# PEMS ソフトウェア

 $\label{eq:phylogenetic} \textbf{P} hylogenetic \ \textbf{E} stimation of \ \textbf{M} etagenomic \ sequence \ using \ \textbf{S} OM$ 

# ユーザーズガイド

 $\mathbf{2}$ 

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
<b>Network</b> :	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。

					x
	USER (D:) PEMS	•	<b>↓</b> PEMS	の検索	٩
整理 ▼ ライブラリに追加 ▼	共有 ▼ 書き込む	新しいフォルダー		<u>►</u> = ▼	0
<ul> <li>☆ お気に入り</li> <li>ダウンロード</li> <li>デスクトップ</li> <li>漫 最近表示した場所</li> </ul>	SOM.zip ZIP 書庫 528 MB		PEMS.zip ZIP 書庫 34.2 KB		
2個の項目					

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。

		- (1)	PEMSの検索	
整理 ▼ 😭 開く	ライブラリに追加 ▼ 共有 ▼	・ <sup>・</sup> ア 電子メールで送信する »		
■ デスクトップ ^ 1 最近表示した場所	<b>50M</b> ファイル フォルダー	PE PEMS.ex MS UNTROD	e , Inc.	
<ul> <li>⇒イブラリ</li> <li>ドキュメント</li> <li>ビクチャ</li> <li>ビデオ</li> <li>ミュージック</li> </ul>	PE PEMS.ico アイコン MS 2.18 KB	PEMS.ini 構成設定 76 バイト	~	
<b>SOM</b> ファイル フォル	更新日時: 2011/09/28 8:44 ダー			

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。

	PEMS.exe
	PEMS
MS	UNTROD, Inc.

STEP 1 lead a Multi Factor	etagenomic sequence using S	OM (PEMS) Ver.5.5.0	[UNTROD, Inc.]
Multi Fasta	FIIE	Gene Count 0	Dimension 136
Pattern File	Pattern.txt	V To Uppercase	V Parallel
STEP 2 : Phylogenetic Estim	nation		
Threshold S	tart Category View	]	
STATUS Elapsed Time	SOM Filename	Segment Length	n Cell Size
		5000	0 × 0
			Segment Number
Current Job			

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。 【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)

開く				<b></b>
	7− → GENOME (K:) → Bacteria	<b>→</b> \$9	Bacteriaの検索	٩
整理 ▼ 新しいフォルダー			:==	• 🔳 🔞
☆ お気に入り	名前	更新日時	種類	サイズ
🚺 ダウンロード 📲	Bacteria.fa	2007/08/01 12:51	FA ファイル	9,696 KB
📃 デスクトップ				
📃 最近表示した場所	-			
📃 録画一覧				
🍃 ライブラリ				
🖹 ドキュメント				
🔄 ピクチャ				
■ ビデオ	-			
ファイル	名( <u>N</u> ): Bacteria.fa	-	Fasta file (*.fa;*.	fas;*.fna) 👻
			開<(0)	キャンセル

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。

Phylogenetic Estimation of Me	tagenomic sequence using	SOM (PEMS) Ver.5.5.0	[UNTROD, Inc.]
STEP 1 : Load a Multi Fasta Multi Fasta Pattern File	File Bacteria.fa Pattern.txt	Gene Count 4 V To Uppercase	Dimension 136 V Parallel
STEP 2 : Phylogenetic Estim	ation art Category View		
STATUS Elapsed Time	SOM Filename	Segment Length 5000	n Cell Size
Current Job			Segment Number

## ■ 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。

名前を付けて保存				×
G C マレ・コンピューター ・ GENOME (ド	<:) 🕨 Bacteria 🕨 Result	👻 🍫 Re	esultの検索	٩
整理 ▼ 新しいフォルダー				•= • 📀
🏭 SYSTEM (C.) 名前	^ 更新	f日時	種類	サイズ
ARCHIVE (D:) USER (E:) @ CD-RW ドライン リムーバブル デ	検索条件に一致する項	目はありません。		
GENOME (K:) ≡	m			•
ファイル名( <u>N</u> ): myResult				-
ファイルの種類( <u>T</u> ): txt (*.txt)				-
▲ フォルダーの非表示			保存(S)	キャンセル

【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。

hylogenetic Estimation of STEP 1 : Load a Multi Fas	Metagenomic sequence using SO sta File	M (PEMS) Ver.5.5.0	[UNTROD, Inc.]
Multi Fasta	Bacteria.fa	4	136
Pattern File	Pattern.txt	📝 To Uppercase	🔽 Parallel
STEP 2: Phylogenetic Es	timation		
Threshold	Start Category View		
STATUS			
Elapsed Time	SOM Filename	Segment Length	n Cell Size
00:02:56	Verrucomicrobia/out_genus	5000	79 × 34
Current Job			Segment Number
C	ompleted Saving Data of Ranking		

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。

名前を付けて保存				<b>— X</b>			
בשעב א 📕 - 🕞	ーター 🕨 GENOME (K:) 🕨 Bacteria 🕨 Resu	lt 👻 🍫 🧖	esultの検索	٩			
整理 ▼ 新しいフォルタ	Ÿ—			≣ ▾ 🔞			
SYSTEM (C:) 🔺	名前	更新日時	種類	サイズ 🔺			
👝 ARCHIVE (D:)	myResult_Verrucomicrobia_All1st.txt	2011/10/05 22:10	TXT ファイル				
👝 USER (E:)	myResult_Verrucomicrobia_All.txt	2011/10/05 22:10	TXT ファイル				
💿 CD-RW ドライン	myResult_Verruconne. bia.txt	2011/10/05 22:10	TXT ファイル				
👝 リムーバブル デ	myResult_Top.txt	2011/10/05 22:10	TXT ファイル				
GENOME (K:)	StravBesult_Spirecht_tes_All1st.txt	2011/10/05 22:10	TXT ファイル				
	myResult_Spirochaetes_All.txt	2011/10/05 22:10	TXT ファイル				
😘 ネットワーク	myResult_Spirochaetes.txt	2011/10/05 22:10	TXT ファイル	+			
	•			•			
ファイル名( <u>N</u> ): myRe	sult_Top.txt			-			
ファイルの種類( <u>T</u> ): txt (*	.txt)			-			
● フォルダーの非表示	<ul> <li>フォルダーの非表示</li> <li>マオルダーの非表示</li> </ul>						

【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。

	Category Selector					_ <b></b>		
(	Gene No.	Total 101833	(	Gene Keyword		Find		
	<ul> <li>aradoz25OH01AKEJJ rank=0000031 x=115.0 y=1037.5 length=389</li> <li>☐ Prokaryotes : 100.0</li> <li>☐ Gammaproteobacteria : 100.0</li> <li>☐ Gammaproteobacteria-Lamprocystis : 100.0</li> </ul>							
	Load	Search	Result	SOM Category				

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。

Category Selector					x
Gene No. 1024 🚔 🗡	Total 101833	Ge	ene Keyword ZSO	Find	
⊟#G08ZSOH01AQ ÈProkaryotes ÈBacteroid	YA8 rank=00030 : 100.0 detes : 100.0 eroidetes-Bacter	12 x= 105.1 - 2400 oides : 100.0	A length=503		
Load	Search	Result	SOM Category		

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					
Search Gene No. 1 🔄 Segment 1 🚖	Gene Name gi 49474831 ref N From 1	IC_005956.1  Bar To 5000	tonella henselae str. Houston-1, c Ranking Range 1 😜	Length 1931047 Search	Segments 387
Result Sequence Ranking Cell Infomation		Cell Y	Distance Average	Std. D	lev.

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。

💀 Search					
SOM					
SOM Map					
Search					
Gene No.	Gene Name			Length	Segments
1 🌲	gi 49474831 ref N	C_005956.1  Barto	onella henselae str. Ho	uston-1, ci 1931047	387
Segment	From	То	Ranking Range		
1 🚖	1	5000	1	Search	
Result Sequence					
Ranking	Cell X	Cell Y	Distance	Average Std. Dev	
1 🚔	0	75	73.68864	559.4601 148.666	3
Cell Information			Neighbors I	infomation	
9.928 Alp 8.584 Alp 9.871 Alp 9.090 Alp 8.144 Alp 8.499 Alp 8.534 Alp 8.477 Alp 9.063 Alp	haproteobacteria- haproteobacteria- haproteobacteria- haproteobacteria- haproteobacteria- haproteobacteria- haproteobacteria-	Bartonella Bartonella Bartonella Bartonella Bartonella Bartonella Bartonella	▲ 16 16 16	Alphaproteobacteria-Te Alphaproteobacteria-Rh Alphaproteobacteria-Ba	asakiella izobiales tonella
8.312 Alp 8.427 Alp	haproteobacteria- haproteobacteria- haproteobacteria-	Bartonella Rhizobiales	- Neighbors F	Range 1 👤	

【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ー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。

12 名前を付けて保存	<b></b>
🕞 🕞 🗸 « SampleData 🖡 Bacteria 🖡 SOM 🛛 🗸 🍫 SOMの検索	٩
整理 ▼ 新しいフォルダー	= • 🔞
ビデオ     ▲     名前     更新日時	種類
ミュージック 検索条件に一致する項目はありません。	
🜏 ホームグループ	
■ コンピューター ▲ SYSTEM (C:) ■ USER (D:)	
	÷.
ファイル名( <u>N</u> ): <mark>BacteriaSOM</mark>	-
ファイルの種類( <u>T</u> ): txt (*.txt)	•
▲ フォルダーの非表示	キャンセル

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。

Samplet 🖉 🖉	Data 🖡 Bacteria 🖡 SOM	👻 🍫 SOMO	の検索 👂
整理 ▼ ライブラリに)	追加 ▼ 共有 ▼ 書き込む	新しいフォルダー	III 🔹 🚺 🔞
3 最近表示した場点▲	名前	更新日時	種類
	BacteriaSOM_Ant.txt	2011/09/28	8:24 TXT ファイル
🍃 ライブラリ	BacteriaSOM_Col.txt	2011/09/28	8:24 TXT ファイル
🖹 ドキュメント 💡	🏸 BacteriaSOM_Fin.bin	2011/09/28	8:24 FDT4 Data F
📔 ピクチャ	BacteriaSOM_Max.txt	2011/09/28	8:24 TXT ファイル
📕 ビデオ	BacteriaSOM_Par.txt	2011/09/28	8:24 TXT ファイル
👌 ミュージック			
🜏 ホームグループ 🖕	•		•
5 個の項目			

# 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。

		Gene Count	Dimension	
Multi Fasta	Bacteria.ra	4	130	
Pattern File	Pattern.txt	🔽 To Uppercase		🔽 Parallel
STATUS				
	SOM Eilename	Segment Length	Cell Size	•
Elapsed Time	Contrinciano			198
Elapsed Time 00:02:58	Alphaproteobacteria/out_genus	5000	424 ×	109
Elapsed Time 00:02:58 Current Job	Alphaproteobacteria/out_genus	5000	424 × Segment Nu	mber

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】