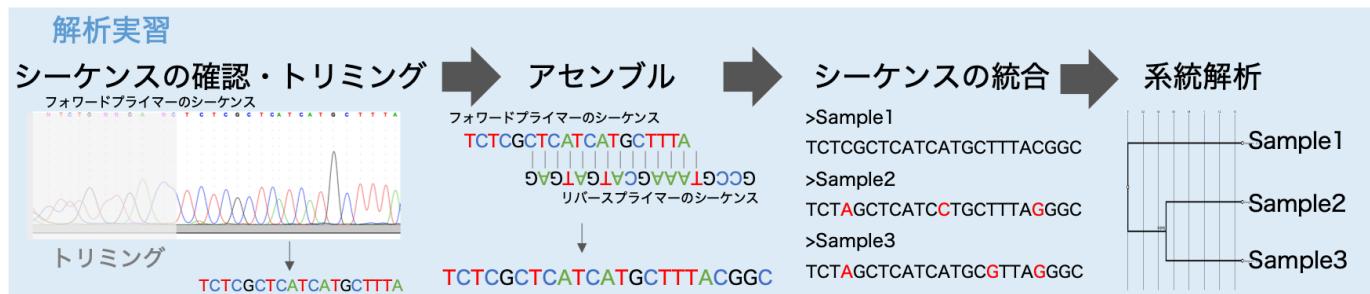


解析実習（分子系統解析）

今回、サンガーフラッシュ法で得られたシーケンスを使って系統解析（系統樹の作成）をおこなっていきます。

解析の流れ

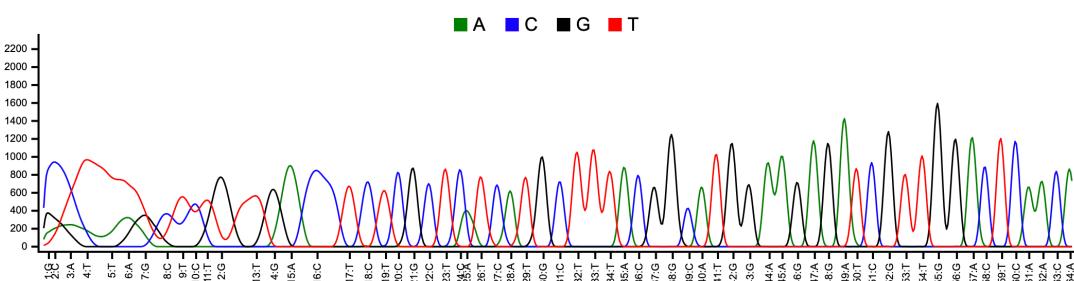
1. シーケンスの確認・トリミング
2. アセンブル
3. シーケンスの統合
4. 系統樹の作成
5. 類似シーケンスの検索



1. シーケンスの確認・トリミング

1.1 サンガーシーケンスの波形ファイル（AB1ファイル）の確認

サンガーシーケンスをおこなうと波形ファイル（各塩基のシグナル強度のデータ）が得られます。そのデータをみることで、シーケンスの精度を評価できます。



波形データファイル（AB1ファイル）は、以下のURLの各班のフォルダのなかのab1__PCR領域名 フォルダに入っています。

シーケンスデータの保管フォルダ:

https://drive.google.com/drive/folders/1eqNgUu_JTyvVZYPuPNX8amMCn5AVo5KX?usp=sharing

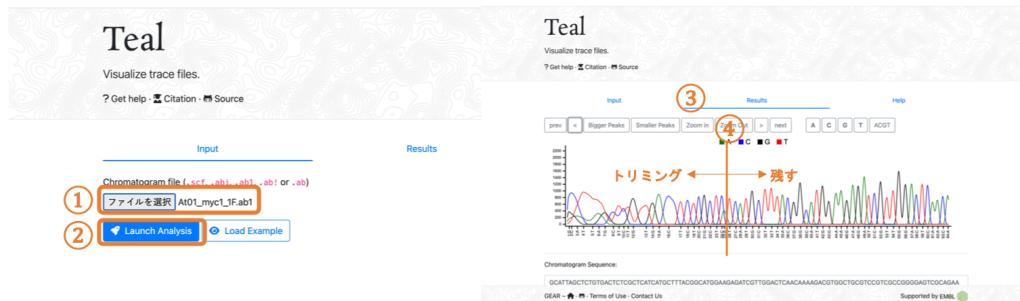
保管フォルダ内のフォワードプライマーで得られたAB1ファイルをひとつ選んで、ダウンロードしてください。

そのAB1ファイルをWebツール「Teal」にアップロードし、確認してみましょう。

<https://www.gear-genomics.com/teal/>

AB1ファイルの確認手順:

1. Inputタブで「ファイルを選択」をクリックし、ダウンロードしたAB1ファイルを選ぶ
2. 「Launch Analysis」をクリックする
3. 自動的にResultsタブに切り替わり、波形データと塩基配列が表示される
4. 波形の状態を目視で確認し、塩基のシグナルが明確な範囲を決定する



1.2 シーケンスのトリミング（切り出し）

シーケンスのトリミングには「EMBOSS: extractseq」を使います。

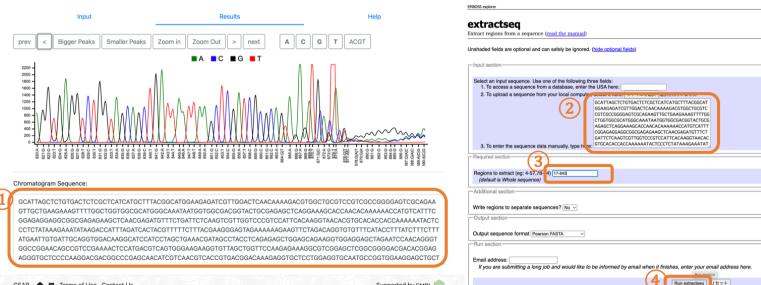
<https://www.bioinformatics.nl/cgi-bin/emboss/extractseq>

トリミングの手順

1. 「Teal」のResultsタブに表示されている塩基配列をコピーする
2. コピーした配列を「EMBOSS: extractseq」のテキストボックスに貼り付ける
3. 切り出す範囲を"Regions to extract"のフォームに入力する

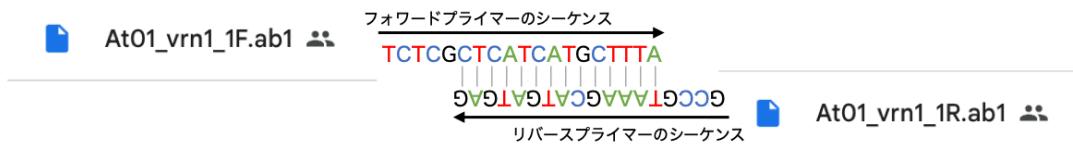
4. 「Run extract」をクリックする

5. 出力された塩基配列をコピーし、テキストエディタ（Wordやメモ帳など）に貼り付けておく



練習

相補鎖側の配列（リバースプライマーで読んだシーケンス）に対しても1.1-1.2の操作をおこなってください

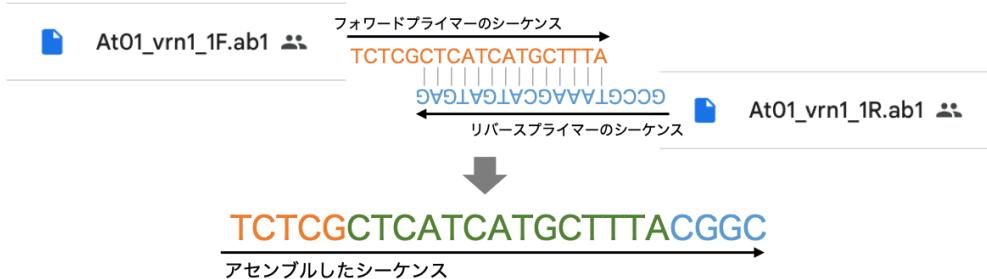


2. アセンブル

2.1 アセンブル

部分的に相同性のある2つ以上の配列を統合し、1つのより長い配列を構築することをアセンブルと言います。

今回の実習では、ひとつの遺伝領域をフォワードプライマーとリバースプライマーを使って二方向からシーケンスを取得しました。その二つのシーケンスをアセンブルし、ひとつの塩基配列にしましょう。



サンガーシーケンスのアセンブルには「CAP3」と呼ばれるソフトウェアがよく使われています。今回、そのWebサービス版を使用します。

<https://doua.prabi.fr/software/cap3>

アセンブルの手順:

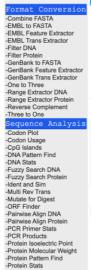
- ステップ1.2で得たフォワード側とリバース側の塩基配列を「CAP3」のテキストボックスに貼り付けて、「SUBMIT」をクリックする
- アセンブルが終わると、自動的にResultsページに移行する
- 「Assemble details」でどのようにアセンブルされたかを確認する。このときフォワード側のシーケンス名に "+"、リバース側のシーケンスが "-" が付いていることを確認する
 - +: 入力したシーケンスがアセンブルに使われたことを表す
 - : 入力したシーケンスの相補鎖配列がアセンブルに使われたことを表す
- 「Contigs」にアセンブル結果の塩基配列（FASTA形式で記述されている）がoutputされています

Number of segment pairs = 2; number of pairwise comparisons = 1
'+' means given segment; '-' means reverse complement

Overlaps	Containments	No. of Constraints Supporting Overlap
***** Contig 1 *****		
EMBOSS_001F-	EMBOSS_001R+ is in EMBOSS_001F-	
DETAILED DISPLAY OF CONTIGS		
***** Contig 1 *****		
EMBOSS_001F-	TAATCAGCTGCTACCAAGACCTCACCTGAGCTGTCATCTTGAGAGGCCAGC	
EMBOSS_001R+	TCAGCTGCTACCAAGACCTCACCTGAGCTGTCATCTTGAGAGGCCAGGAGA	
consensus	TAATCAGCTGCTACCAAGACCTCACCTGAGCTGTCATCTTGAGAGGCCAGGAGA	

2.2 相補鎖変換

アセンブル結果が相補鎖配列の場合（フォワード側が "+"、リバース側が "-" でアセンブルされた場合）、以下のWebツールを使って、その配列を相補鎖変換する
https://www.bioinformatics.org/sms2/rev_comp.html



Reverse Complement

Reverse Complement converts a DNA sequence into its reverse, complement, or reverse complement. You may want to work with the case of each input sequence character is maintained. This is useful for comparing sequences.

Paste the raw sequence or one or more FASTA sequences into the text area below.

Check the browser compatibility page before using this program.

Submit | Clear | Reset | reverse-complement |

*This page requires JavaScript. See browser compatibility.
You can mirror this page or use it off-line.

Reverse Complement results

```
>Contig1 reverse complement
TCTCGCTCATCATGCTTTACGGCATGGAGAGATCGTTGACTCAAAAGACGTGGCT
GGCTCCGGCGCCGGGGAGTCGGAGAAGTTGCTGAAAGAATTGTTGGCTGGGGCATGG
GCAAATTAATGTGGCGACCGTACCGGGAGAGCTCAGGAAGAACCCACACAAAAAACCAT
GTCATTCTGGAGAGGGCGCGAGAGCTCAAGAGATGTTCTGATTTCTCAAGTCG
TTGGTCCCCTCCATTCAAAAGGTAAACACGTGACACACACACACACACACACACACAC
AGAAATATAAGACCAATTAGATACAGTTTTCTTCTGAGAGGAGTGAAGAAAAGAAA
GTTCTAGAGGGTGTTCATACCTTAACTCTTCTTATGAAATTGTGATTTCAGGTGGA
CAAGGCATCCATCCATGCTGAAGATAGCTGAACTACCTCAGAGAGCTGGAGCAAGAGTGG
GGAGCTGAGAAATCCAACAGGGTGGCGGGAGACAGCCGTCGAGAAACTCTCATGACGTCAGTGG
GAAGAAAGGTGTTAGGGTTCAAGAGGAAGGGCGCTGGAGCTGGGGGAGCACACCGA
GAAGGGTGTGCTCCCAAGAGCAAGGGACAGGGACAGGGACACATCGTCAACCGTACCCGTGACGGACAA
AGAGGTGCTCTGGAGGTGCAATGCCCTGGAGCTGCTCTGGGGGGTGAAGGGGGACGGCT
CCTCGCTCAAGATACAGACTCAGGTGAGGTCTGGTAGACGTGATTA
```

3. シーケンスの統合

この後の系統解析のためには、各サンプルの塩基配列を集め必要があります。

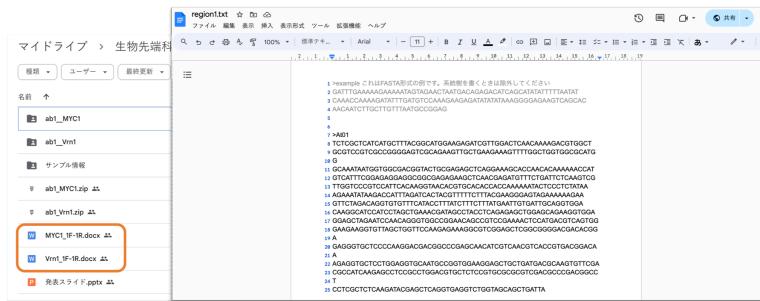
Google ドライブの共有フォルダ内に、各サンプルのアセンブル配列を集めるためのファイル（領域名.docx）を置いています。

Google ドライブ共有フォルダ：

https://drive.google.com/drive/folders/1eqNgUu_JTyvVZYPuPNX8amMCn5AVo5KX?usp=sharing

そのファイルにアセンブル配列をコピー＆ペーストしてください。

重要：それぞれの配列の名前 (FASTA形式の配列名) をサンプル名に変更してください



4. 系統樹の作成

系統樹の作成もWebサービス（例えば、NGPhylogeny.fr; 下記URL）を使って簡単におこなえます。

<https://ngphylogeny.fr/>

今回は、どのように系統樹を作成しているかを理解するために「A la Carte」モードで実行していきましょう。「A la Carte」では、系統樹作成の各ステップのツールを自分で指定できます。

NGPhylogenyの系統樹作成は4ステップでおこなわれます。

1. マルチプルアライメント (Multiple Alignment) : サンプル間の塩基配列が揃うように並べる
2. アライメントの整形 (Alignment Curation) : 欠失データなどが多いボジションの情報は除外し、使用可能な塩基のみを残す
3. 系統樹の推定 (Tree Inference) : サンプル間の塩基の類似度を算出し、系統関係を推定する
4. 系統樹の描画 (Tree Rendering) : 推定した系統樹を描画する

1. マルチプルアライメント

```
s1  GAATTCTCGCTCAATCATGCG
s2  TCTAGCTCAATCTGCGCCATT
s3  AATTCTAGCTCAATCATGCGCC
s4  TTCTAGCT--ATCCGTAGGCCA
```

2. アライメントの整形

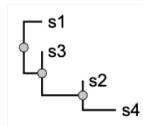
```
s1  GAAATCTCGCTCAATCATGCG
s2  TCTAGCTCAATCTGCGCCATT
s3  AATTCTAGCTCAATCATGCGCC
s4  TCTAGCT--ATCCGTAGGCCA
```

使用する領域

3. 系統樹の推定

	s1	s2	s3	s4
s1	1.00	0.86	0.93	0.80
s2	0.86	1.00	0.93	0.93
s3	0.93	0.93	1.00	0.86
s4	0.80	0.93	0.86	1.00

4. 系統樹の描画



実習での系統樹作成の手順:

1. 「A la Carte」をクリック
2. 各ステップは以下のツールを選び、「Create workflow」をクリックする
 - Multiple Alignment: MAFFT
 - Alignment Curation: BMGE
 - Tree Inference: FastME
 - Tree Rendering: Newick Display
3. 前ステップ「3. シーケンスの統合」で準備したFASTA形式データをすべてコピーし、Input dataのテキストボックスに貼り付ける
4. 各ステップのツールのオプションを次のようにする
 - MAFFT: デフォルトのまま（変更しない）
 - BMGE: 「Gap Rate cut-off [0-1]」の値を0にする。今回、挿入欠失変異（InDel）を考慮せずに、SNPのみを考慮して系統解析をおこなう
 - FastME: 「Bootstrap branch supports」をYesにする。このオプションで出力されるブートストラップ値は、系統樹の各枝の信頼度の指標になる
 - Newick Display: デフォルトのまま（変更しない）

5. 類似シーケンスの検索

次世代シーケンサー第3次シーケンサー（ロングリードシーケンサー）が登場して以降、多くの生物でゲノムが解読されるようになってきました。解読されたゲノム配列はDNAデータバンクなどで公開されています。

おもなDNAデータバンク:

- NCBI (National Center for Biotechnology Information; アメリカ)
- DDBJ (DNA Data Bank of Japan; 日本)
- ENA (European Nucleotide Archive; イギリス)

実習で使ったタルホコムギやイネ、それらの近縁種も代表的な系統のゲノム配列が公開されており、今回の解析で利用できます。

コムギの公開ゲノム配列:

- Triticum aestivum (パンコムギ) https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_018294505.1/
- Triticum aestivum subsp. spelta (スペルトコムギ) https://www.ncbi.nlm.nih.gov/datasets/genome/GCA_903994165.1/

イネと近縁種の公開ゲノム配列:

- Oryza sativa Japonica Group (品種: Nipponbare) https://www.ncbi.nlm.nih.gov/datasets/genome/GCA_003865235.1/
- Oryza sativa indica subgroup (品種: ZH8015) https://www.ncbi.nlm.nih.gov/datasets/genome/GCA_034818605.1/

ここでは、公開ゲノム配列から実習の遺伝領域の塩基配列を得てみましょう。

手順:

- 上記の公開ゲノム配列ページに移動する
- 「BLAST the reference genome」をクリックする
- 実習で取得した遺伝領域の塩基配列（※1）を“Enter Query Sequence”的テキストボックスに貼り付ける
※1 長い塩基配列が得られている系統であれば、どの系統でもOK
- 画面下のほうにある「BLAST」をクリックする
- 検索結果のうち、上位の結果をクリックする
- 「GenBank」をクリックする
- 表示されたページの「FASTA」をクリックする