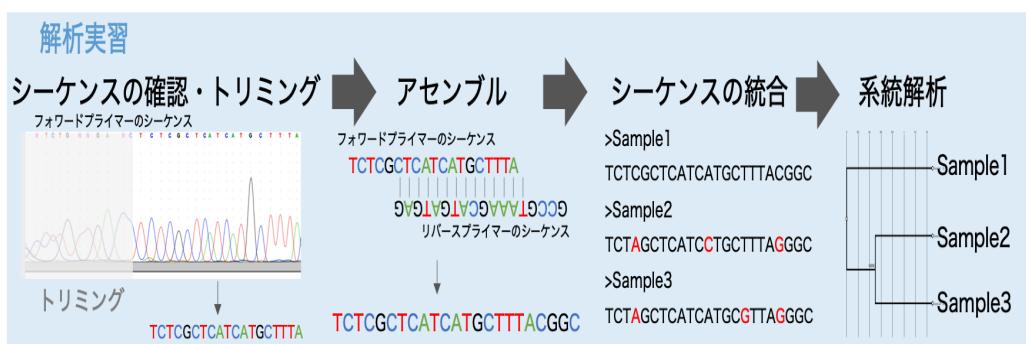


解析実習（分子系統解析）

今回、サンガー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

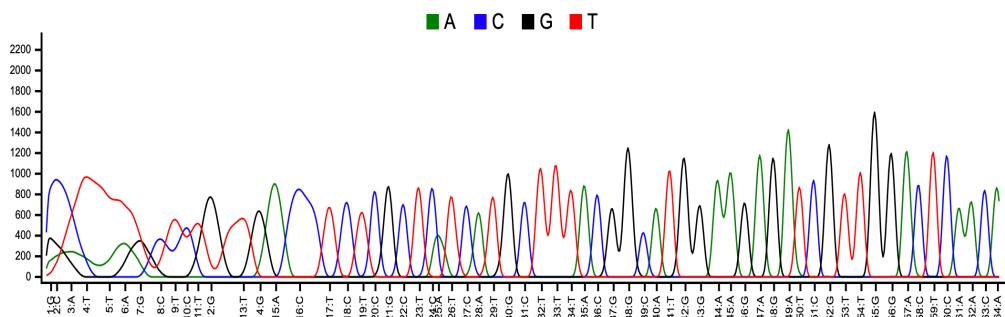
1. [シーケンスの確認・トリミング](#)
2. [アセンブル](#)
3. [シーケンスの統合](#)
4. [系統樹の作成](#)
5. [類似シーケンスの検索](#)



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>

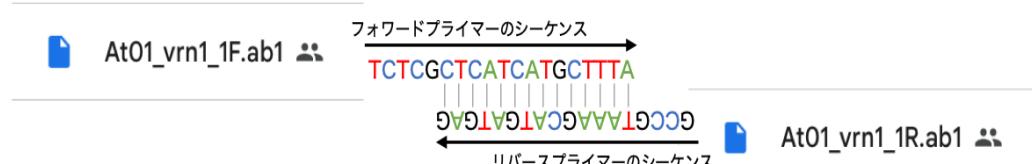
トリミングの手順

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

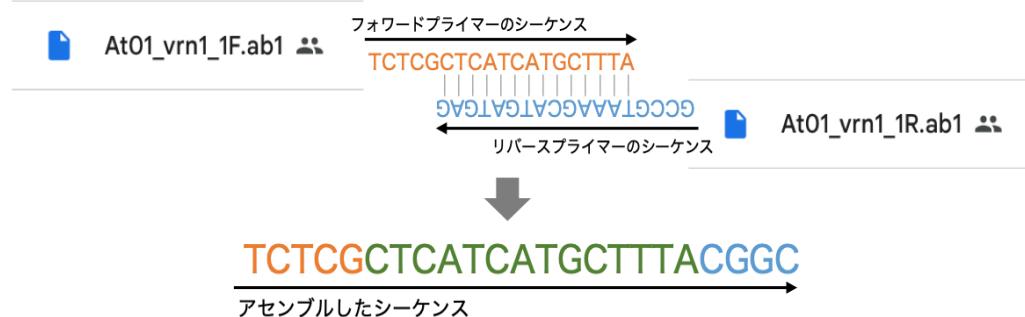


2 アヤンブル

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:

1. 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
 2. アセンブルが終わると、自動的にResultsページに移行する
 3. 「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
 4. 「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）が表示されています

PRABI-Doua
Pôle Rhône-Alpes de Bioinformatique Site Doua

CAP3 Sequence Assembly Program

Enter your sequences in FASTA format (no more than 50 kb):

```
>EMBOSS_001
TCCTGGCTCATCTGCTTACGGCGATGAAAGATGTTGGACTACAAA
AAAGATGCGCT
>EMBOSS_001R+TCCTGGCGGGAGTCCGAAAGTGTGAGAAAGTTTGGCTG
GTGGCGATGAT
>CAAAATAATGTTGCCACGCTACTCGAGAGCTCAGGAAGCACCAAC
```

SUBMIT | **CLEAR**

This form allows you to assemble a set of contiguous sequences (contigs) with the [CAP](#) program.

If you use CAP3 in any published work, please cite the following reference:
 Huang, X., and Madan, A. (1999) CAP3: A DNA sequence assembly program. *Genome Res.*, **9**, 868-877.
 For a more advanced use of CAP3, it is recommended to install the original software on your local computers.

1

2

3

4

CAP3 assembly:

Running ... done.
Results:

Contigs
Single sequences
Assembly details
Your sequence file

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-	TAATCGACGTGACCAACGACCTCACCTGAGCTGTATCTTGAGAGCGAG	
EMBOSS_001R+	TCACGCTGCTACCCAGACCTCACCTGAGCTGTATCTTGAGAGCGAG	
consensus		
TAATCGACGTGACCAACGACCTCACCTGAGCTGTATCTTGAGAGCGAG		

2.2 相補鎖交換

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

GeneScript
The Biology CRO

Sequence Manager

Format Conversion

- Combine FASTA
- EMBL to FASTA
- FASTA to EMBL
- EMBL Trans Extractor
- FASTA Trans Extractor
- Filter Protein
- GenBank to FASTA Extractor
- GenBank Trans Extractor
- Range Extractor DNA
- Range Extractor RNA
- Reverse Complement
- Three to One

Analysis

- Codon Position
- Codon Usage
- Dna Pattern
- DNA Pattern Find
- Fuzzy Search
- Fuzzy Search Protein
- Multi Rev Trans
- ORF Finder
- Parwises Align DNA
- Protein Conservation Protein
- Primer Stats
- Protein Isoelectric Point
- Protein Molecular Weight
- Protein Purity

Reverse Complement

Reverse Complement converts a DNA sequence into its reverse, complement, or reverse complement. The case of each input sequence character is maintained. You may want to work with the results of this analysis.

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

```
>Contig1 reverse complement
TCTCGCTCATCATGCGTTAACGGCATGGAAGAGATCTGGACTCAACAAAAGACGTGGCT
GGCTCGCTGCAGGGGACTCGCAGAACGTTGCTGAAGAAAGTTTGGCTGGCCGATGG
GCAAAATATTGTCGCGACGGTACTCGGAGCTGCAAGGAAACCAACACAAAAACCATG
GTCATTCGGAGAGGGCCGCCAGAACGACTCACAGGATGTTCTGATTCAGGCT
TTGGTCGGCTCATTCACAGGTACACGTGCACACACACAAAAAAACTACCTCTATAA
AGAAATATAAGACATTAGATACACTACGTTTTCTTAAGGAGGAGTAAAGAAAAGAA
GTTCTAGACAGCTGTGGTACATTTTCACTTTTCACTTTTGTAAATTGTATTGAGGTGGA
CAAGGCATCTACCTGCTGAGGAACTAGATGCTACCTAGAGGAGCTGGAGCAGAACTGG
GGAGCTGAAATCCACACGGTGGGCCAACAGCTGGCAAAACTCTACGATCTGAGGT
GAAAGAGCTGGTAGCTGTTCCAGAAGAGACCGCTGGAGCTCGGGGGGAGCACAGGA
GAGGGCTGCTCCCCAACAGGACGAGCCGGCAGAACATGTCACCTCACCGTGCAGGACAA
AGAGGCTGCTGGTAGCTGGAAGGAGCTGGCTGAGGACGTCGTTGACGACAGTGTGCA
CGCCATCAAAGGCTCTCCCTGGGAGCTCTCGCTGGCTGGCGGCTGACGCCCCAGGCC
CCTCGCTCATCACAGCTCACGGTCTGGTACGCTGATGAGCTGATGATTA
```

Please 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.

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る必要があります。

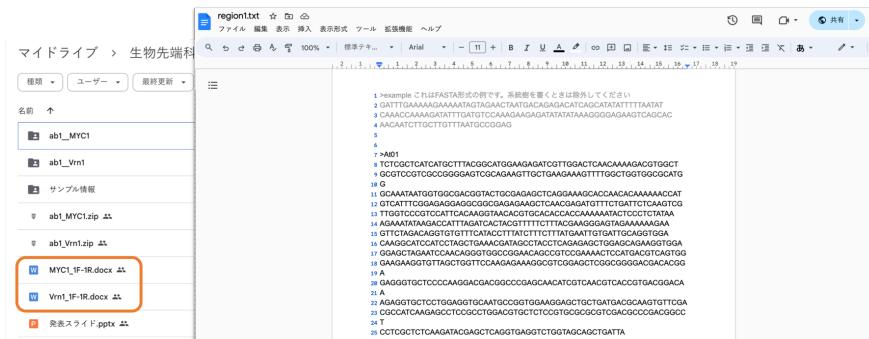
Google ドライブの共有フォルダ内に、各サンプルのアセンブル配列を集めるためのファイル（領域名.docx）を置いています。

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ポ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

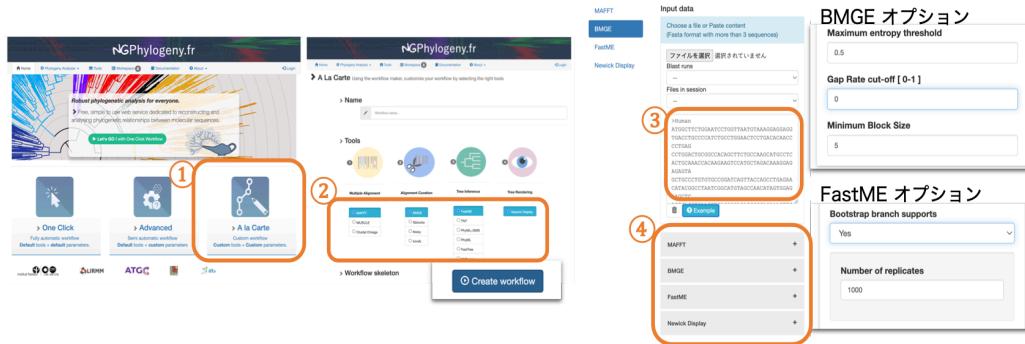


実習での系統樹作成の手順:

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）



NGPhylogenyのサーバーが混んでいる場合、数分～数時間程度解析がはじまらない場合があります。
代替手段として、[Google Colabの系統解析パイプライン](#)を使用することも可能です。

5. 類似シーケンスの検索

次世代シーケンサーや第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/
- *Aegilops tauschii* (タルホコムギ)
https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_002575655.2/
- *Aegilops sharonensis* (別のコムギ近縁野生種)
https://www.ncbi.nlm.nih.gov/datasets/genome/GCA_904067115.1/

イネと近縁種の公開ゲノム配列:

- *Oryza sativa* (AAゲノム)
 - Japonica group
 - 品種: Nipponbare https://www.ncbi.nlm.nih.gov/datasets/genome/GCA_003865235.1/
 - 品種: Kitaake https://www.ncbi.nlm.nih.gov/datasets/genome/GCA_009797565.1/
 - Indica Group
 - 品種: 93-11 https://www.ncbi.nlm.nih.gov/datasets/genome/GCA_003865215.1/
 - 品種: Tetep https://www.ncbi.nlm.nih.gov/datasets/genome/GCA_004348155.2/
 - 品種: Zhenshan 97 https://www.ncbi.nlm.nih.gov/datasets/genome/GCA_001623345.3/

- 近縁種 (Oryza sativaと同じAAゲノムをもつ種)
 - Oryza rufipogon https://www.ncbi.nlm.nih.gov/datasets/genome/GCA_046718375.1/
 - Oryza glaberrima https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_000147395.1/
 - Oryza barthii https://www.ncbi.nlm.nih.gov/datasets/genome/GCA_000182155.4/
 - Oryza longistaminata https://www.ncbi.nlm.nih.gov/datasets/genome/GCA_009805545.1/
 - Oryza meridionalis https://www.ncbi.nlm.nih.gov/datasets/genome/GCA_000338895.3/
 - Oryza glumipatula https://www.ncbi.nlm.nih.gov/datasets/genome/GCA_000576495.2/
- 近縁種 (別のゲノムをもつ種)
 - Oryza punctata [BB・BBCC] https://www.ncbi.nlm.nih.gov/datasets/genome/GCA_000573905.2/
 - Oryza officinalis [CC] https://www.ncbi.nlm.nih.gov/datasets/genome/GCA_008326285.1/
 - Oryza granulata complex [GG] https://www.ncbi.nlm.nih.gov/datasets/genome/GCA_005223365.2/

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"の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

The screenshot shows the NCBI Genome assembly IWGSC CS RefSeq v2.1 page. A red circle labeled ① highlights the 'View annotated genes' button. A red box labeled ② highlights the 'BLAST the reference genome' button. A red box labeled ③ highlights the 'Enter Query Sequence' input field containing a DNA sequence. A red box labeled ④ highlights the 'BLAST' search button. A red box labeled ⑤ highlights the 'Sequences producing significant alignments' table. A red box labeled ⑥ highlights the 'GenBank' link for the top result. A red box labeled ⑦ highlights the 'FASTA' link for the GenBank page.