

PEMS Software

Phylogenetic **E**stimation of **M**etagenomic sequence using batch-learning

Self-Organizing Map (BLSOM)

Users' Guide

1. Outline

The software **PEMS** is intended for the phylogenetic estimation of user-input gene sequences in multi-FASTA format using **SOM** (Self Organizing Map), a method for data mining analysis.

The SOM-based clustering of a large amount of gene data requires high-capacity and high-speed computer systems. By using the results of SOM-based clustering by large-scale systems, such as the Earth Simulator, this software performs the phylogenetic estimation of user-input gene sequences on a PC.

2. Specifications

■ Recommended PC configuration

CPU: Intel Core 2 Duo 2.8 GHz or better

Memory: 2 GB or more

HDD: Free space of 4 GB or more (excluding swap space)

Video: Resolution of 1280 x 1024 pixels or more, 16-bit or higher color

■ OS

Microsoft Windows 7 or more

Microsoft .NET Framework 4.0 or better runtime environment

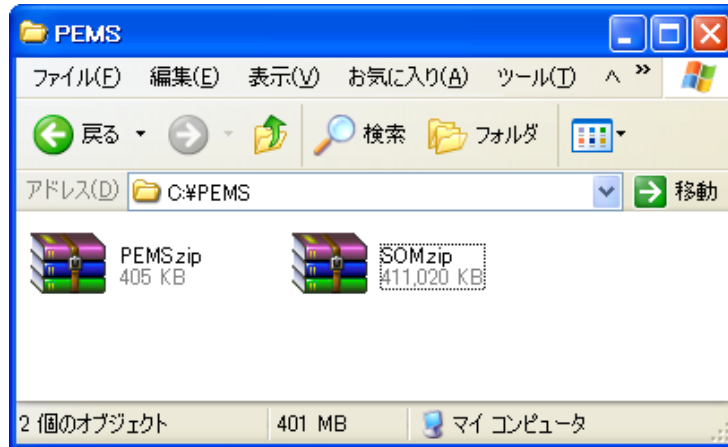
[Note]

This software prevents the response time of the user interface from slowing down by handling the high computational load through multithreading. With a single-core CPU, the software's screen redrawing and button response significantly slow down. For the software, a dual-core or better CPU is required.

3. Installation

■ STEP 1: Installing PEMS

The software comprises an analysis application and SOM dataset.

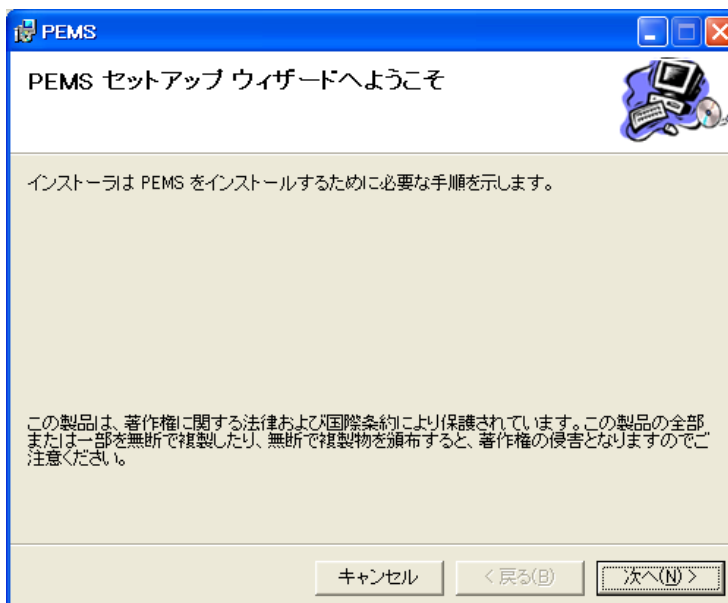


Decompress **PEMS.zip** (compressed file) using appropriate decompression software.

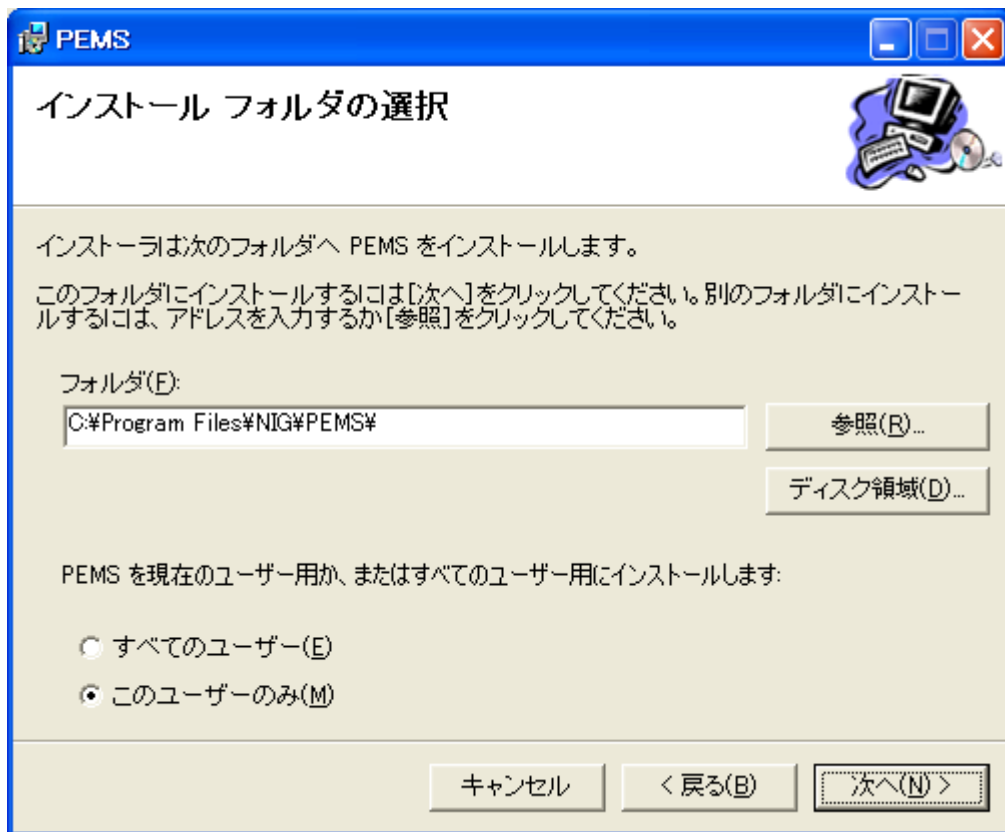


In the decompressed folder, double-click **setup.exe** to start the installation process.

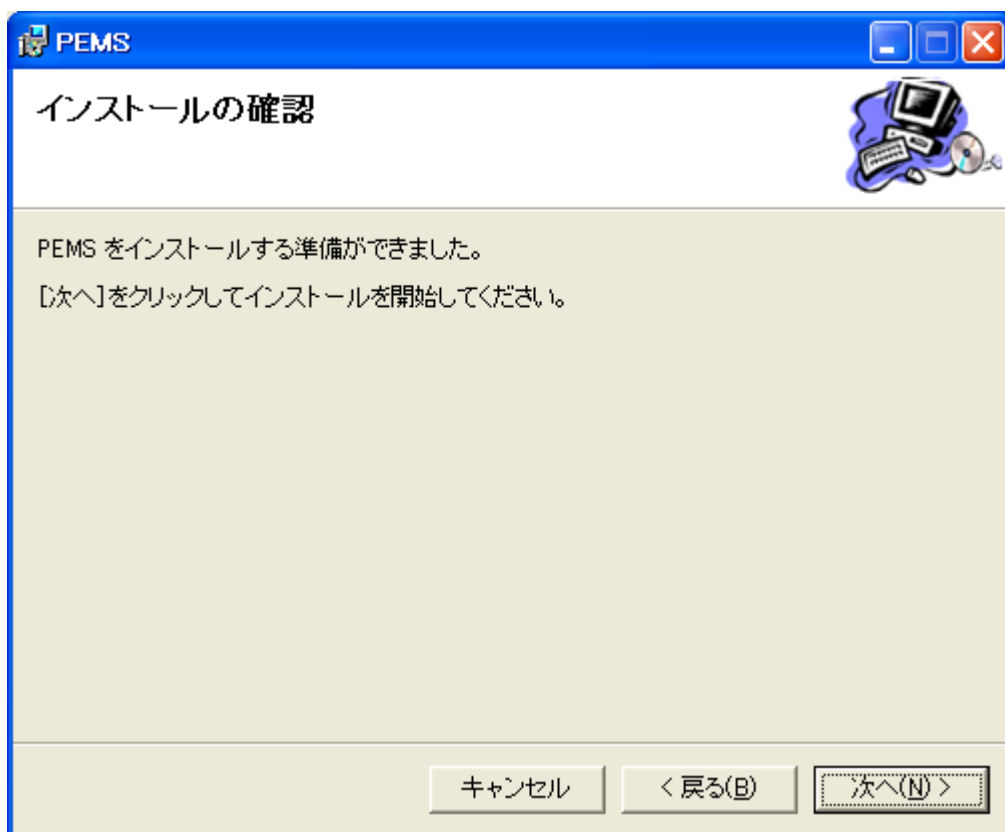
(Depending on the system configuration, the extension **.exe** may be hidden (only **setup** is visible).)



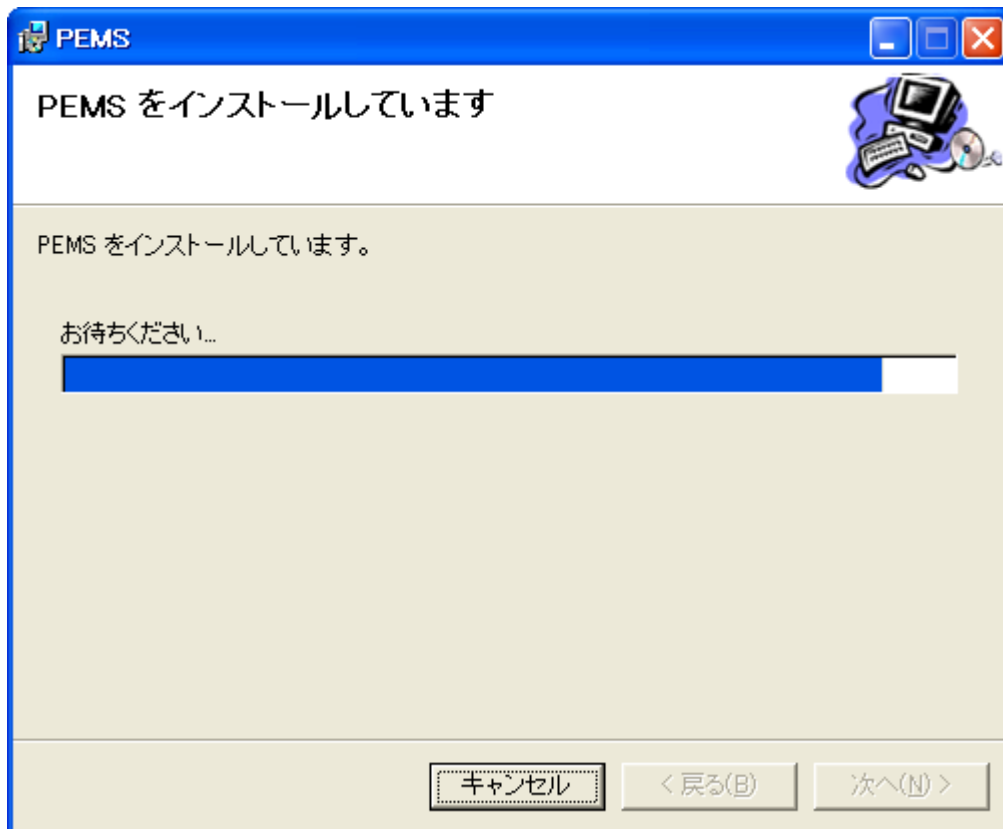
Click [次へ] ([Next]).



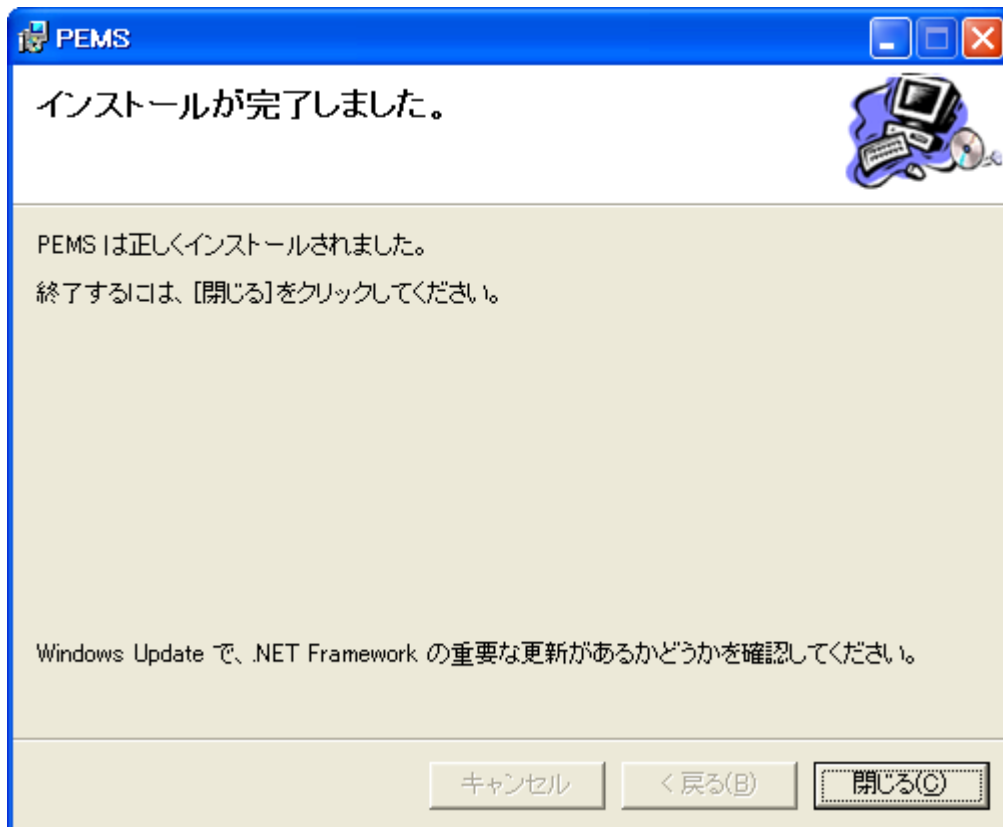
To change the installation folder, specify a desired folder and then click [次へ] ([Next]). The installation confirmation dialog appears.



Click [次へ] ([Next]) to start the installation.



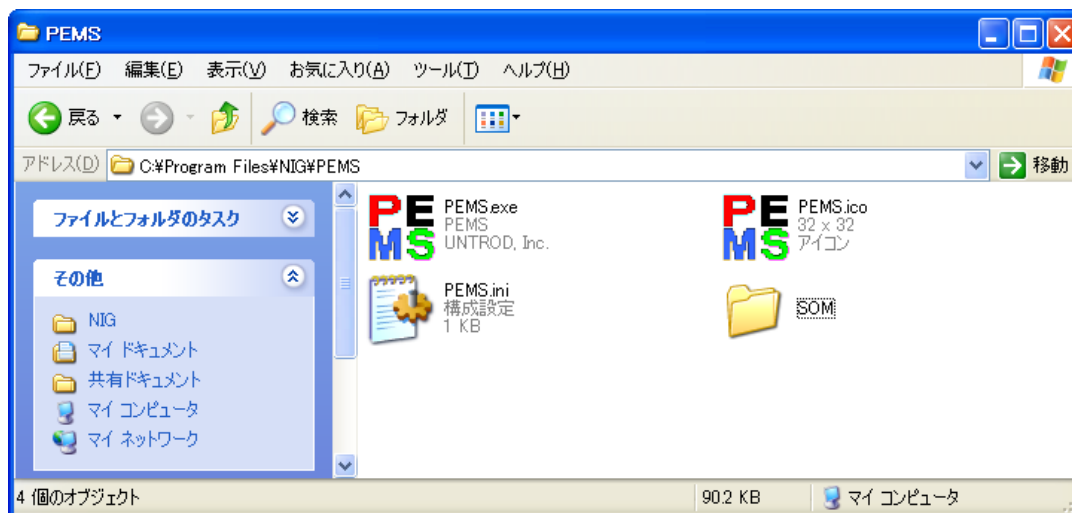
When the installation is complete, click [閉じる] ([Close]) to finish the installation.



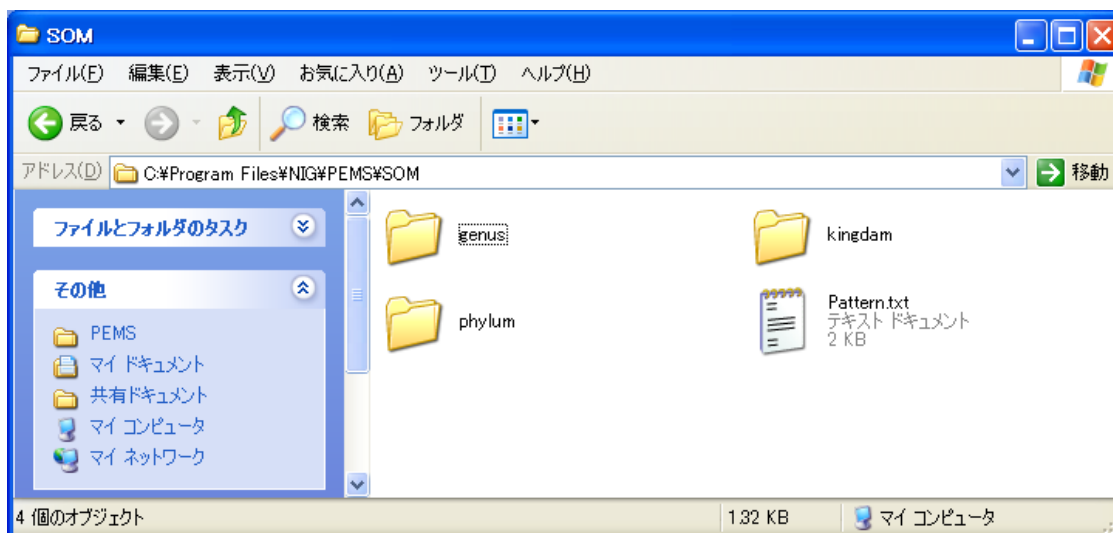
■ STEP 2: Installing SOM data

Decompress **SOM.zip** (compressed file) using appropriate decompression software.

Move the decompressed folder named **SOM** to the same directory as the directory where PEMS execution program **PEMS.exe** is installed in STEP 1.



Depending on the type of decompression software, a subfolder named SOM may be created in the SOM folder. Check that the SOM folder (subfolder) contains the following files and folders.



■ STEP 3: Uninstalling the software

To uninstall the software, go to [Control Panel] -> **[Add/Remove Programs]**. Then, select **PEMS** and click **[Remove]**.

Note that the SOM dataset is not deleted. If the dataset is unnecessary, manually delete the SOM folder containing it.

4. How to use the software

■ **STEP 1:** Starting the PEMS application

Double-click the **PEMS** shortcut icon on the desktop to launch the PEMS application.

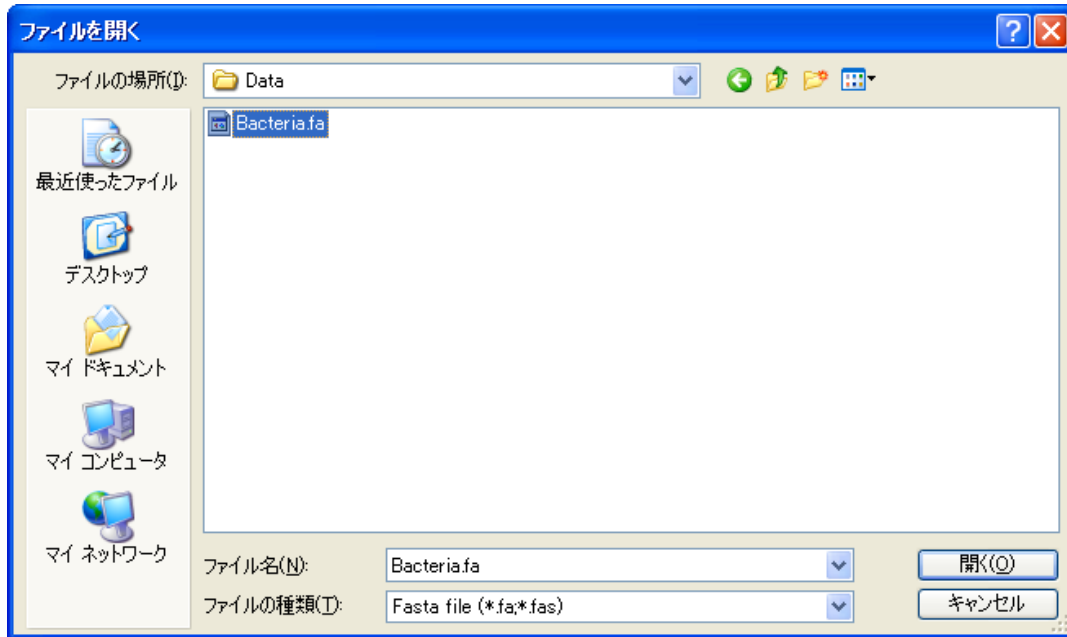


Alternatively, choose **[Start] -> [All Programs] -> [PEMS] -> [PEMS]** to launch the **PEMS** application.

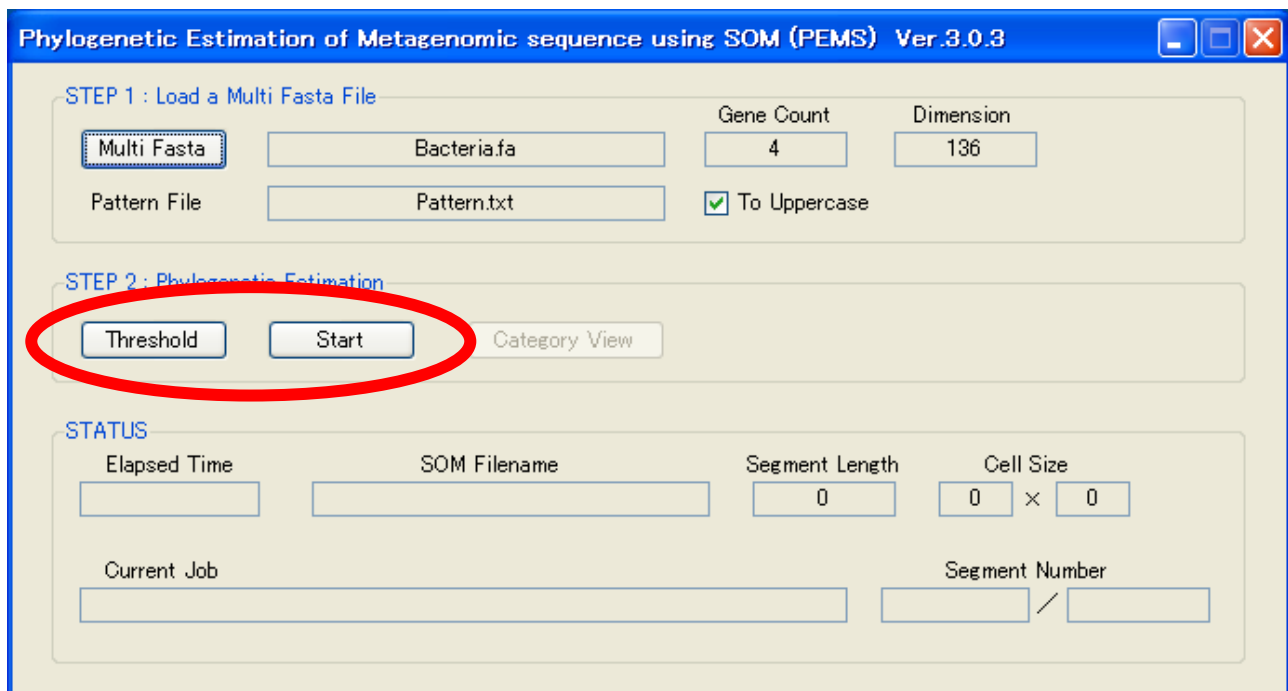
A screenshot of the PEMS application window. The title bar reads "Phylogenetic Estimation of Metagenomic sequence using SOM (PEMS) Ver.3.0.3". The interface is divided into three main sections: "STEP 1 : Load a Multi Fasta File", "STEP 2 : Phylogenetic Estimation", and "STATUS".
In the "STEP 1" section, there is a "Multi Fasta" button, a text input field, a "Gene Count" field with the value "0", and a "Dimension" field with the value "136". Below these are a "Pattern File" field with the value "Pattern.txt" and a checked checkbox labeled "To Uppercase".
The "STEP 2" section contains three buttons: "Threshold", "Start", and "Category View".
The "STATUS" section displays several fields: "Elapsed Time", "SOM Filename", "Segment Length" (value "0"), "Cell Size" (value "0" x "0"), "Current Job", and "Segment Number" (value " / ").

■ STEP 2: Specifying a multi-FASTA file

Click [**Multi Fasta**] to select an existing multi-FASTA file in the folder. Then, click [**開く**] (**[Open]**) (for the multi-FASTA file format, see **Supplemental information**).



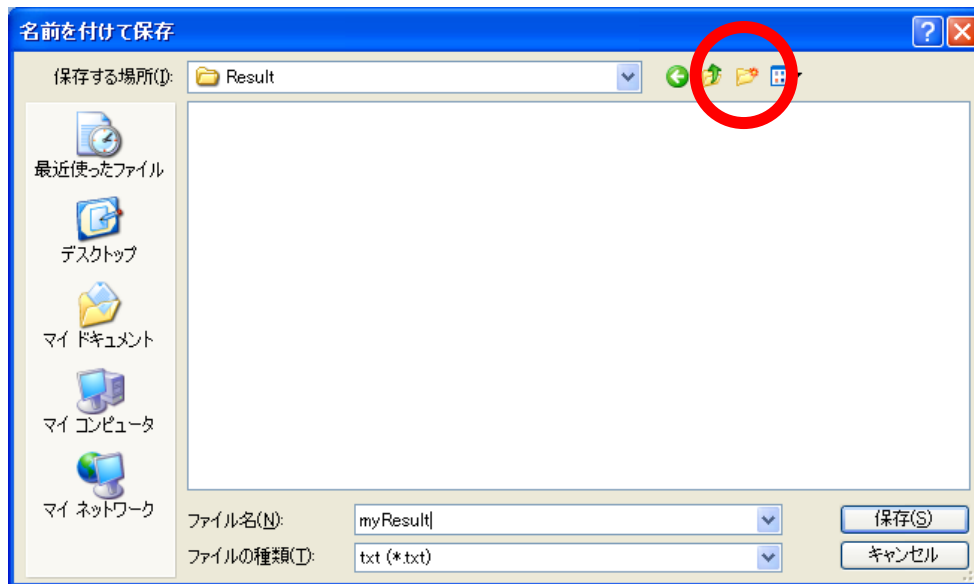
When the file load is successfully completed, the number of genes contained in the multi-FASTA file is indicated in the **Gene Count** box, as shown below. [**Threshold**] and [**Start**] are enabled.



■ STEP 3: Performing automatic estimation

Click **[Start]** and specify the folder and file name for storing analysis results. Then, click **[保存]** (**[Save]**).

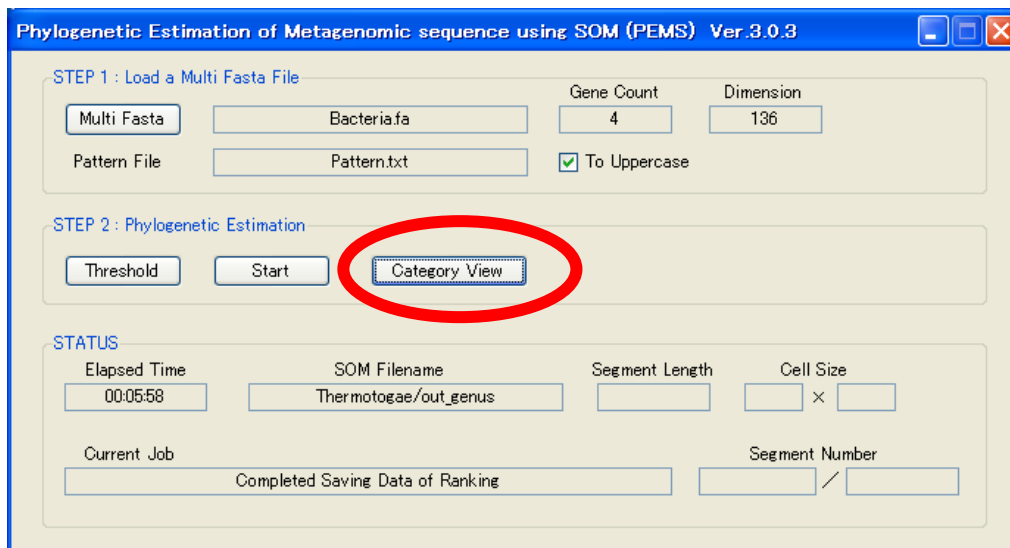
The extension .txt is automatically appended to the file name; the file is saved with the same name regardless of whether or not the file extension .txt is manually added.



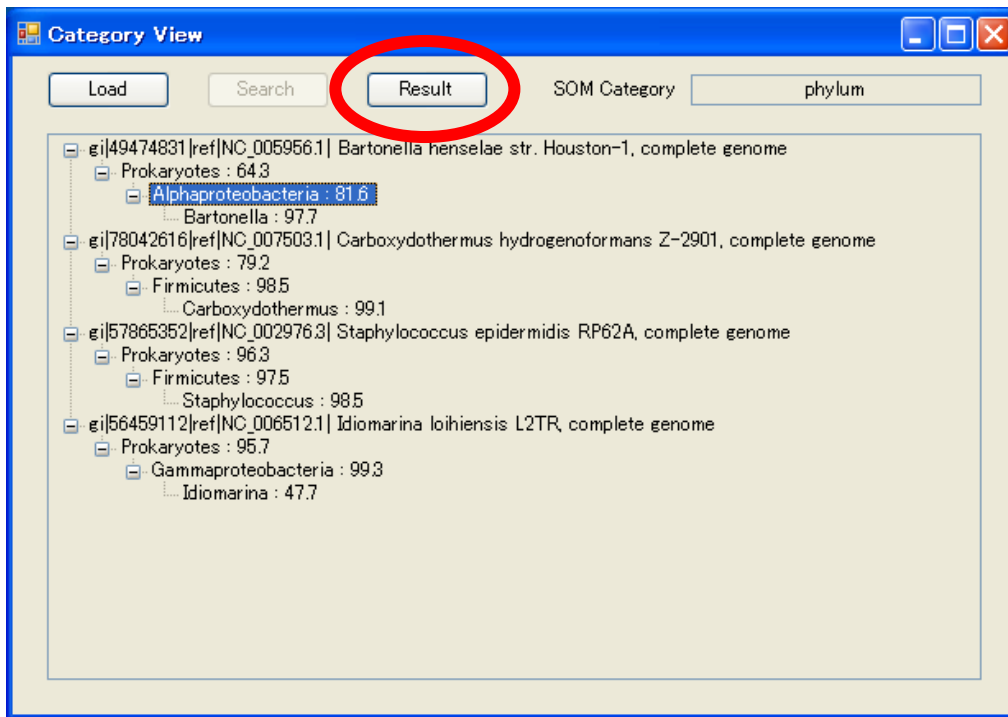
To create a new folder for storing analysis result data in this save dialog, click the folder icon circled in red above to create a folder with an appropriate name.

Before clicking **[Start]** to start computation for automatic estimation, you can click **[Threshold]** to change the threshold from its default (40) to any other value. For the meaning of this parameter, see the section “**Overview of analysis.**” Set a suitable value for the purpose of analysis.

Clicking **[Start]** starts computation for automatic estimation. When the entire computation is complete, related information is displayed in the **STATUS** area as shown below.



Click **[Category View]** to display the results of automatic estimation in the **Category View** window.



In this window, the name of each user-specified gene is at the top of the tree, and the results of phylogenetic estimation for the gene are displayed in the form of a top-down subtree.

Each subtree includes an estimated category name and the aggregate frequency of the category estimated by SOM (in percent).

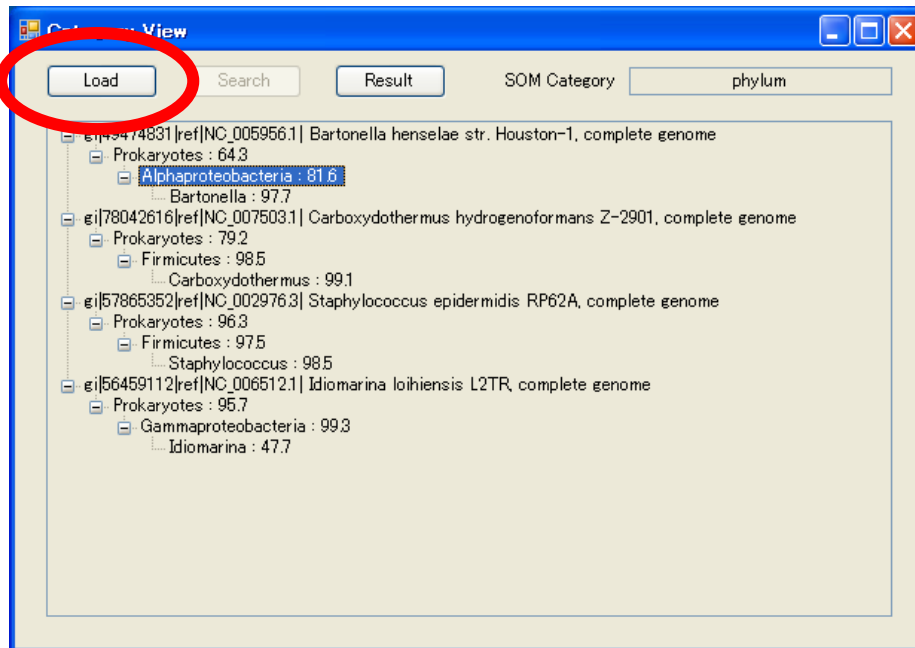
The results displayed in the tree are saved in a text file. Click **[Result]** to view the contents of the file on a text editor.

It is important to assess estimation results from various perspectives by referring to analysis result data files in the folder, as well as information displayed in the tree.

■ STEP 4: Performing estimation for individual segments

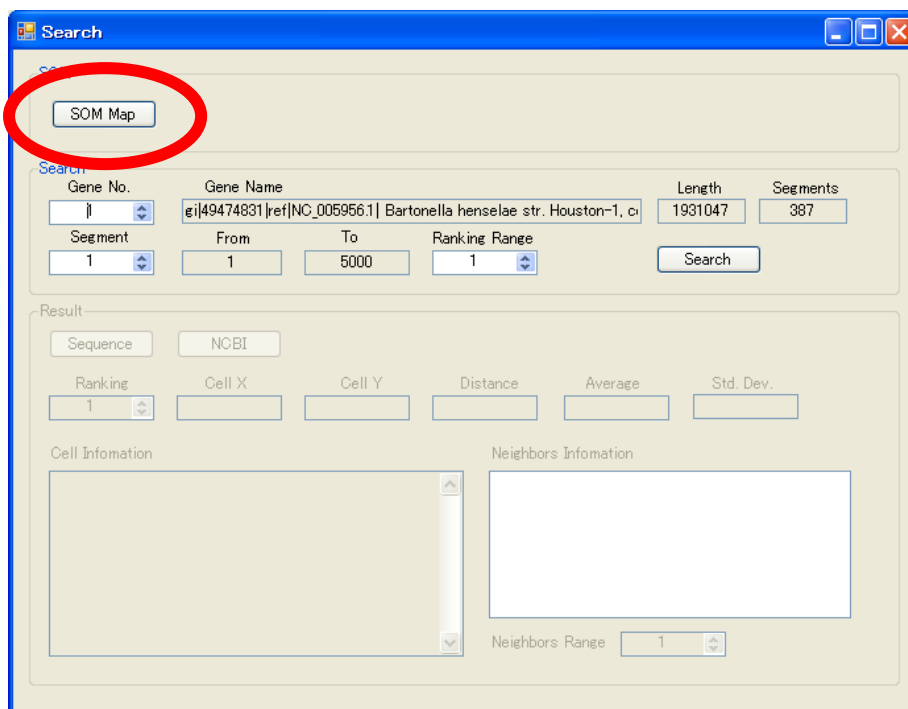
In the tree, click an entry to be analyzed in more detail. Clicking a branch displays the name of the SOM map used for the estimation in the upper right box.

At the kingdom or phylum level, the same SOM map applies to different gene branches for phylogenetic estimation. At the genus level, different SOM maps are often used for different genes.

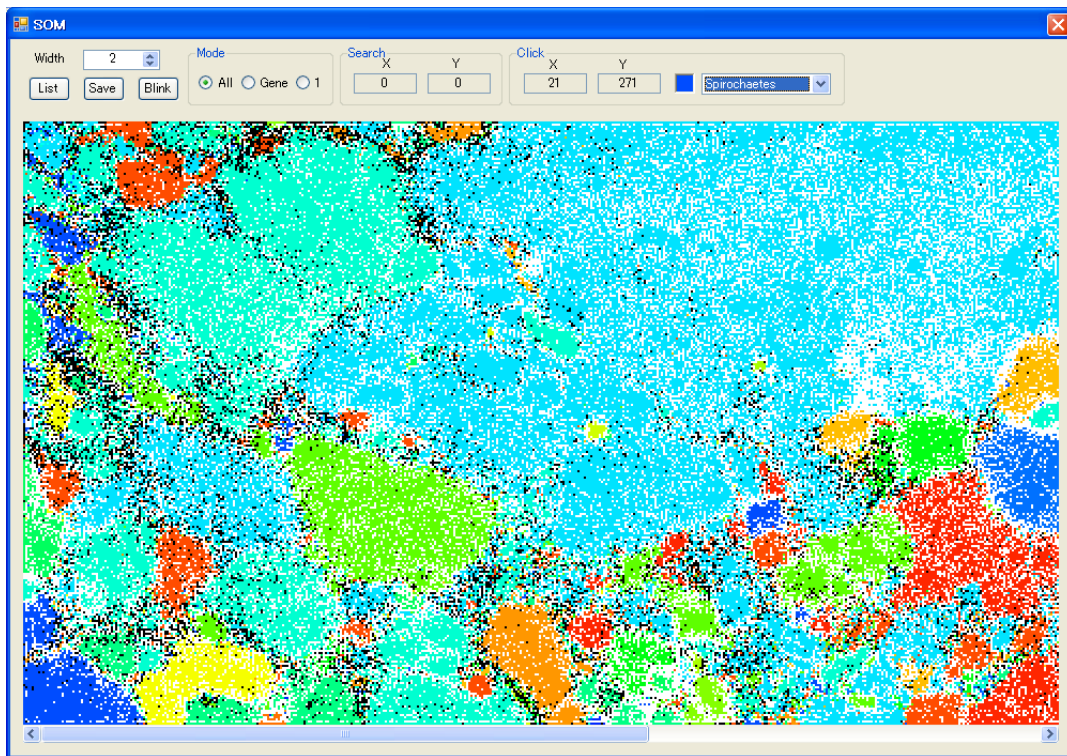


With a branch selected, click [**Load**] to load the dataset for the SOM map displayed in the upper right box and perform computation for phylogenetic classification. Upon computation completion, [**Search**] is enabled.

Clicking [**Search**] displays the window below.



Click [**SOM Map**] in this window to display the classification map based on SOM cells that was used for the phylogenetic estimation in a two-dimensional color grid.



The window can be resized by dragging its lower right corner.

If the SOM cell size is too large to display the entire map even by resizing the window, change the value in the **Width** box to decrease the number of pixels per cell.

When the number of pixels per cell is increased for detailed observation, use the scroll bars at the right and bottom of the map to move to a desired area.

Click [**List**] in the SOM window to display a list of category names and their corresponding colors.

Click [**Blink**] with **All** selected in the **Mode** area to display all segments contained in all user-input genes in reverse colors on the SOM map.

Click [**Blink**] with **Gene** selected in the **Mode** area to display all segments contained in a gene specified by the **Gene No.** box in the **Search** window in reverse colors on the SOM map.

Click [**Blink**] with **1** selected in the **Mode** area to display only a segment specified by the **Segment** box in the **Search** window that is contained in a gene specified by the **Gene No.** box in the same window. This segment is indicated with cross marks in the reverse color on the SOM map.



In all modes, relevant cells are displayed in reverse colors while pressing and holding the mouse button. Releasing the mouse button returns them to their original colors.

Click [**Save**] to access the color selection dialog for specifying the color of user-input genes. Select an appropriate color and click [OK].



When the file save dialog appears, enter an appropriate file name and then click [Save]. In this way, the SOM map image and the list image are saved at the same time.

By clicking on a cell on the SOM map, the address and color of the clicked SOM cell as well as the category name of the gene segment assigned to the SOM cell are displayed in the upper right area.

SOM cells with no gene segment assigned are displayed in white.

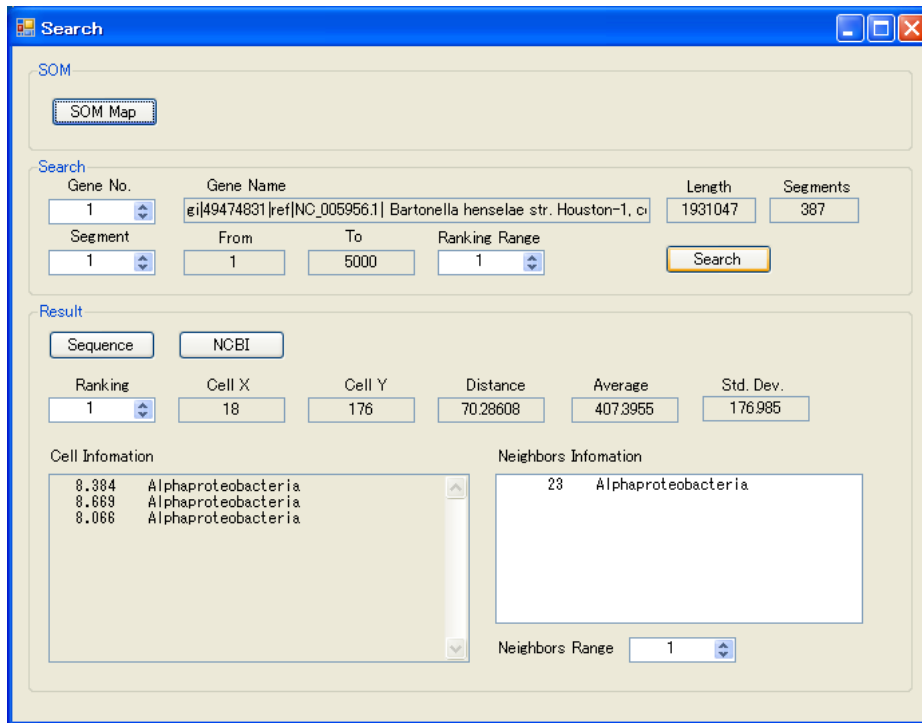
SOM cells with gene segments in different categories assigned are displayed in black.

SOM cells with gene segments in a single category assigned are displayed in the color corresponding to the category.

Set the number of a desired gene to the **Gene No.** box in the **Search** window. Set the number of a desired segment in the specified gene to the **Segment** box.

The distance between the pattern frequency vector of the specified gene segment and each weight vector in the SOM is computed to search for SOM cells with high similarity. The **Ranking Range** box specifies the number of similarity ranks from the top to be included in SOM cell search.

After the completion of parameter setting, click [**Search**] to perform the search process.

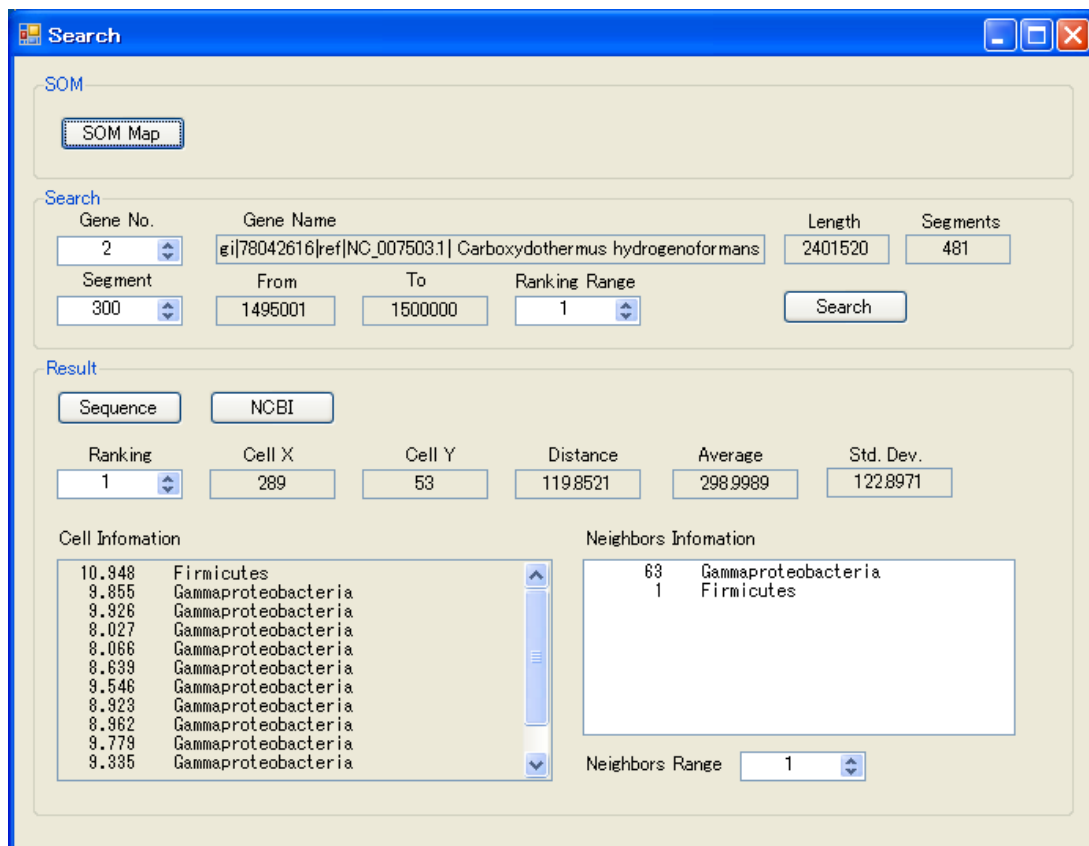


The **Result** area displays search results.

Changing the value in the **Ranking** box changes the displayed information to that of SOM cells at different similarity ranks.

The **Cell Information** field displays the results of clustering stored in the subject SOM cell.

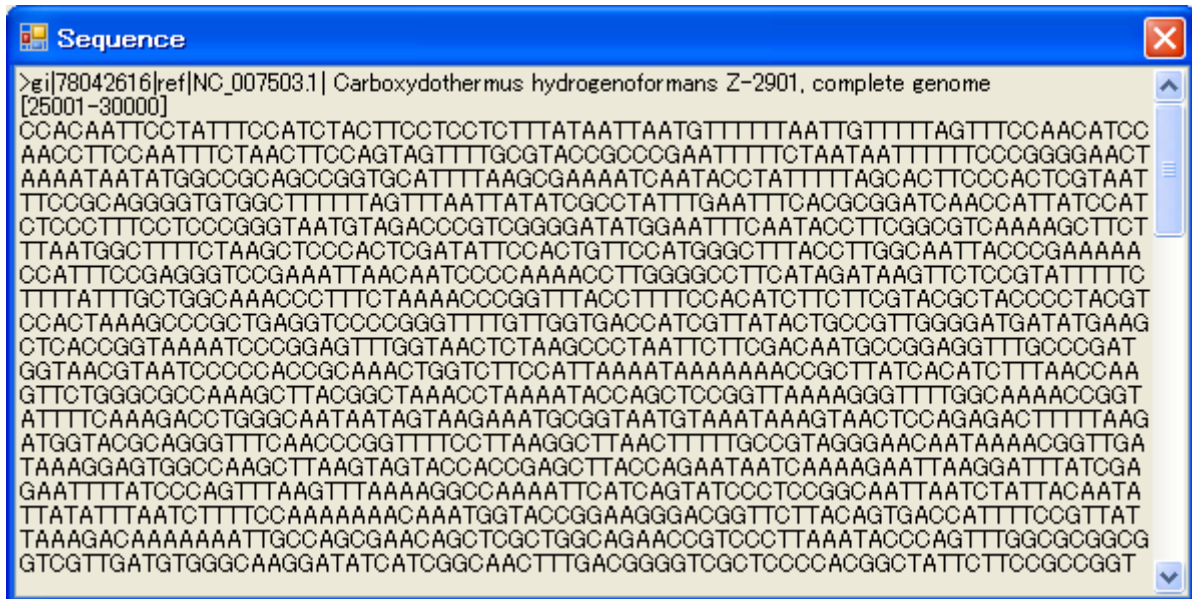
The **Neighbors Information** field displays the sum of frequencies of each category stored in the subject SOM cell and its neighboring SOM cells in ascending order.



Changing the value in the **Neighbors Range** box can change the range of neighboring SOM cells. If this value is 1, the sum of frequencies of each category stored in nine SOM cells in total (the subject SOM cell and its neighboring cells) is displayed.

During search operation, [**Sequence**], etc. in the **Result** area is disabled.

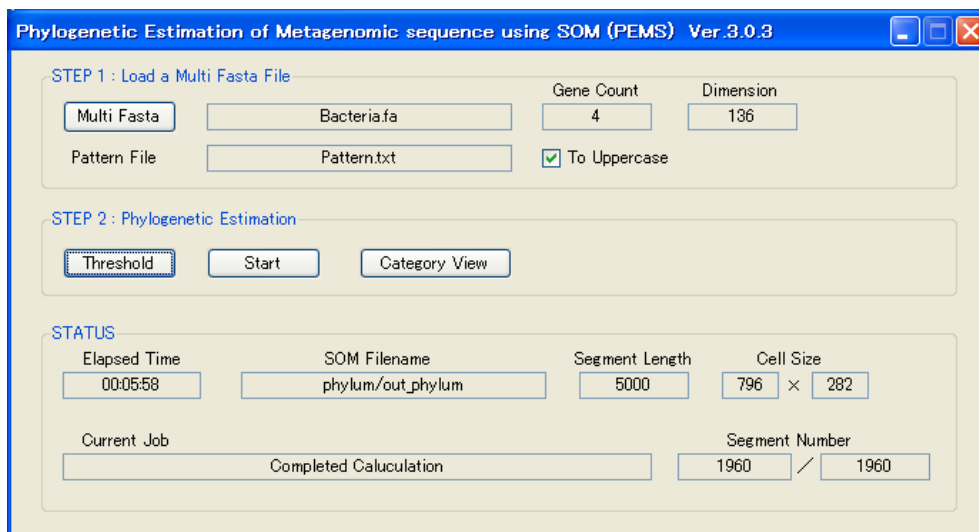
Click [**Sequence**] to display the sequence of the selected gene segment in a separate window. The first line indicates the gene annotation, the second line the segment start and end points, and the third and subsequent lines the sequence.



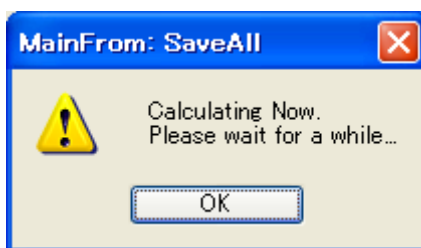
```
Sequence
>gi|78042616|ref|NC_007503.1| Carboxydotherrmus hydrogenoformans Z-2901, complete genome
[25001-30000]
CCACAATTCCTATTTCCATCTACTTCCTCCTCTTTATAATTAATGTTTTTAAATGTTTTAGTTTTCCAAACATCC
AACCTTCCAATTTCTAACTTCCAGTAGTTTTGCGTACCGCCCGAATTTTTCTAATAATTTTTCCCGGGAACT
AAAATAATATGGCCGACGCCGGTGCATTTAAGCGAAAATCAATACCTATTTTAGCACTTCCCACTCGTAAT
TTCCGCAGGGGTGTGGCTTTTTAGTTTAATTATATCGCCTATTTGAATTTACGCGGATCAACCATTATCCAT
CTCCCTTTTCTCCCGGGTAATGTAGACCCGTCGGGGATATGGAATTTCAATACCTTCGGCGTCAAAAAGCTTCT
TTAATGGCTTTTCTAAGCTCCCCTCGATATCCACTGTTCCATGGGCTTTACCTTGGCAATTACCCGAAAAA
CCATTTCCGAGGGTCCGAAATTAACAATCCCCAAAACCTTGGGGCCTTCATAGATAAGTTCTCCGTATTTTTC
TTTTATTTGCTGGCAAACCCTTTCTAAAACCCGGTTTACCTTTTCCACATCTTCTTCGTACGCTACCCCTACGT
CCACTAAAGCCCGCTGAGGTCCCCGGGTTTTGTTGGTGACCATCGTTATACTGCCGTTGGGGATGATATGAAG
CTCACCGGTAAAATCCCGGAGTTTGGTAACCTAAGCCCTAATCTTCGACAATGCCGGAGGTTTGGCCGAT
GGTAACGTAATCCCCACCGCAAACCTGGTCTTCCATTAATAAAAAAACCCTTATCACATCTTTAACCAA
GTTCTGGGCGCCAAAGCTTACGGCTAAACCTAAAATACCAGCTCCGGTTAAAAGGTTTTGGCAAACCCGGT
ATTTTCAAAGACCTGGGCAATAATAGTAAGAAATGCGGTAATGTAAATAAAGTAACTCCAGAGACTTTTTAAG
ATGGTACGCAGGGTTTCAACCCGGTTTTCTTAAGGCTTAACTTTTGCCGTAGGGAACAATAAACCGTTGA
TAAAGGAGTGGCCAAGCTTAAGTAGTACCACCGAGCTTACCAGAATAATCAAAAAGAAATTAAGGATTTATCGA
GAATTTTATCCAGTTTAAGTTTAAAAGGCCAAAATTCATCAGTATCCCTCCGGCAATTAATCTATTACAATA
TTATATTTAATCTTTTCCAAAAAACAATGGTACCGGAAGGGACGGTTCTTACAGTGACCATTTTCCGTTAT
TAAAGACAAAAAATTGCCAGCGAACAGCTCGCTGGCAGAACCGTCCCTTAAATACCCAGTTTGGCGCGGCG
GTCGTTGATGTGGCAAGGATATCATCGGCAACTTTGACGGGGTCGCTCCCCACGGCTATTTCCCGCCGGT
```


STEP 5: Exiting the application

To exit the PEMS application, click the Close button (red cross) at the top right corner of the main window.



When attempting to close the main window during computation, the dialog below appears. Click **[OK]** and wait for the computation process to end.



To change the background color of the window, open the **PEMS.ini** file in the installation folder and edit the RGB value on the sixth line in hexadecimal (e.g. FE907E).

5. Overview of analysis

The process of phylogenetic estimation by the software is outlined below.

1. The sequence of each user-specified gene is divided into segments with a given length, and the pattern frequency vector of each segment is computed.
2. A SOM cell with a weight vector that is the most similar to the pattern frequency vector of each segment is identified.
3. The frequencies of categories in this SOM cell and its neighboring SOM cells are computed.
4. The estimated category frequencies are sorted in descending order.
5. If SOM data is available under an estimated category, the process proceeds to the lower level, and the steps above are performed.

Nucleotide patterns used for computation in Step 1 above are described in the **Pattern.txt** file in the folder where the SOM dataset is located. This file is related to the result of SOM data analysis and must not be edited/revised.

The segmentation length for SOM clustering is automatically specified. If the entire length of a user-specified gene for analysis is shorter than the specified segment length but is larger than the default threshold value, the analysis is carried out through frequency vector normalization. This threshold value can be changed by editing the value on the fifth line in the **PEMS.ini** file in the installation folder.

Genes longer than the specified segment length are divided into segments with a given length for frequency analysis. Odd segments at the end of divided gene sequences are also analyzed through frequency vector normalization if their length is larger than the threshold value.

When any nucleotide pattern other than specified patterns is present in a gene segment, it is excluded from counting; only the frequency vectors of specified patterns are normalized.

In Steps 2 to 4 above, phylogenetic estimation is performed for each segment in each gene, and the results are saved into files corresponding to SOM data used for the estimation.

For example, when a save file is named **myResult.txt**, the result of phylogenetic estimation based on the SOM analysis file for out_kingdom at the kingdom level is saved into the **myResult_kingdom.txt** file; the result of phylogenetic estimation based on the SOM analysis file for Actinobacteria at the genus level is saved into the **myResult_Actinobacteria.txt** file.

An example of contents written to the files is given below.

```
#gj|78042616|ref|NC_007503.1| Carboxydotherrmus hydrogenoformans Z-2901, complete genome
[1155001-1160000]    2 16    Prevotella          21    Flavobacteriales    6    Pedobacter    4
[1520001-1525000]    22 39    Flavobacteriales    21
```

The line starting with **#** indicates the gene annotation, and the next and subsequent lines indicate the result of computation for each segment.

In the example above, **[1155001-1160000]** at the beginning shows the segment start and end points. The next set of figures, **2 16**, shows the X and Y coordinates of the SOM cell with a weight vector that is the most similar to the pattern frequency vector of the segment (the shortest inter-vector distance). The frequencies of categories in this SOM cell and its neighboring eight SOM cells are summed and sorted in descending order. The results are indicated next to the SOM cell coordinates.

The result of simply summing up the category frequencies of segments in a gene is saved with a file name with `_Category Name_All` inserted, like **myResult_Actinobacteria_All.txt**.

Multiple categories may or may not take first place in the frequency ranking at the same time. In this document, a single top category is called **true top category**.

The result of summing up only the frequencies of true top categories of segments in a gene is saved with a file name with `_Category Name_All1st` inserted, like **myResult_Actinobacteria_All1st.txt**.

The result of extracting the category with the highest frequency among the summed frequencies of true top categories at the same level in each gene is saved with a file name **myResult_Top.txt**. The same information as this file is displayed in a tree view by clicking **[Display]**.

The counting result of the number that have been estimated for each category in each Kingdom/Phylum/Genus is saved with a file name **myResult_Hist.txt**.

In automatic computation with **[Save All]**, the rate of the frequency of a top category in each segment is computed.

For example, suppose that there is the following segment described in `myResult_phylum.txt`.

```
#gj|49474831|ref|NC_005956.1| Bartonella henselae str. Houston-1, complete genome  
[165001-170000] 82 196 Alphaproteobacteria 25 Bacteroidetes 8 Chlorobi 2
```

The rate of the frequency of Alphaproteobacteria, top category, is:

$$(25/(25 + 8 + 2)) * 100 = 71.4 \%$$

If the rate of the frequency is higher than the value specified in the **Threshold** box, this segment is used for phylogenetic estimation for Alphaproteobacteria at the next lower level.

Segments in the same gene that have different top categories are classified into different categories for phylogenetic estimation at the lower level.

Unlike the case above, if multiple categories take first place in the frequency ranking upon classification of segments for phylogenetic estimation at the lower level, each of such segments is classified into multiple categories for phylogenetic estimation.

6. Supplemental information

■ Multi-FASTA file format

The annotation of each gene is prefixed with a > character.

The gene annotation line starts with a > character and is written in one row. Except for these rules, the file is free-format.

A gene annotation line is followed by a line describing the sequence of the gene.

If the sequence extends over multiple rows, there is no restriction on the number of characters per row. However, computation efficiency can be improved by breaking the line at appropriate points, instead of writing the sequence in an extremely long line. The sequence length is not restricted, but its upper limit depends on the memory size of the PC to be used.

If the specifications above are met, there is no restriction on the number of genes contained in a file. However, an excessively large file size consumes much processing time and memory. Therefore, it is recommended to divide genes into multiple files appropriately and to adjust the number of genes contained in one file.

The file extension should be .fa or .fas.

[Example]

>gj|49474831|ref|NC_005956.1| Bartonella henselae str. Houston-1, complete genome

```
GTGAAGAGAGAAAAAAATCCTTTCATTTACACATGCTTTCTGATGCAACGGGAGAAACATTAATATCTG
TTGGAAGAGCTGTAGCATCACAATACACGATGAGTCAGGCAACAGAACATATCTATCCGATGATTCGTAA
CAAAACACAGTTGCAGAGAGCTCTTGATGAGATACAACAAGAGCCTGGGATAGTTCTTTACACAATCATT
GACAAAAAAATTAATTTCTTCTTAGAAAAAGATGTGAAAAAATAGAAATTCCTGTATT
```

>gj|78042616|ref|NC_007503.1| Carboxydotherrmus hydrogenoformans Z-2901, complete genome

```
AACCTGAAAAAAGTGTGAAAAAATTTTGTGGATTTGTGGATAAAACAAGGTTTTTGCTAATTTTTGCTA
ATAAAAAAATTTATAAAGAGATTCGTGAAAGCAAAGATTGTGGATAACGAAAAACTCAAGAAAAATTTTC
TTGACGGGTCTTATCCCATCTTCTATAATTTAGGTGTATCTATAGGGGATATAGGCTTTTTTAGATAGAT
TAGGAGGTGTGAAAATGAAAAGAACCTACCAACCAAAAAACCGGCGGCG
```

■ Collaboration with the **Dendrogram** analysis software

When the **Dendrogram** analysis software (AHCD) has been installed in the default directory C:\Program Files\NIG\Dendrogram\AHCD.exe, **[Dendrogram]** appears at the start of **PEMS**, as shown below.

Click **[Dendrogram]** to launch the **Dendrogram** application. The name of the file analyzed with **PEMS** is passed to the **Dendrogram** and added to the file dialog.

