<td>5</td>
</tr>
<tr>
<td>8</td>
<td>8</td>
<td>11</td>
<td>16.786</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>5</td>
<td>7</td>
<td>17.693</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>5</td>
<td>6</td>
<td>25.480</td>
<td>9</td>
</tr>
<tr>
<td>11</td>
<td>9</td>
<td>15</td>
<td>26.992</td>
<td>7</td>
</tr>
<tr>
<td>12</td>
<td>1</td>
<td>8</td>
<td>33.968</td>
<td>6</td>
</tr>
<tr>
<td>13</td>
<td>5</td>
<td>9</td>
<td>40.461</td>
<td>10</td>
</tr>
<tr>
<td>14</td>
<td>1</td>
<td>5</td>
<td>49.336</td>
<td>12</td>
</tr>
</tbody>
</table>
Fig. 1. Clustering results provided by HCA dendrogram

Fig. 2. The two-dimensional perceptual map constructed by MDS with normalized raw STRESS = 0.009