PEMS ソフトウェア

Phylogenetic Estimation of Metagenomic sequence using SOM

ユーザーズガイド

第 6.6.0 版

1. 概要

本ソフトウェア PEMS の目的は、データマイニングの解析手法の一つである SOM (Self Organizing Map)を活用して、ユーザが入力したマルチファスタ形式の遺伝子シーケンスの系統推定を行うことである。

大量の遺伝子データを SOM でクラスタリングするためには、大容量で高速な計算処理能力を有する計算機システムが必要となる。そこで本ソフトウェアでは、SOM によるクラスタリング処理については 、地球シミュレータなどの大規模なシステムで計算された結果を活用し、その処理結果を用いて PC 上でユーザが入力した遺伝子シーケンスの系統推定を行う。

2. 仕様

■ PC の推奨仕様

CPU:	Intel Core 2 Duo 2.8GHz 以上
Memory:	2GB 以上
HDD:	4GB 以上の残容量（スワップ領域は除く）
Video:	解像度 1280 x 1024 ピクセル以上、色 16 ビット以上
Network:	NCBI データベースにアクセス可能なインターネット環境

■ OS

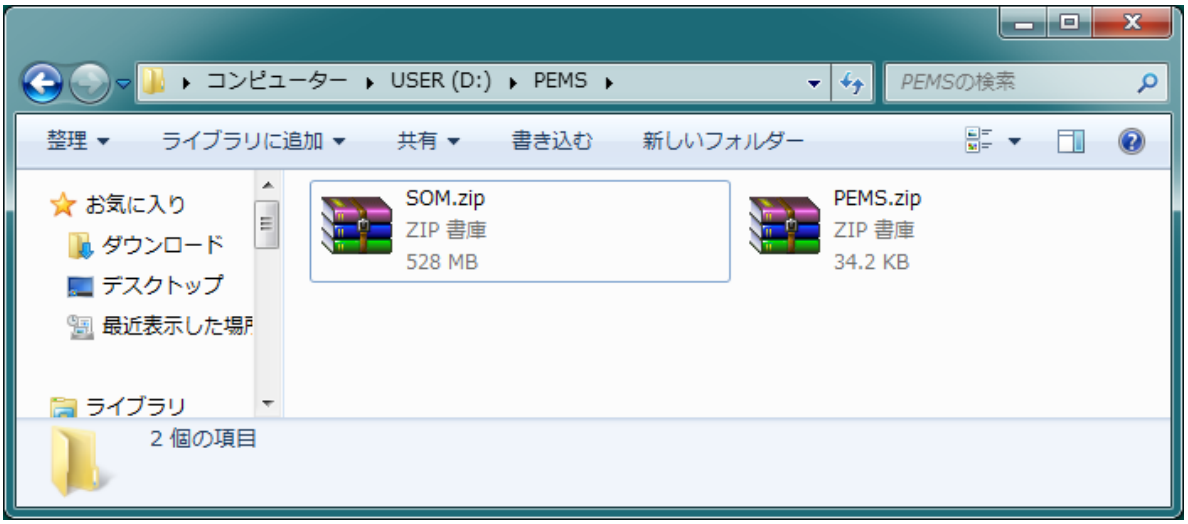
Microsoft Windows 7 (64 ビット版を推奨)
Microsoft .NET Framework 4.0 以上のランタイム

【注意】
本ソフトウェアは、負荷の高い計算をスレッドで処理することによって、ユーザインターフェースのレスポンスの低下を防いでいる。シングルコアの CPU の場合、本ソフトウェアの画面の再描画やボタンなどの反応が著しく低下する。このため、CPU に関してはデュアルコア以上の CPU を必要とする。

3. インストール

■ STEP 1 ： PEMS のインストール

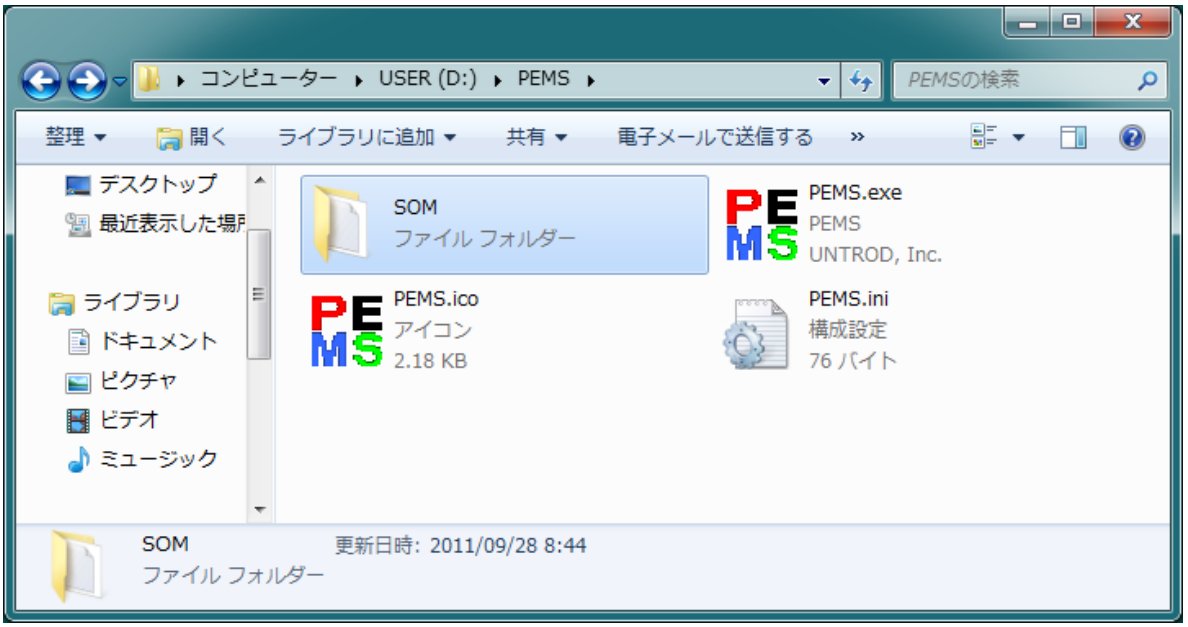
本ソフトウェアは、解析アプリケーションと SOM のデータセットから構成されている。



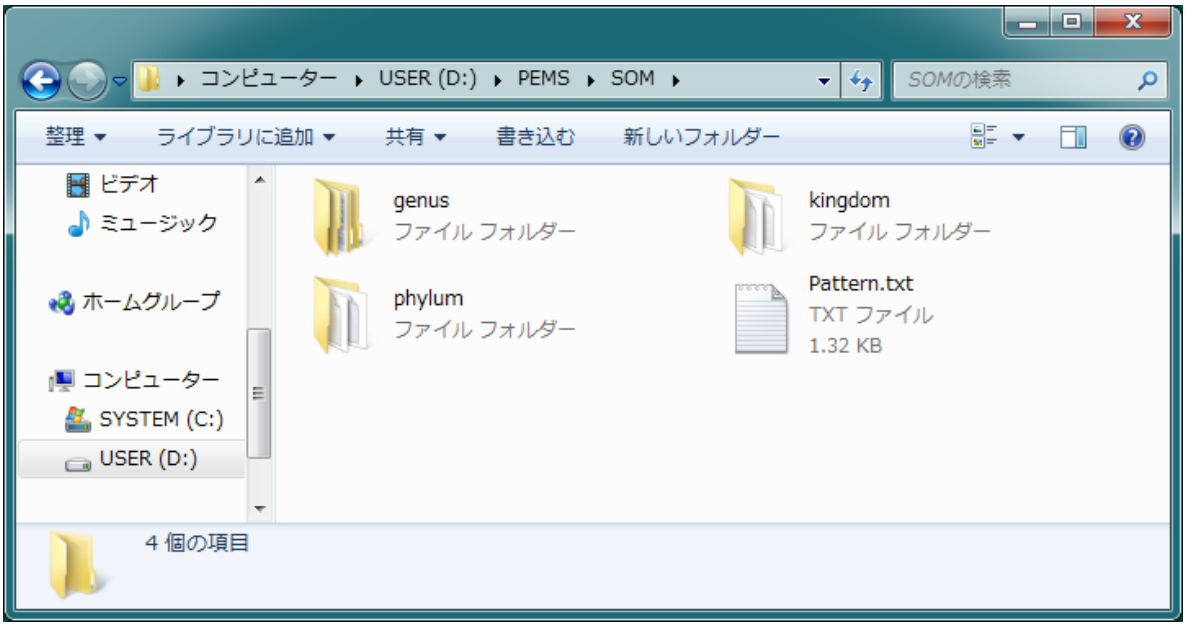
PEMS.zip の圧縮ファイルを適当な解凍ソフトウェアを利用して解凍する。

■ STEP 2 ： SOM データのインストール

SOM.zip の圧縮ファイルを適当な解凍ソフトウェアを利用して解凍する。
解凍したフォルダ SOM を、STEP1 で解凍した PEMS ソフトウェアの実行プログラム PEMS.exe と同じディレクトリに移動する。



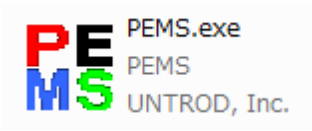
解凍ソフトウェアによっては、SOM フォルダの中にさらに SOM フォルダを生成する場合があるので、SOM フォルダの内部が以下のようなファイル構成になっていることを確認する。

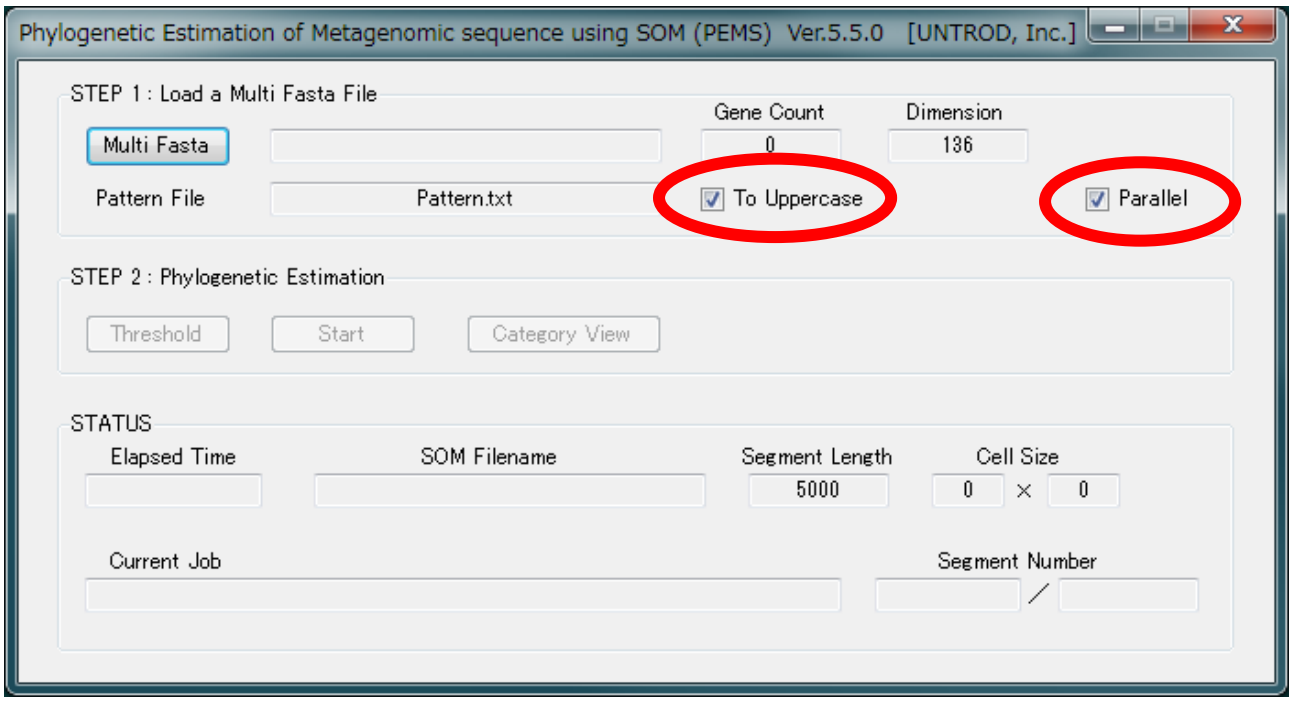


4. 使用方法

■ STEP 1：PEMS アプリケーションの起動

デスクトップの PEMS.exe と書かれたアイコンをダブルクリックして PEMS アプリケーションを起動する。



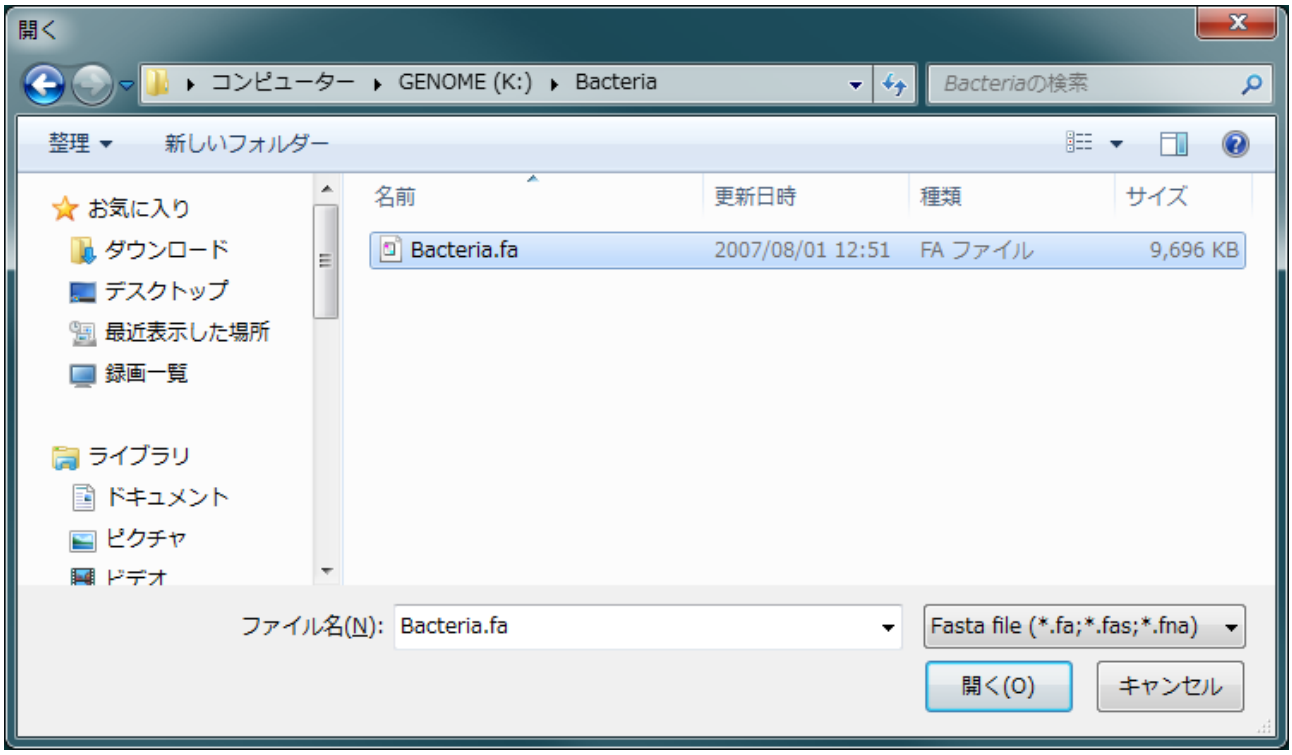


マルチコアを搭載したＰＣ上で計算させる場合は、【Parallel】 チェックボックスをオンに設定すると、計算の処理速度が向上する。

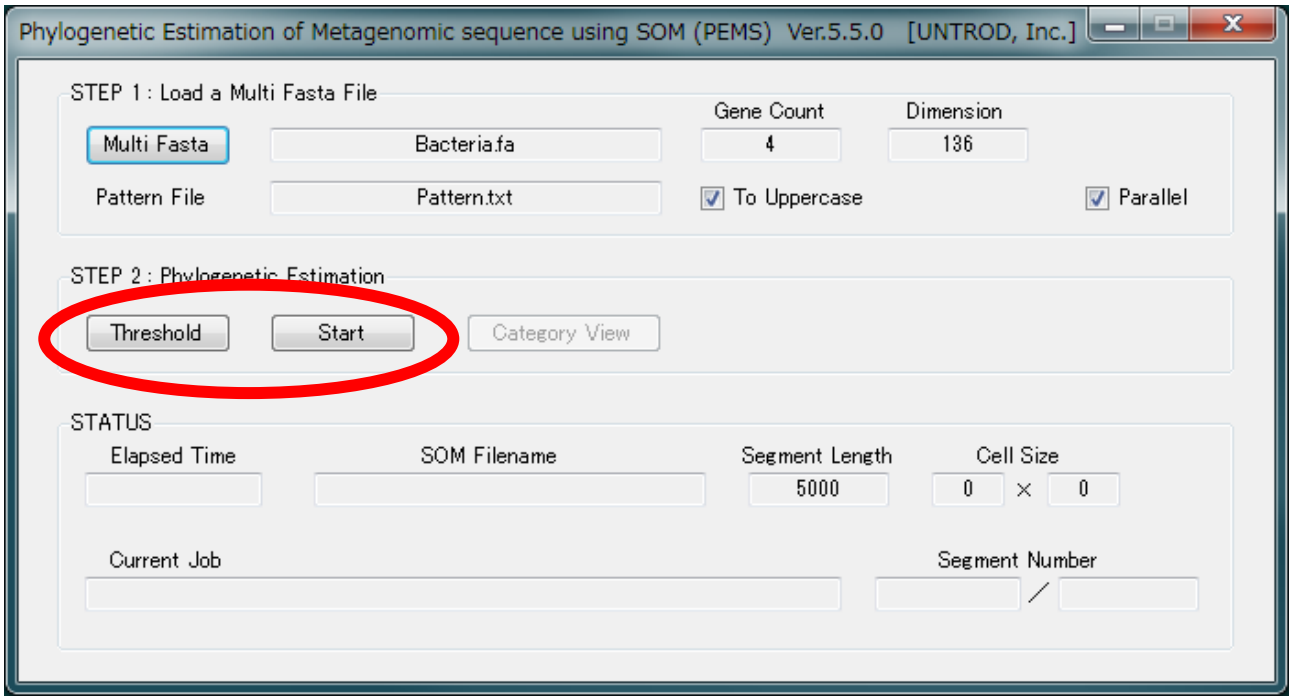
Fasta ファイル内に記述された遺伝子の各文字を、すべて大文字に自動変換して計算したい場合には、【To Uppercase】 のチェックボックスをオンに設定する。

■ STEP 2：マルチファスタファイルを指定する

【Multi Fasta】ボタンを押して、あらかじめフォルダに保存しておいたマルチファスタファイルを指定して**【開く】**ボタンを押す。（マルチファスタのファイル形式については、**補足**を参照）



ファイルの読込みが正常に完了すると、次図のようにマルチファスタのファイル内に含まれる遺伝子の数が【Gene Count】欄に表示され、【Threshold】や【Start】のボタンが使えるようになる。



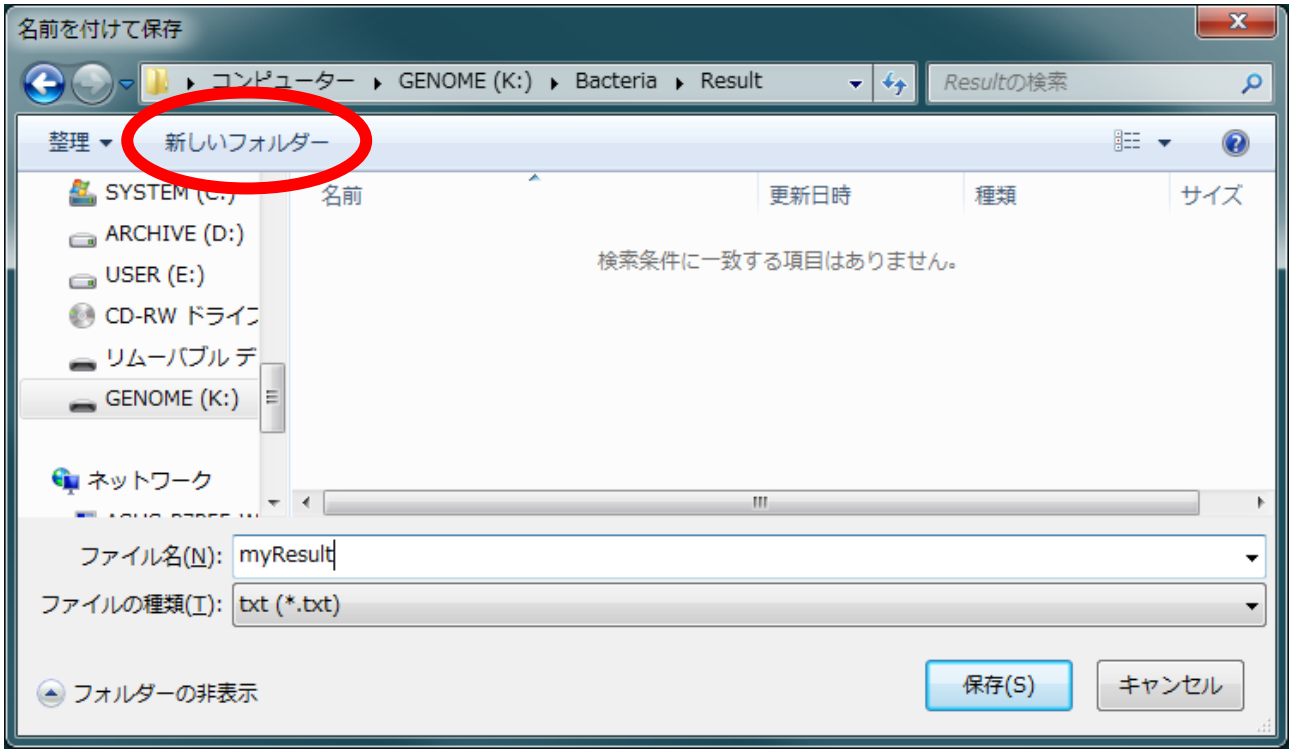
■ STEP 3：自動推定の実行

【Start】ボタンを押して、解析結果を保存するフォルダとファイル名を指定して**【保存】**ボタンを押す。

ファイル名には、.txt の拡張子が自動的に付加されるので、ファイルの拡張子に .txt を付けても付けなくても、同じファイル名として保存される。

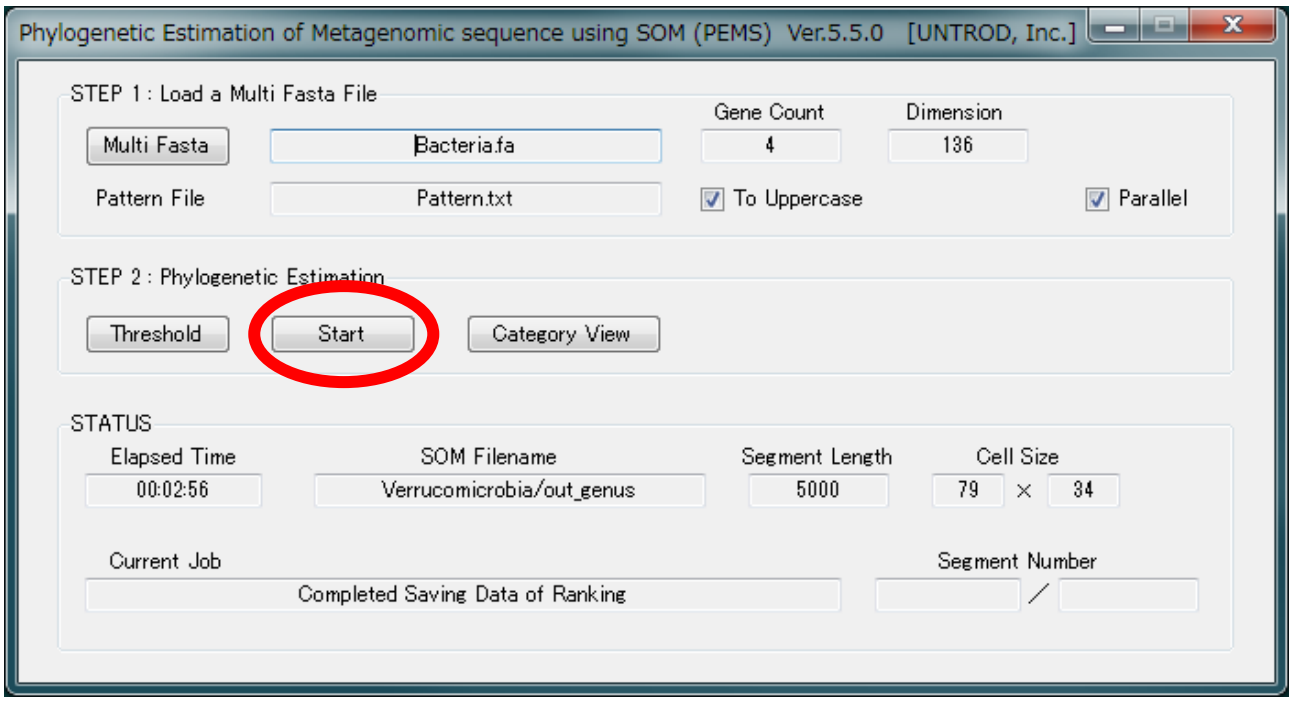
この保存ダイアログ内で、解析結果のデータを格納するためのフォルダを新規に作成したい場合には、下図の赤丸で示したフォルダのアイコンをクリックし、適当な名前のフォルダを作成する。

フォルダを作成した場合は、そのフォルダ内に移動してから、保存するファイルのファイル名を指定する。

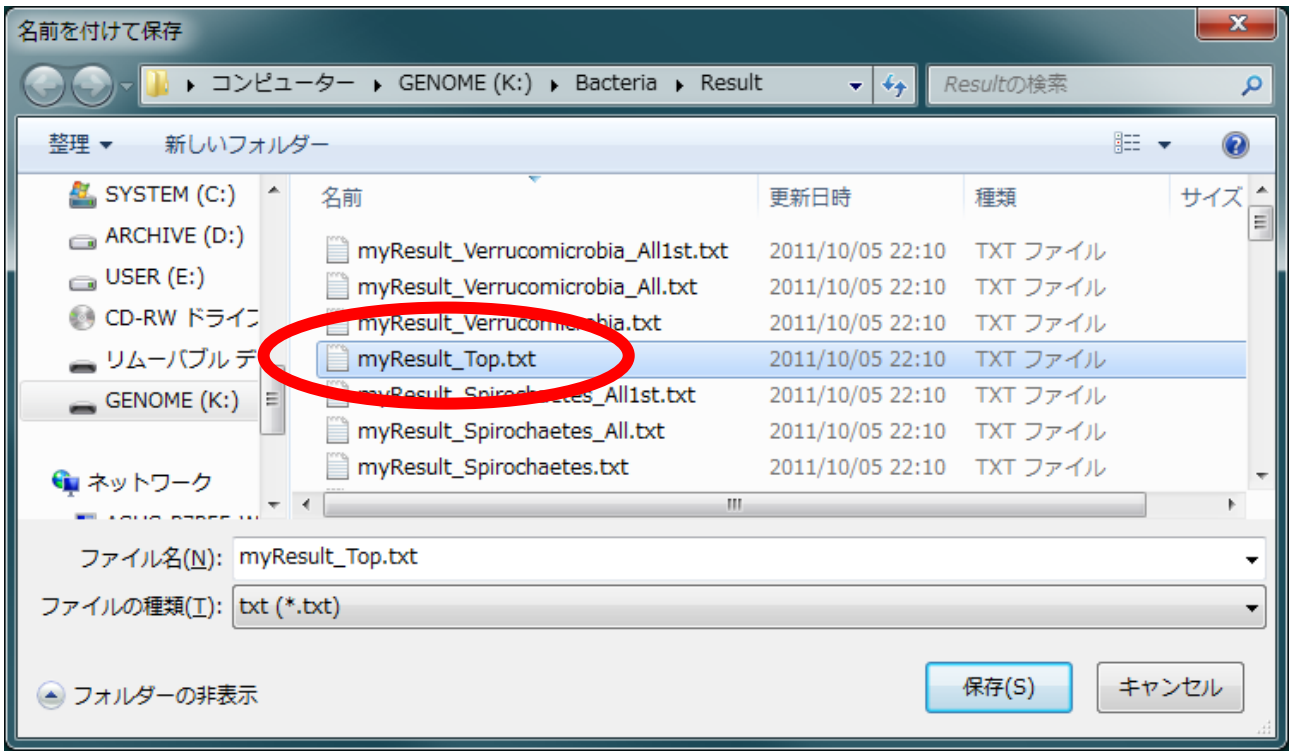


【Start】ボタンを押して推計計算を実行する前に、【Threshold】のボタンを押して、閾値の数値ボックスの値をデフォルトの 40 から異なった値に変更することも可能である。この値の意味については、**解析手法の概要**の項目を参照し、解析目的に適した値に設定する。

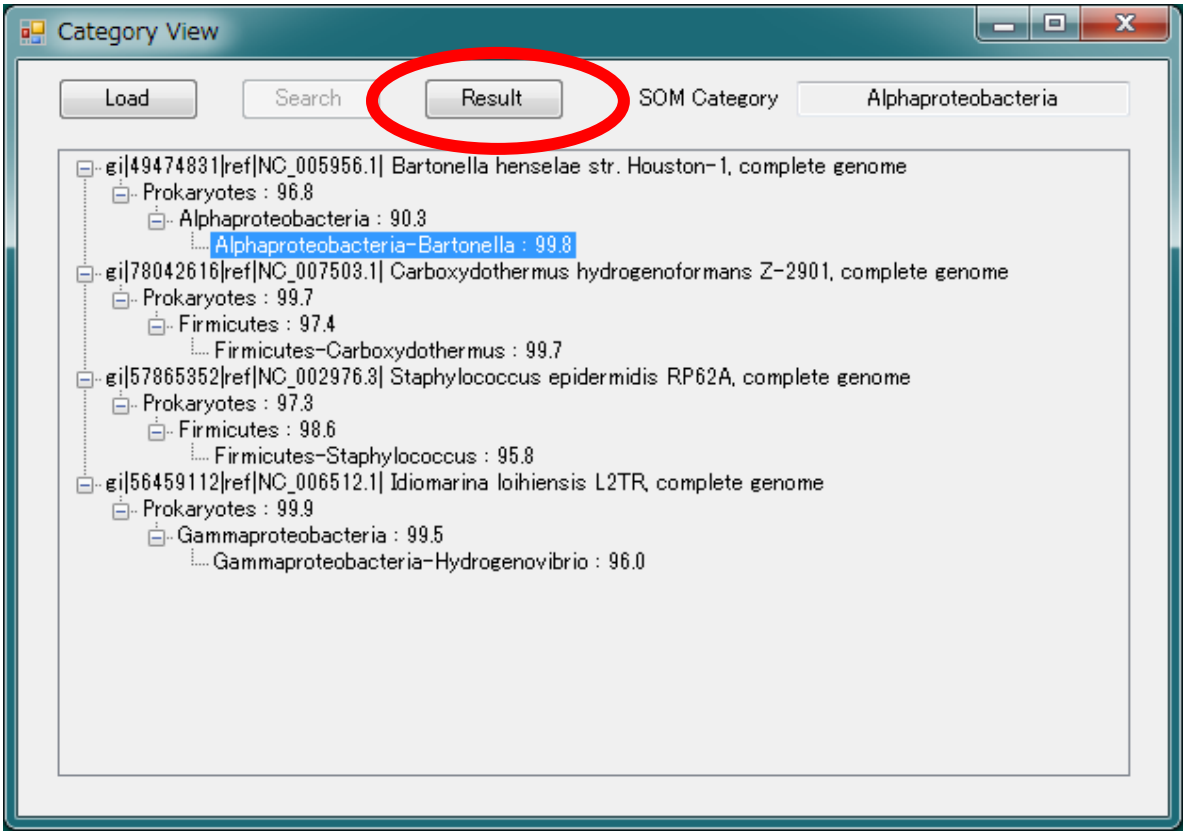
【Start】ボタンを押すと自動推定の計算が始まり、計算が全て終了すると次図のように、一番下の【STATUS】欄に計算終了のメッセージが表示される。



すでに計算済みのデータがあり、そのデータを読み込んでデータを設定したい場合は、【Start】ボタンを押し、下図のように、*****_Top.txt のファイルを指定してから**【保存】**ボタンを押す。



【Category View】ボタンを押すと自動推定された結果が【Category View】ウィンドウに表示される。



このウィンドウには、ユーザが指定した各遺伝子名がツリーの最上位に表示され、その遺伝子の系統推定の結果が上位階層から順に、サブツリーとして表示される。

サブツリーには、推定されたカテゴリ名と SOM が推定したカテゴリの総和頻度の割合がパーセントで表示される。

ツリーに表示された結果は、テキストファイルとして保存されており、【Result】ボタンを押すとファイルの内容がテキストエディタ上に表示される。表示される結果は、ツリーと同じ構造データなので、ツリー上で選択した項目には依存せず、同じファイルの内容が表示される。

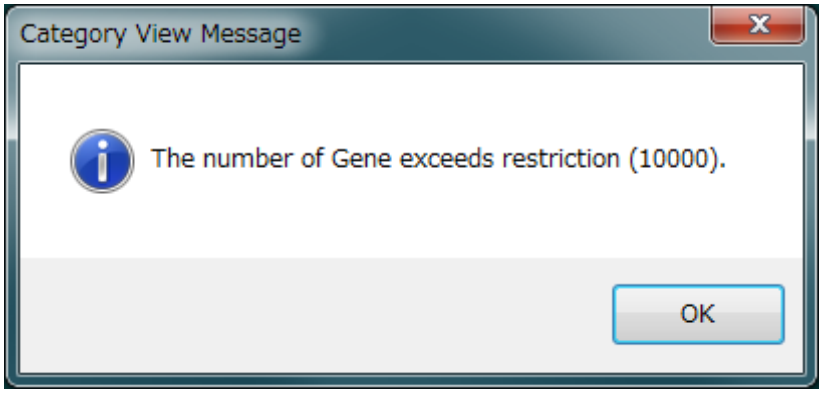
あらかじめ、ウィンドウズ上で *.txt の拡張子を持つファイルを、適当なテキストエディタと対応付けておけば、【Result】ボタンを押すと、****_Top.txt ファイルが対応付けられたエディタで開く。

対応付けがなされていない場合は、ウィンドウズ標準の**メモ帳**が使われる。

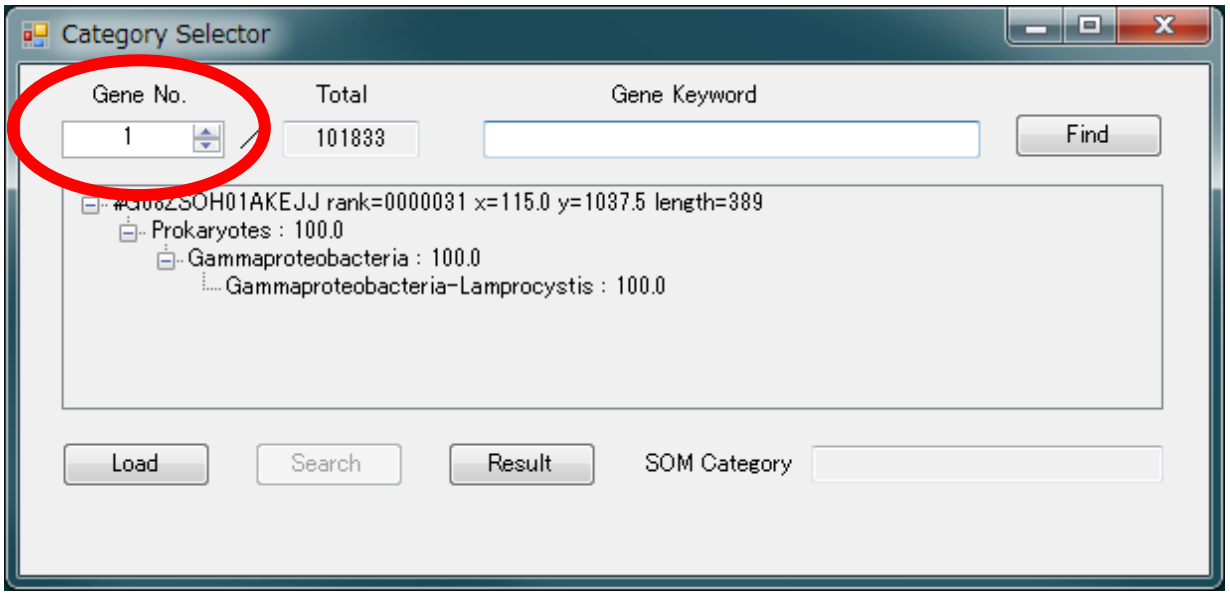
巨大なファイルを開く場合には、フリーソフトの Sakura Editor を推奨する。

推定結果については、ツリーに表示された結果だけでなく、解析結果のデータが保存されたフォルダ内の各ファイルを参照するなどして、多角的に判断することが重要である。

解析の対象となる遺伝子セグメントの総数が、あらかじめ設定された閾値を超えると、以下のようなメッセージが表示される。



【OK】ボタンを押すと、最初の遺伝子に関する階層ツリーのみが表示される。

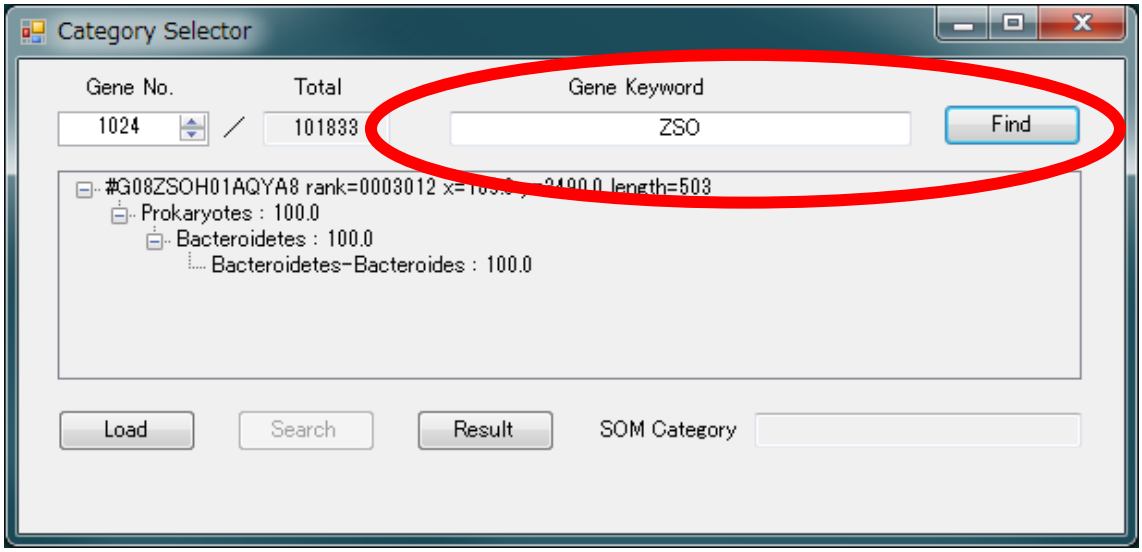


Gene No. の数値スピンに遺伝子セグメントの番号を入力するか、数値スピンの右側の三角形の小さなボタンを押すことで、遺伝子セグメントの番号を変更することができる。

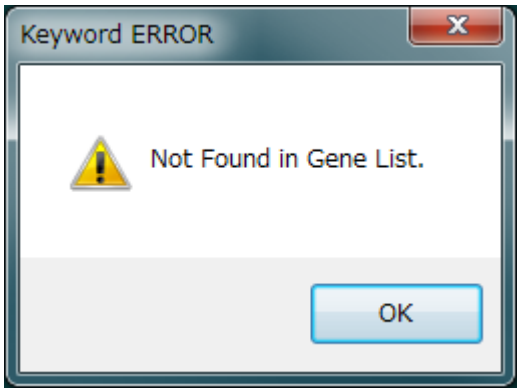
また、Gene Keyword のテキストボックスに適当な文字列を入力してから【Find】ボタンを押すと、入力した文字列を遺伝子名の中に含む遺伝子セグメントが検索され、該当する遺伝子セグメントの階層ツリーが表示される。

再度【Find】ボタンを押すと、次の検索結果が表示される。【Find】ボタンが選択された状態で、Enter キーを押し続けることで、高速な連続検索が可能となる。

上記で述べた二種類の階層ツリーの切り替えを決める閾値を変更したい場合には、PEMS.ini の６行目の数値を書き換える。



入力した文字列が、どの遺伝子名にも含まれない場合は、以下のようなメッセージが表示される。



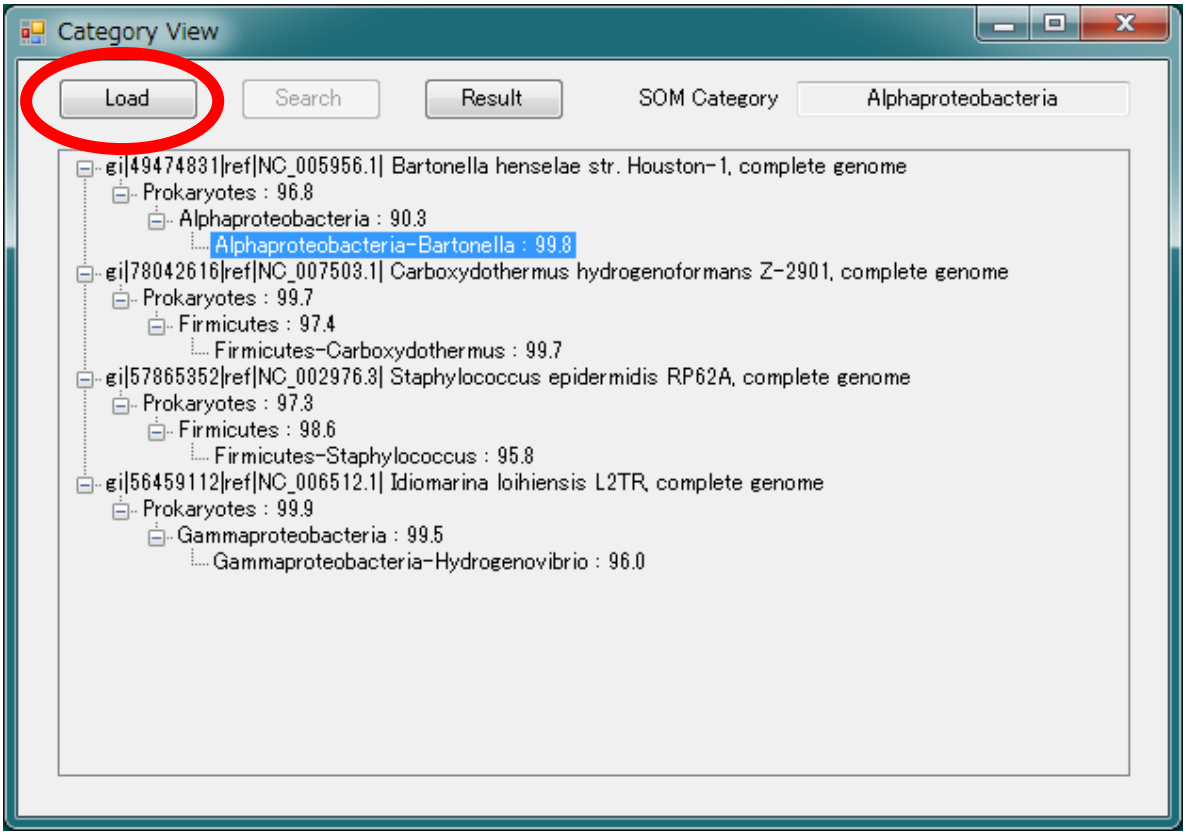
【Result】ボタンを押して表示された ****_Top.txt ファイルから、適当な文字列をコピーして、上記の検索用の Gene Keyword のテキストボックスにペーストして検索すると、効率のよい探索が可能となる。

■ STEP 4：個別セグメントの推定の実行

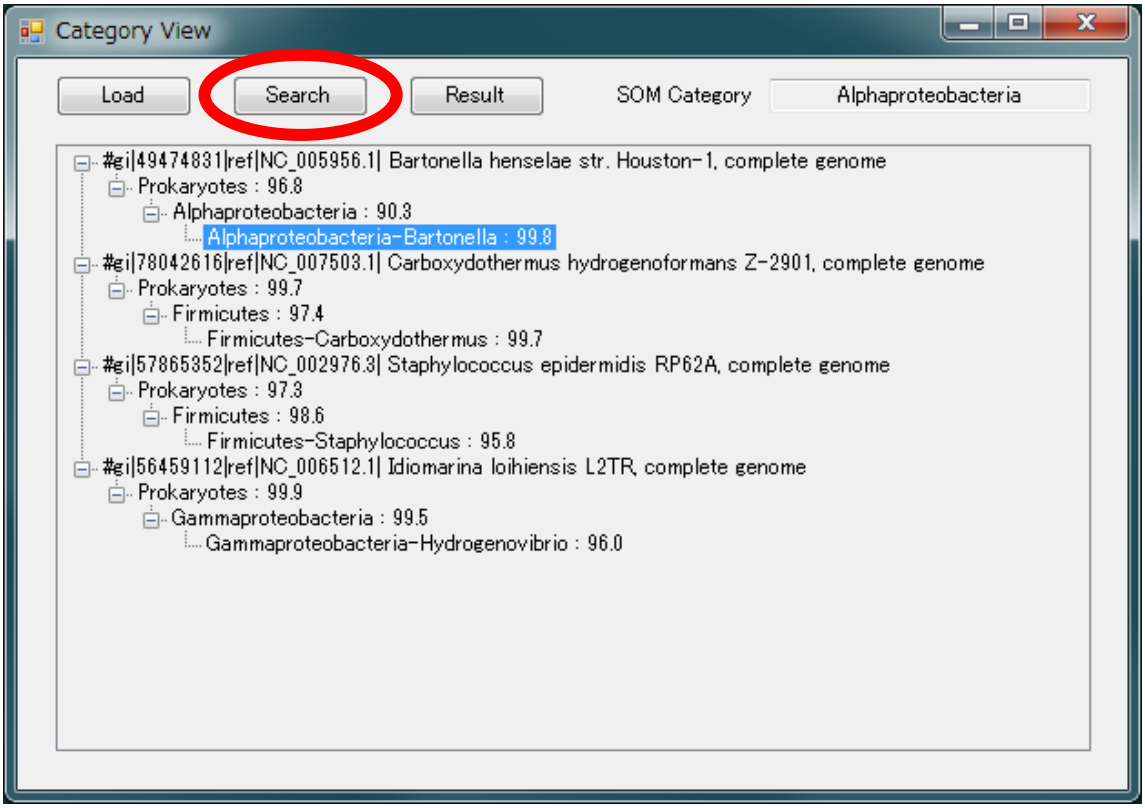
ツリーの中からさらに詳しく分析を行ってみたい項目をクリックする。枝をクリックすると、右上のテキストボックスに、推定に使用した SOM マップの名前が表示される。

異なった遺伝子の枝ではあっても、大分類（kingdom）や中分類（phylum）の枝の場合は、同じ SOM マップによって系統推定が行われている。小分類（genus）の枝の場合は、遺伝子ごとに異なった SOM マップが使われていることが多い。

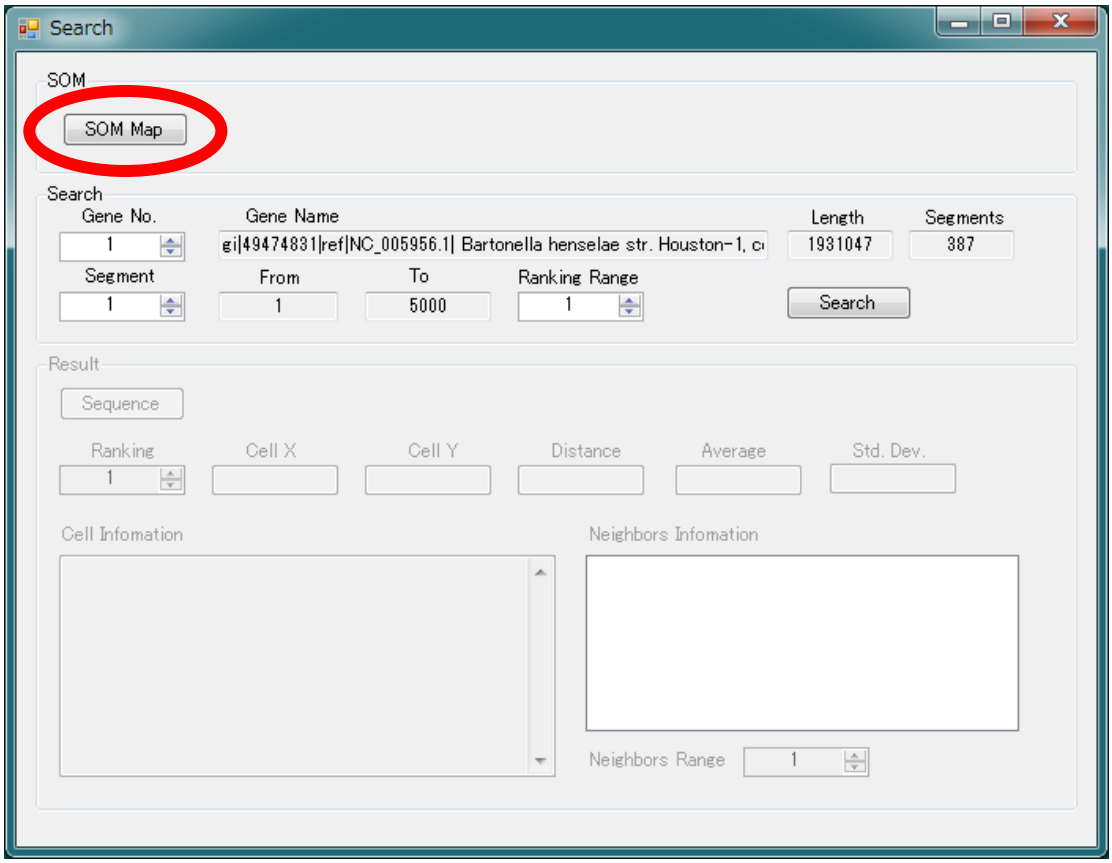
枝を選択したら、【Load】ボタンを押して、右上のテキストボックスに表示された SOM のデータセットをロードして、分類計算を実行する。



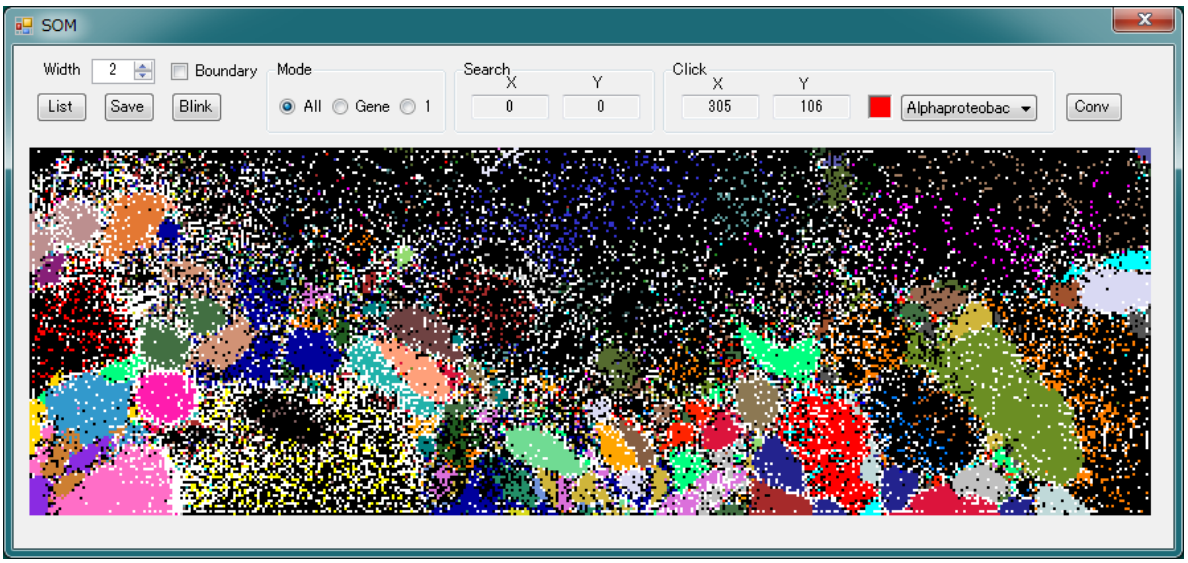
計算が完了すると、【Search】ボタンが利用可能となる。



【Search】ボタンを押すと、下記に示したウィンドウが表示される。



Search ダイアログの【SOM Map】ボタンを押すと、系統推定に使用した SOM セルの分類マップが 2 次元のカラー格子として表示される。



ウィンドウの右下をドラッグすることで、ウィンドウの大きさを調整できる。
SOM のセルサイズが大きすぎて、ウィンドウを広げても全体のマップが表示できない場合には、【Width】の数値ボックスの値を変更して、セル 1 個当たりの表示ピクセル数を少なくする。

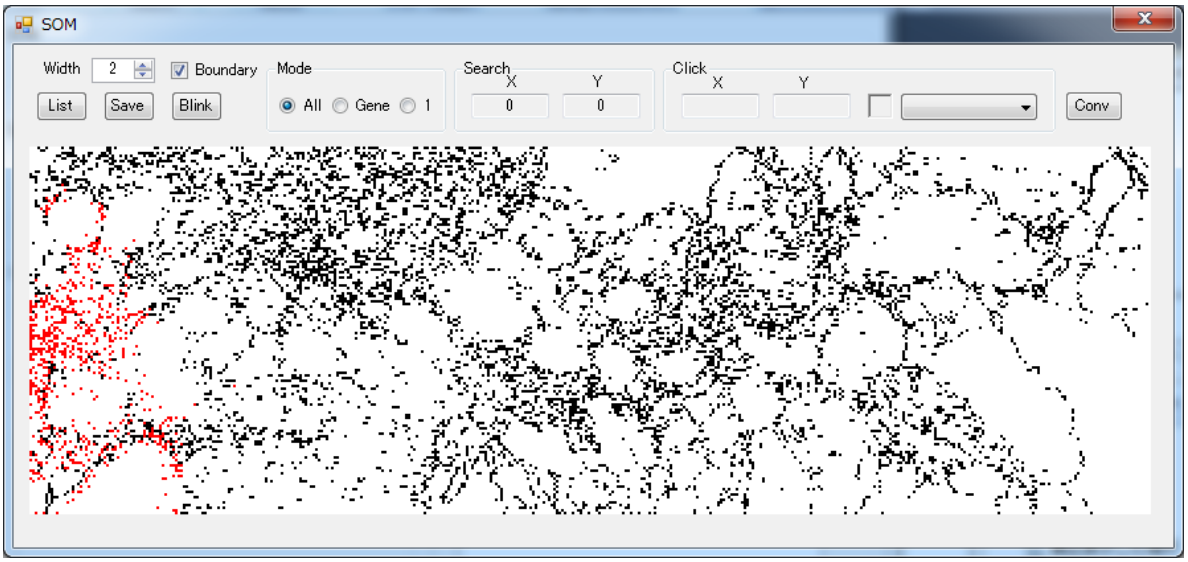
また、セル 1 個当たりの表示ピクセル数を大きくして細部の様子を見たい場合は、マップの右横と下側に付随しているスクロールバーをドラッグすることで、見たい領域を移動することができる。

Boundary のチェックボックスがオフの場合は、同じ SOM セル上に種類の異なった遺伝子が登録されていれば、そのセルには黒色が割り当てられる。遺伝子がまったく割り当てられなかった SOM セルには白色が割り当てられる。

Boundary のチェックボックスがオンの場合は、同じ SOM セル上に種類の異なった遺伝子が登録されていても、そのセルの色は黒ではなく、自分を含めて周囲２５個（５×５）のセル内に一番多く登録された遺伝子の色を仮に割り当てる。

そして、横方向に順に色の変化を調べ、色が変わっている境界セル（白色との境界変化は除外する）に黒色を割り当てる。

その後、黒色以外のセルの色をすべて白色に変更してから、ユーザの遺伝子セグメントが最大応答した SOM セルに赤色を割り当てる。



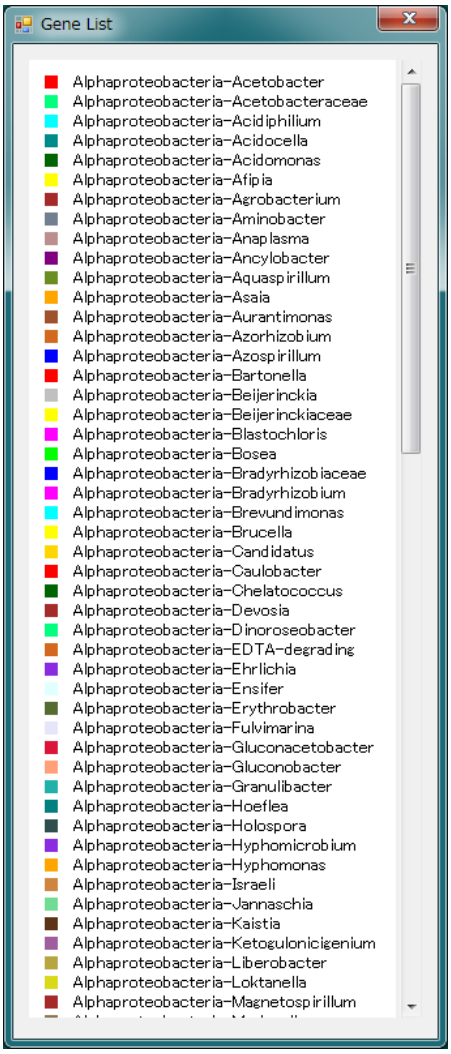
【Mode】を All に設定して【Blink】ボタンを押すと、ユーザが入力したすべての遺伝子に含まれる全てのセグメントが、SOM のマップ上で反転色で表示される。

【Mode】を Gene に設定して【Blink】ボタンを押すと、Search ダイアログの Gene No. の数値ボックスの値に設定した番号の遺伝子に含まれるすべてのセグメントが、SOM のマップ上で反転色で表示される。

Search ダイアログの Search ボタンを押して、検索を行った後であれば（詳細は後述）、【Mode】を 1 に設定して、【Blink】ボタンを押すと、Search ダイアログの Gene No. の数値ボックスの値に設定した遺伝子のうちの、Segment の数値ボックスに設定したセグメントのみが、SOM のマップ上で十字状に反転色で表示される。

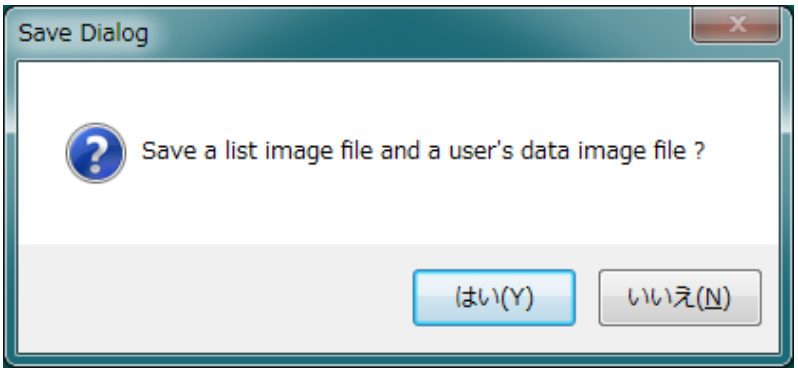
いずれのモードにおいても、マウスを押し下げている間、該当セルが反転色で表示され、マウスを離すと元の色に戻る。

本体ウィンドウ側の【List】ボタンを押すと、カテゴリの名前と色との対応の一覧表が表示される。



【Save】ボタンを押すとファイル保存のダイアログが開くので、適当なファイル名を指定してから保存ボタンを押し、SOM マップの画像を保存する。

SOM マップ画像の保存が完了すると、以下のようなダイアログが表示されるので、リストの画像や、ユーザが与えた遺伝子セグメントがマップされた場所の画像を保存したい場合は 【はい】ボタンを押す。



SOM マップ内の任意のセルをマウスでクリックすると、クリックした場所の SOM セルの番地とセルの色、その SOM セル内に分類された遺伝子セグメントのカテゴリ名が右上に表示される。

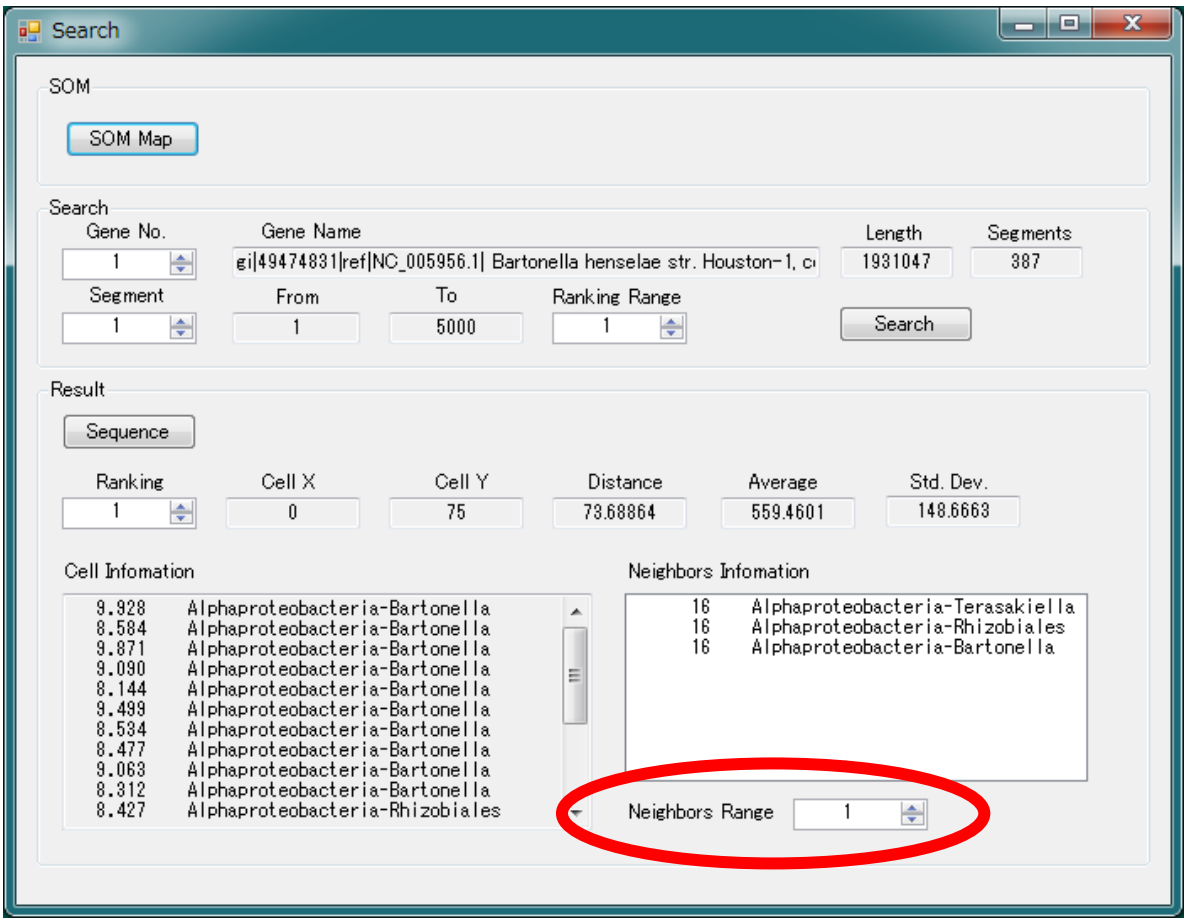
どの遺伝子セグメントも割り当てられていない SOM セルは、白色で表示される。
異なった種類のカテゴリの遺伝子セグメントが割り当てられている SOM セルは、黒色で表示される。
1 種類のカテゴリの遺伝子セグメントしか割り当てられていない SOM セルは、そのカテゴリの色で表示される。

Search ダイアログで、調べたい遺伝子の番号を【Gene No.】の数値ボックスに設定する。指定した遺伝子内のセグメントのうち、調べたいセグメントの番号を【Segment】の数値ボックスに設定する。

指定された遺伝子のセグメントのパターン頻度ベクトルと、各 SOM 内のウェイトベクトルとの距離を計算して、類似度の高い SOM セルを検索するが、このとき、類似度の上位何位までの SOM セルを検索するかを【Ranking Range】の数値ボックスに設定する。

各パラメータの設定が完了したら【Search】ボタンを押して検索処理を実行する。

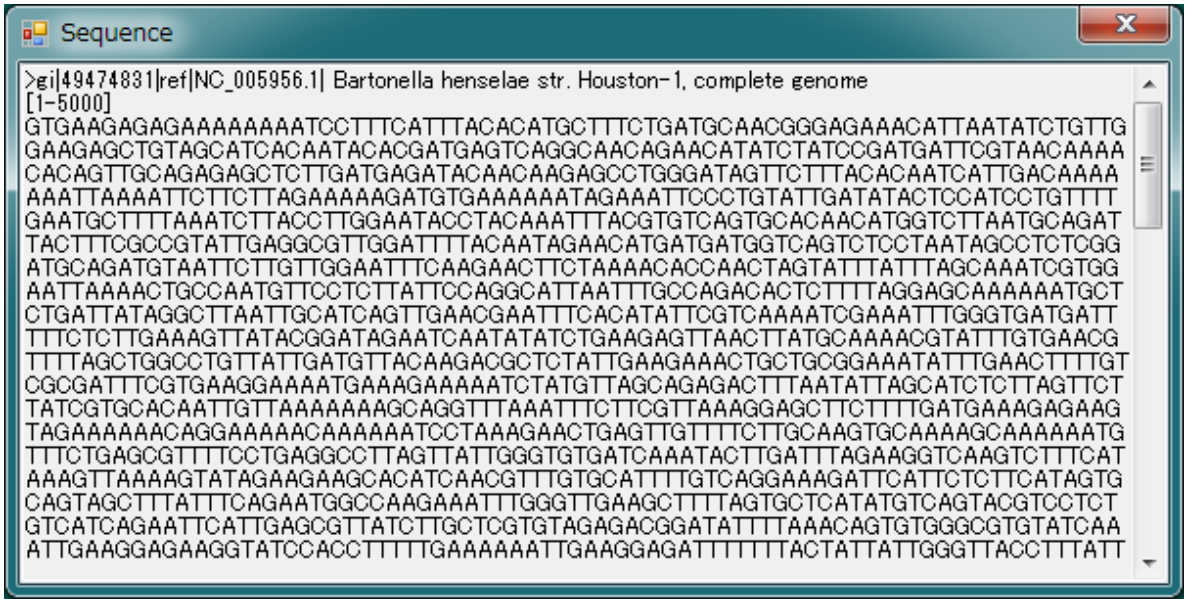
【Result】のグループ内に検索結果が表示される。
【Ranking】の数値ボックスの値を変更すると、類似度の順位の異なる SOM セルの情報に切り替わる。
【Cell Information】の欄に、該当の SOM セル内に格納されているクラスタリング結果が表示される。
【Neighbors Information】の欄に、該当の SOM セルを含む近傍の SOM セル内に格納されている カテゴリの頻度の総和が昇順に表示される。



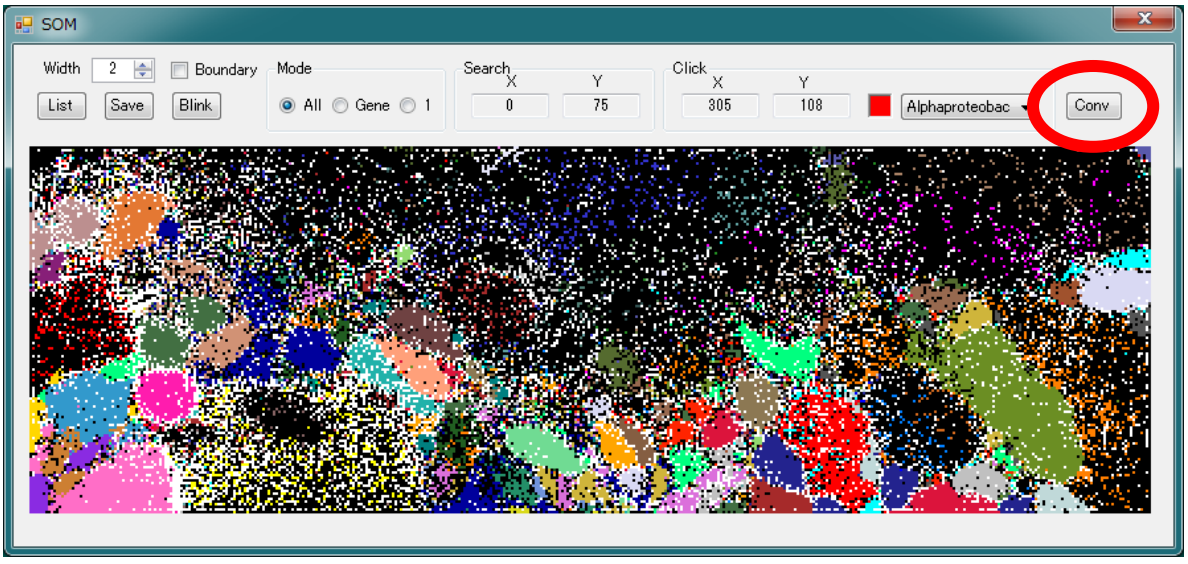
【Neighbors Range】の数値ボックスの値を変更すると、SOM セルの近傍の範囲を変更することができる。この値が1のときは、該当の SOM セルと周囲の近傍セルを含む合計 9 個の SOM セル内に登録されたカテゴリの頻度の集計結果が表示される。

検索を実行すると、【Result】のグループ内の【Sequence】ボタンなどが使用可能となる。

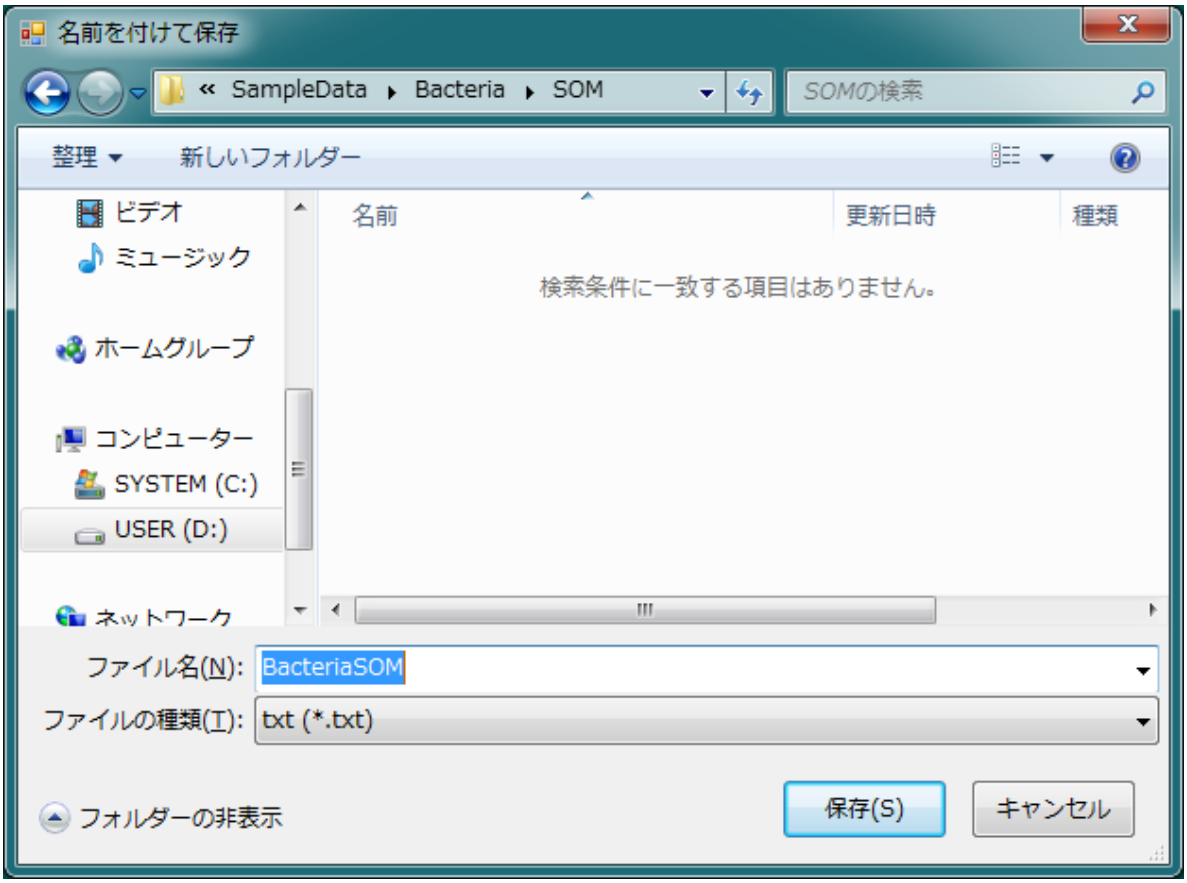
【Sequence】ボタンを押すと、指定した遺伝子セグメントのシーケンスが別ウィンドウとして開き、１行目に遺伝子名、２行目にセグメントの開始・終了位置、３行目以降にシーケンスが表示される。



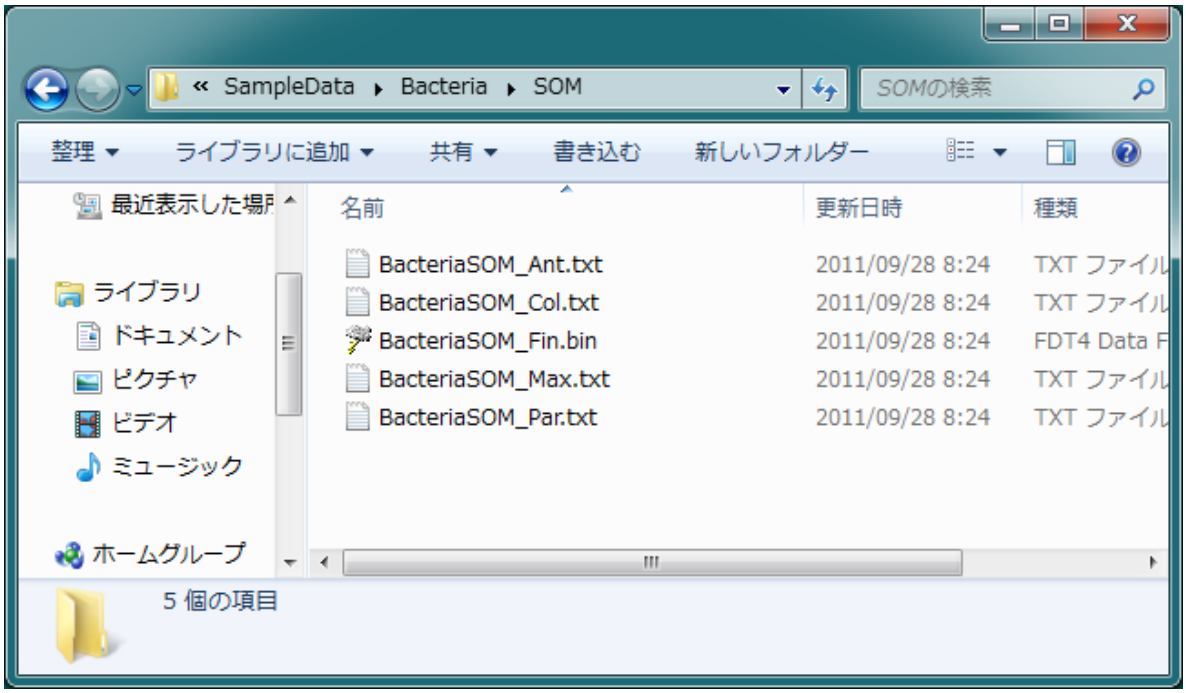
SOM マップのダイアログの一番右の Conv ボタンを押すと、SOMStudio の SOMViewer で閲覧するためのファイル一式を生成することができる。



ファイルの保存ダイアログが表示されるので、必要に応じて、フォルダの移動やフォルダの作成などを行ってから、保存するファイル名を指定してから**【保存】**ボタンを押す。

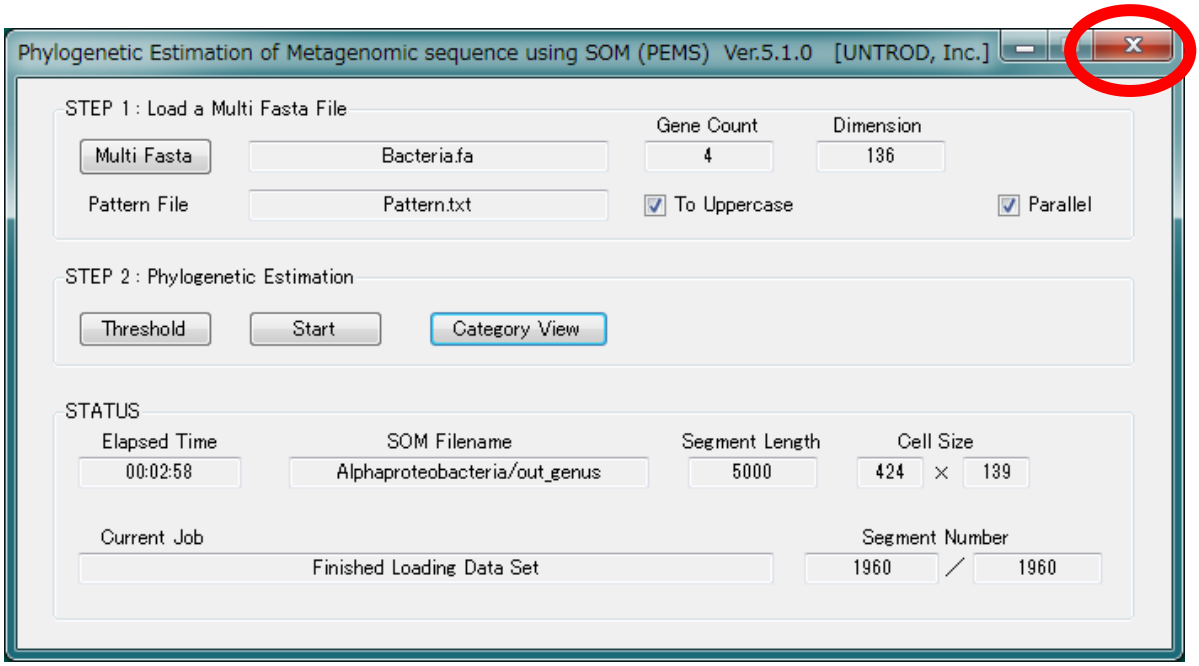


保存を開始すると Conv ボタンが使用不可となり、保存が完了すると再び Conv ボタンが使用可能となり、以下のようなファイルが生成される。

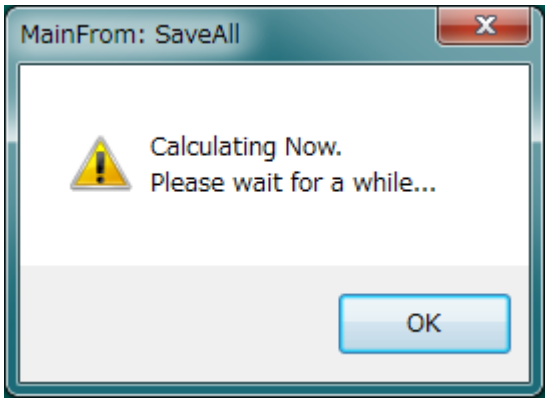


STEP 5：アプリケーションの終了

PEMS アプリケーションを終了する場合は、本体ウィンドウ右上の赤いバツ印のボタンを押す。



計算が終了する前に本体ウィンドウを閉じようとすると、下図のようなダイアログが表示されるので、【OK】ボタンを押し、計算が終了するまでしばらく待つ。



ウィンドウの背景色を変更したい場合は、インストールしたフォルダ内の PEMS.ini ファイルの8行目にRGBの値を16進数で追加表記(例：FE907E)することで変更できる。

5. 解析手法の概要

本ソフトウェアの系統推定のプロセスの概要は以下のようになっている。

1. ユーザが指定した各遺伝子のシーケンスを、決められた長さのセグメントに分断し、各セグメント内のパターン頻度ベクトルを計算する。
2. 各セグメントのパターン頻度ベクトルと最も類似度の高いウェイトベクトルを持つ SOM セルを検出する。
3. 検出された SOM セルを含む、近傍内の 各 SOM セルに登録されたカテゴリの頻度を計算する。
4. 推定されたカテゴリ頻度を降順にソートする。
5. 推定されたカテゴリの下の階層の SOM データが用意されている場合には、下の階層に進み、上記の一連のプロセスを実行する。

上記の1の計算処理において使用するヌクレオチドのパターンについては、SOM のデータセットをインストールしたフォルダ内の Pattern.txt ファイルに記載されている。このファイルは SOM データの解析結果と関係しているため、編集して変更してはならない。

分断するセグメントの長さは、 SOM のクラスタリング処理の際に用いたセグメント長が自動的に指定される。ユーザが解析用に指定した遺伝子の全長が、このセグメント長に満たない場合であっても、デフォルトで設定されている閾値よりも遺伝子が長ければ、頻度ベクトルを正規化補正して解析を行う。この閾値を変更したい場合には、インストールしたフォルダ内の PEMS.ini ファイルの 5 行目の数字(長さの閾値)を編集することで変更が可能である。

また、セグメント長よりも長い遺伝子の場合には、遺伝子をセグメント長ごとに分断して頻度解析を行うが、分断の最後の残余セグメントについても、この閾値の長さ以上であれば、頻度ベクトルを正規化補正して解析の対象とする。

遺伝子のセグメント内に指定されたパターン以外のヌクレオチドパターンが現れた場合は、それらのパターンはカウントから除外され、指定されたパターンの頻度ベクトルのみが正規化補正される。

上記の2～4のプロセスにより、各遺伝子内の各セグメントについての系統推定が行われ、推定に使用した SOM データに対応してファイルに保存される。

例えば、ユーザが保存ファイル名として、myResult.txt を指定した場合は、kingdom 階層の out_kingdom の SOM 解析ファイルに基づいて系統推定した結果は、myResult_kingdom.txt ファイルに保存され、genus 階層の Actinobacteria の SOM 解析ファイルに基づいて系統推定した結果は、myResult_Actinobacteria.txt ファイルに保存される。

ファイルには以下のような内容が記述される。

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

で始まる行が遺伝子名で、次の行から各セグメントの計算結果が示される。

最初の [1155001-1160000] は、セグメントの切り出し位置を示している。
次の 2 16 の 2 つ組みの数字は、このセグメントのパターン頻度ベクトルと最も類似度の高い(ベクトル間距離の短い)ウェイトベクトルを持った SOM セルのX座標とY座標を示している。
この最大類似度を示した SOM セルを含む近傍 9 個の SOM セル内に登録されていたカテゴリ名の頻度を集計し降順にソートした結果が、SOM セルの座標に続いて表示される。

遺伝子内のセグメントのカテゴリ頻度を、単純に総計した結果が、
myResult_Actinobacteria_All.txt
のように、_カテゴリ名_All の文字列が間に挿入された状態のファイル名で保存される。

頻度順に並べられた、カテゴリの 1 位に同率 1 位のカテゴリが存在する場合と、存在しない場合がある。
ここでは、同率 1 位のカテゴリが存在しない 1 位のカテゴリを**真正 1 位**と呼ぶことにする。

遺伝子内のセグメントで、真正 1 位のカテゴリのみを集計した結果が、
myResult_Actinobacteria_All1st.txt
のように、_カテゴリ名_All1st の文字列が間に挿入された状態のファイル名で保存される。

同一レベルの階層内における、真正 1 位のカテゴリの集計結果を総合して、各遺伝子について、一番頻度の高かったカテゴリを抽出した結果が、myResult_Top.txt というファイル名で保存され、これと同じ内容が、【Display】ボタンによってツリー表示される。

自動的に階層的な計算処理が行われ、各セグメントについて、カテゴリ頻度 1 位の頻度比率を計算する。
例えば、myResult_phylum.txt 内に記載された下記のようなセグメントの場合、

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

カテゴリの第 1 位である Alphaproteobacteria の頻度比率は、
(25 / (25 + 8 + 2)) * 100 = 71.4 %
となる。
この頻度比率が、【Threshold】の数値ボックスで指定した値よりも大きい場合は、次の階層である Alphaproteobacteria での系統推定のためのセグメントとして使われる。

同じ遺伝子内のセグメントどうしではあっても、それぞれのセグメントのカテゴリ頻度 1 位の順位が異なれば、下の階層において、異なったカテゴリでの系統推定のためのセグメントとして分類される。
また、先ほどのカテゴリ頻度の集計処理とは異なり、下位の階層にセグメントを区分けする際に同率 1 位のカテゴリが複数存在する場合には、そのセグメントは、それぞれのカテゴリに重複して分類され、系統推定のためのデータとして使われる。

6. 補足

■ マルチファスタファイルの形式

各遺伝子の名前が > の文字の後に続く。
遺伝子名の行については、1 文字目が > で始まり、1 行で記述されること以外、形式は自由である。

遺伝子名を記載した行の次の行からは、実際のシーケンスが記述される。
シーケンスが複数行にまたがる場合には、1 行の文字数に特に制限はないが、非常に長いシーケンスとして記述するより、適当な文字数で改行を挿入したほうが計算効率が良い。シーケンスの長さの制限は特にないが、長さの上限は、使用する PC の搭載メモリに依存する。

上記の仕様であれば、1 つのファイル内に含まれる遺伝子の数に制限はないが、1 つのファイルの処理時間が長くなることや、メモリ使用量が多くなることなどを考えると、適当に分類して 1 ファイル内に含まれる遺伝子数を調整することが望ましい。

ファイル名の拡張子は、.fa もしくは .fas を指定する。

【例】

```
>gi|49474831|ref|NC_005956.1| Bartonella henselae str. Houston-1, complete genome
GTGAAGAGAGAAAAAAAATCCTTTCATTTACACATGCTTTCTGATGCAACGGGAGAAACATTAATATCTG
TTGGAAGAGCTGTAGCATCACAATACACGATGAGTCAGGCAACAGAACATATCTATCCGATGATTCGTAA
CAAAACACAGTTGCAGAGAGCTCTTGATGAGATACAACAAGAGCCTGGGATAGTTCTTTACACAATCATT
GACAAAAAAATTAAAATTCTTCTTAGAAAAAGATGTGAAAAAATAGAAATTCCCTGTATT
>gi|78042616|ref|NC_007503.1| Carboxydothermus hydrogenoformans Z-2901, complete genome
AACCTGAAAAAAGTGTGAAAAAAATTTTGTGGATTTGTGGATAAAACAAGGTTTTTGCTAATTTTTGCTA
ATAAAAAAAATTTATAAAGAGATTCGTGAAAGCAAAGATTGTGGATAACGAAAAACTCAAGAAAAATTTC
TTGACGGGTCTTATCCCATCTTCTATAATTTAGGTGTATCTATAGGGGATATAGGCTTTTTTAGATAGAT
TAGGAGGTGTGAAAATGAAAAGAACCTACCAACCAAAAAACCGGCGGCG
```