岐阜大学 3年後期 ゲノム生物学

ゲノムアセンブリ法とゲノム解析に関して

*本講義で紹介するソフトウェアのバージョンは2020年2月4日現在のものです。
分からない点があれば、以下サイトを参照ください。

https://qiita.com/danryo_official

講義内容

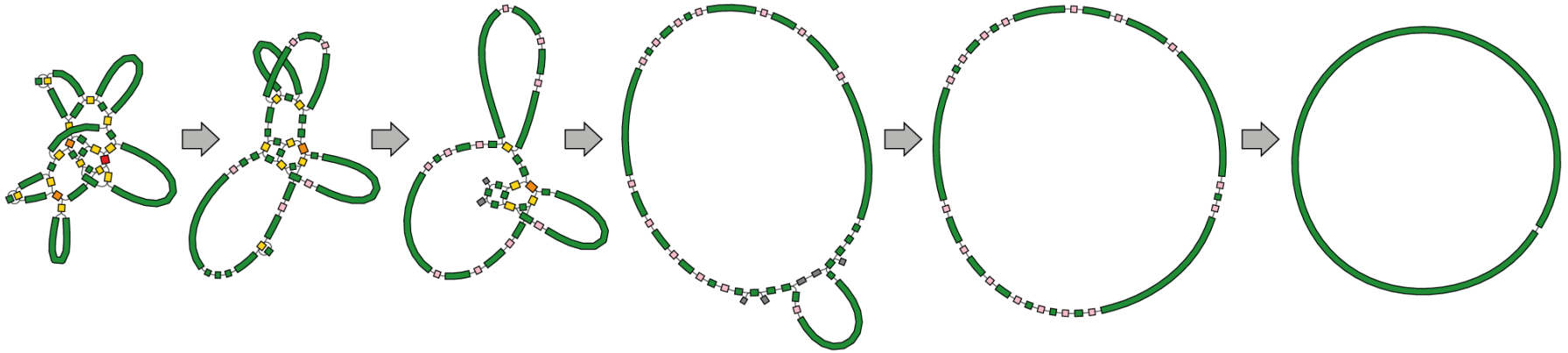
1. 初めに
2. NGSのおさらい (ペアリードシーケンス, メイトペアーシーケンス)
3. de brujin graph法によるオイラーパス問題
4. ゲノムアセンブリをしてみる
5. 完成したゲノムデータの活かし方 -アノテーション-
6. 終わりに

ここで紹介する内容はゲノムアセンブリ法の1例です。
さらに深く学びたい人は**ググって**ください。

*本講義で使用する全てのソフトウェアとコマンドは以下サイトに記しました。
https://qiita.com/danryo_official

はじめに

本講義のテーマは次世代シーケンサーから得られたリード (配列) をコンピュータ内で組み立てる***de novo*アセンブリ**を行うこと、および得られたゲノムデータ、もしくはドラフトゲノムデータを**どのように活用するかを考える**というものです。



Wick RR, Judd LM, Gorrie CL, Holt KE (2017) Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. PLoS Comput Biol 13(6): e1005595. https://doi.org/10.1371/journal.pcbi.1005595

ゲノムアセンブリの歴史

1995年 1.83 Mb の真正細菌 *Haemophilus influenza*のゲノム決定
(生物としては初、ウイルスはこの1年前にシーケンスされた)

1997年 4.6 Mbの大腸菌 *Esherichia coli*のゲノム決定
同年 12.1 Mbの出芽酵母 *Saccharomyces cerevisiae*のゲノム決定

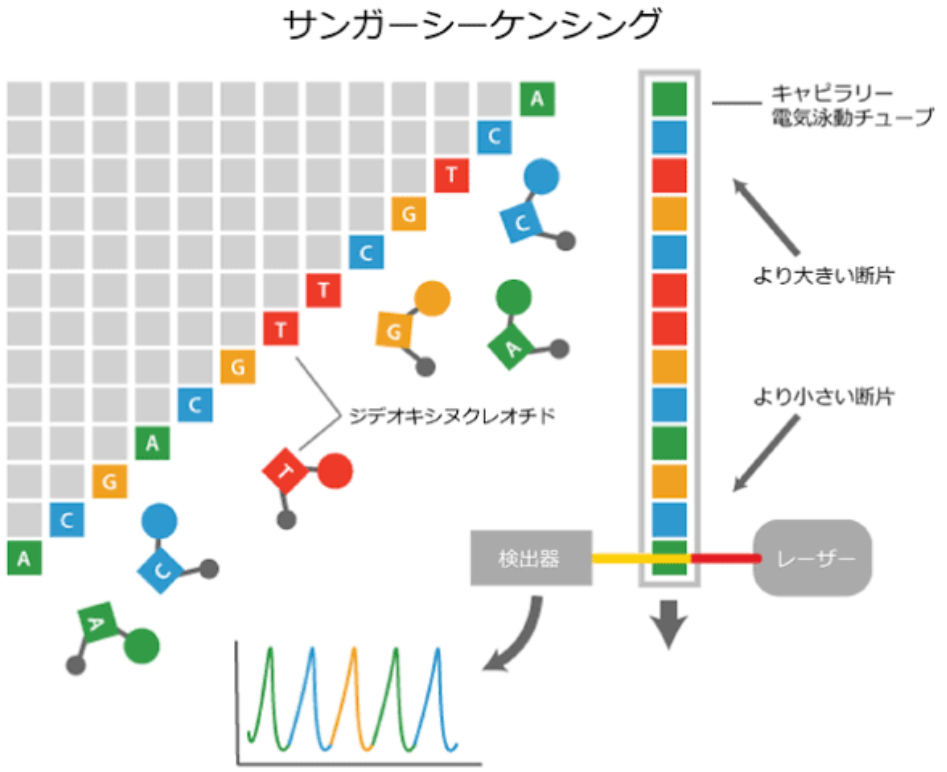
2003年 ヒト *Homo sapiens*のゲノム決定発表

ゲノムシーケンス手法の変化

以前はゲノムシーケンスには**サンガー法**が用いられていた。

つまり、デオキシリボヌクレオチドに放射性標識しておき、ポリアクリルアミド電気泳動により断片長に応じて分離して、オートラジオグラフィーにより検出していた。

その後、蛍光標識やキャピラリー電気泳動といった技術を取り込むことで飛躍的に発展した。



https://www.cosmobio.co.jp/support/technology/a/next-generation-sequencing-introduction-apb.asp

ゲノムシーケンス手法の変化

サンガー法初期

- 放射線の影響で1日に読める量が少ない。

サンガー法後期

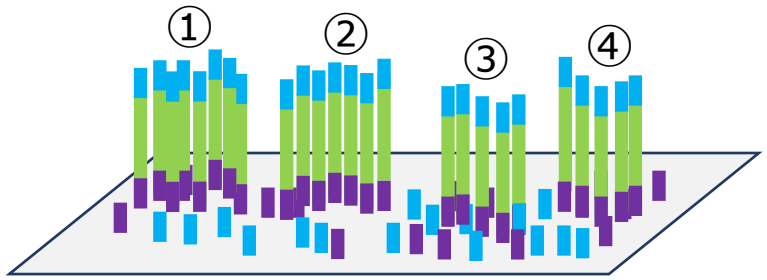
- 蛍光物質になり、スピードUP
- ただ、それでも時間がかかりすぎる。
- ヒト: 10年、バクテリア: 数年

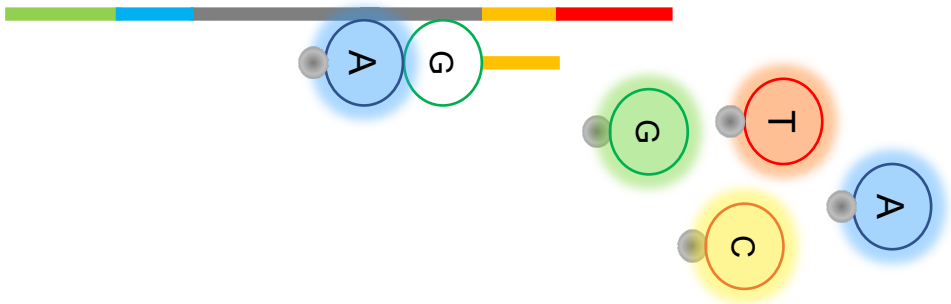
ゲノムシーケンス手法の変化

2004年、ケンブリッジ大学化学科初ベンチャー企業Solexaが現在のSBS (Sequencing By Synthesis) のプロトタイプを発明。

1日に読める限界のリード数が爆発的に上昇。

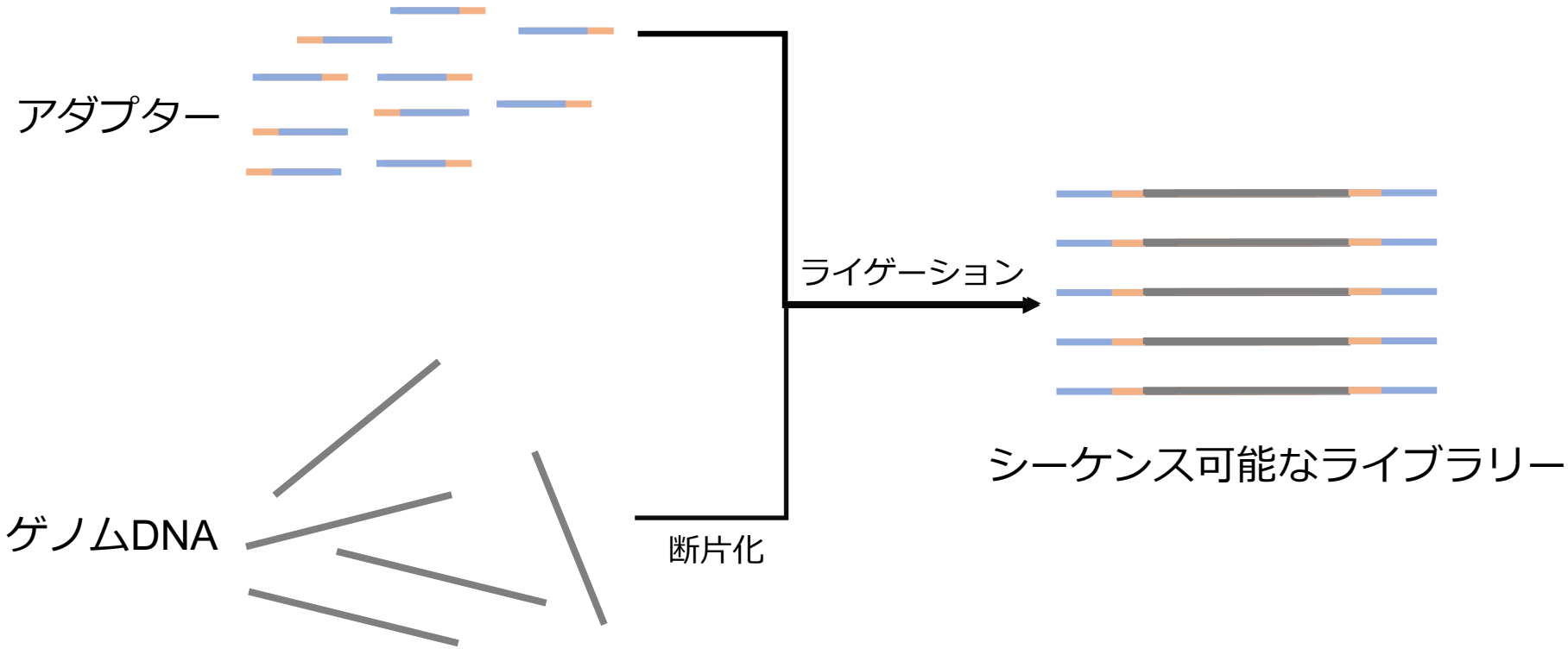
2005年にはSBSを用いた**バクテリオファージphiX-174**のゲノム決定が行われた。





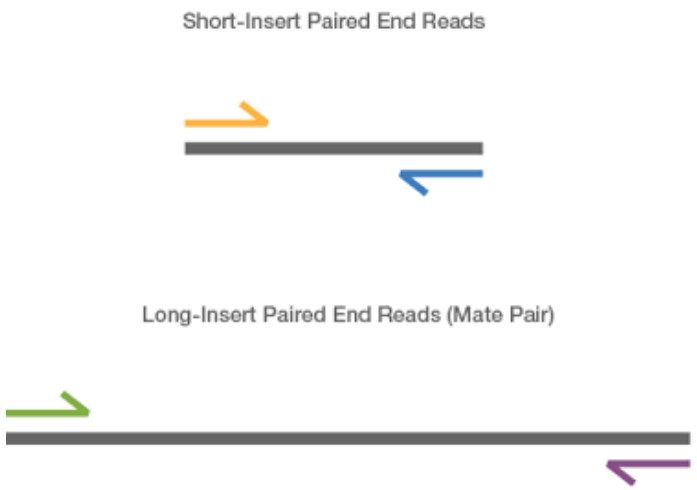
SBSによるゲノムシーケンスについて

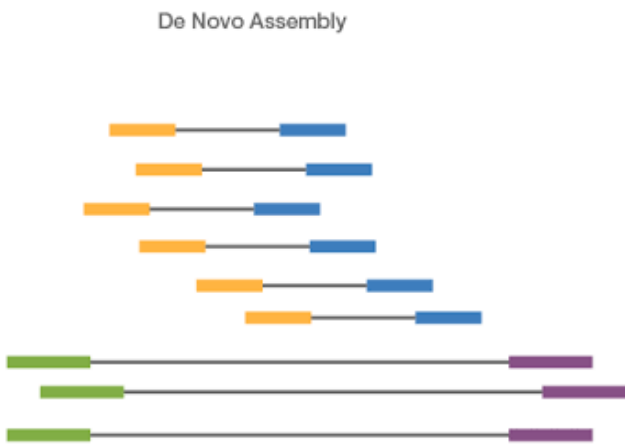
SBSによるアウトプット配列は短いのが基本 (数十 bp~数百 bp)
よって、目的サイズに分断してからゲノムシーケンスを行う。(制限酵素、超音波など)



ゲノムシーケンスでよく用いられる2手法

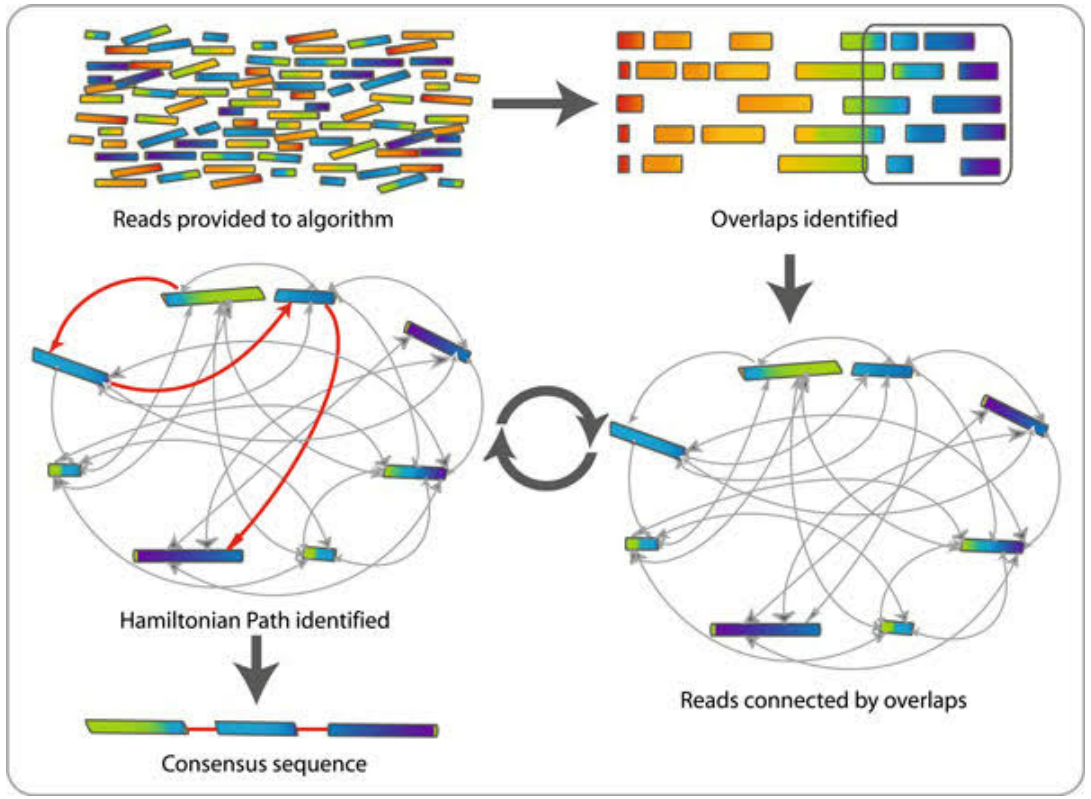
ペアエンドシーケンス:同じ DNA 分子配列を二回読んだペアのシーケンス
メイトペアシーケンス:大きな分子の両端から二つのタグのみをシーケンス





ゲノムを構築する

コンピュータ内ではどのようなアルゴリズムでゲノムを構築しているのか。



2000年代に使われていた方法から説明する。

Overlap-Layout-Concensus法
各リードを頂点(ノード)として、k個の共通連続塩基がある頂点同士を辺(エッジ)で結んだグラフを作成し、全ての頂点を通るパスを探索 (ハミルトンパス問題という)

要は端が同じ配列を探して、すべてくっつくように並べ換える。

ゲノムを構築する

Overlap-Layout-Concensus法の問題点

1. 計算量が膨大。計算量はリード数の2乗に比例する。
2. 短い配列には不適。2000年代で使われた理由はサンガー法が使われていたから。

De Brujin Graph法へのシフト

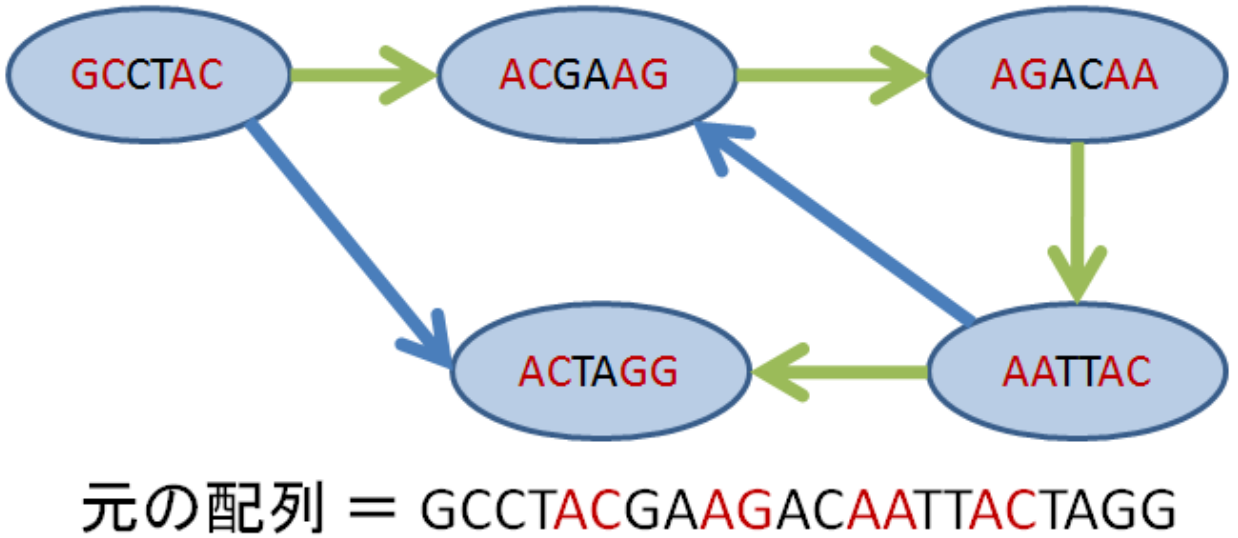
1. 普通のコンピュータでも計算できるようになる。
2. 同時に計算そのものを高速化できる。

ゲノムを構築する

De Brujin Graph法

リードを1塩基ずつずらしたK個の連続塩基からなるk-merグラフを各リードごとに作成する。

全リードの完全一致ノードをマージすることで「de Bruijnグラフ」を作成し、全ての辺を通るパスを探索 (オイラーパス問題という)



https://hoxo-m.hatenablog.com/entry/20100930/p1
http://www.iu.a.u-tokyo.ac.jp/~kadota/20111015_kadota.pdf

ゲノムを構築を体験してみる

多分さっぱり分からないと思うので、実際に手を動かしてみましょう。

ゲノムを構築を体験してみる

ハミルトンパス問題としてゲノムアセンブル問題を解いてみる

ハミルトンパス問題でのルール
3-mer以上端に共通配列があればくっつけるとする。(通常K-mer (Kは定数))

CGTAGCG

TGACGAT

ATGCCGT

GCGATGA

ゲノムを構築を体験してみる

ハミルトンパス問題としてゲノムアセンブル問題を解いてみる

ハミルトンパス問題でのルール
3-mer以上端に共通配列があればくっつけるとする。(通常K-mer (Kは定数))

CGTAGCG

TGACGAT

ATGCCGT

GCGATGA

ゲノムを構築を体験してみる

ハミルトンパス問題としてゲノムアセンブル問題を解いてみる

回答

ATGCCGT
CGTAGCG
GCGATGA
TGACGAT
ATGCCGTAGCGATGACGAT

ゲノムを構築を体験してみる

オイラーパス問題としてゲノムアセンブル問題を解いてみる

オイラーパス問題でのルール
今回はK=3としてリードを3塩基ずつにする。分けたリードから共通して通るパスを見つけて、最も長いパスを探す。

CGTAGCG

TGACGAT

ATGCCGT

GCGATGA

ゲノムを構築を体験してみる

オイラーパス問題としてゲノムアセンブル問題を解いてみる

オイラーパス問題でのルール
今回はK=3としてリードを3塩基ずつにする。分けたリードから共通して通るパスを見つけて、最も長いパスを探す。2-mer分共通領域があれば、くっつけるものとする。

CGTAGCG なら CGT GTA TAG AGC GCG

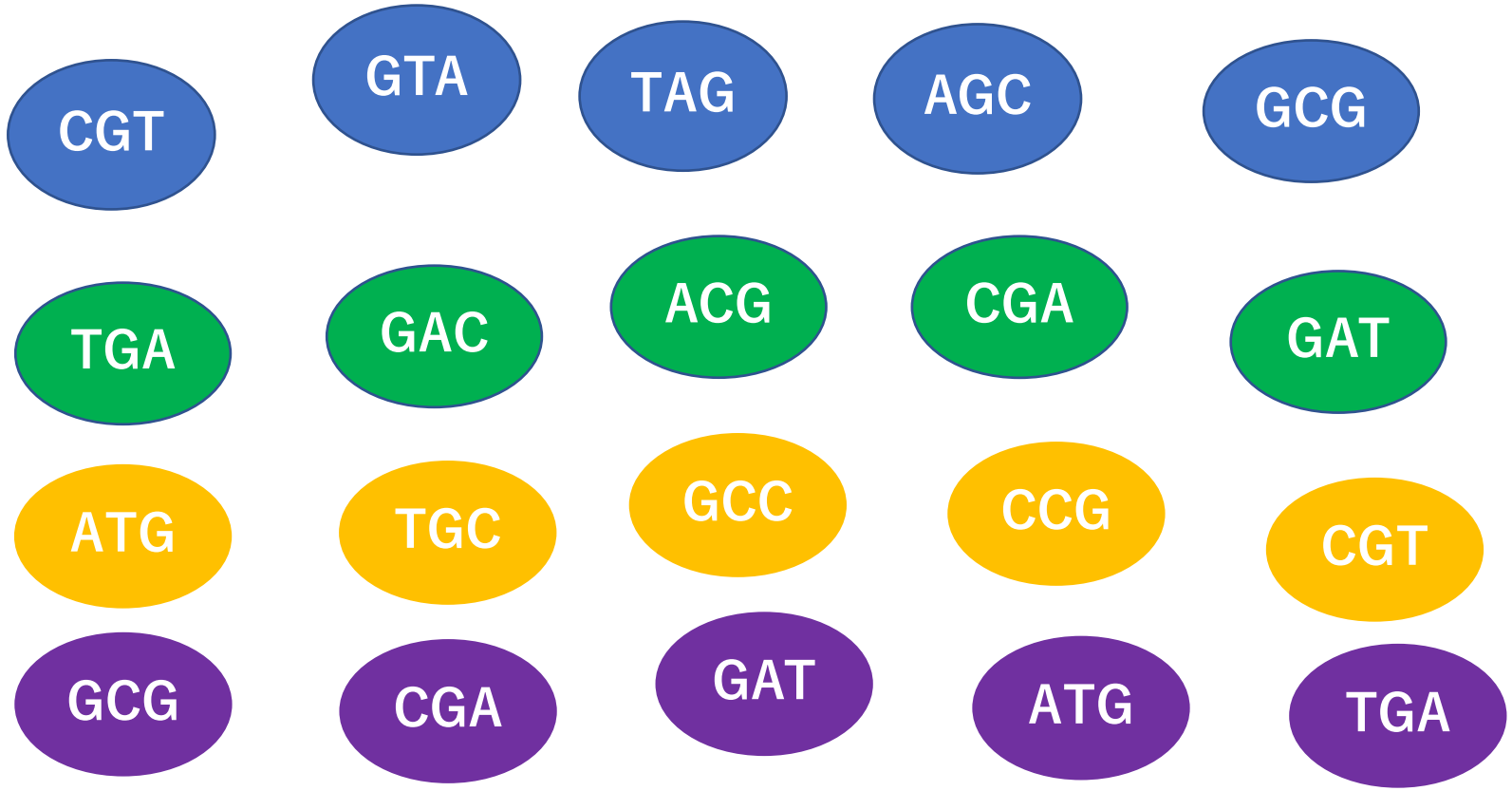
TGACGAT なら TGA GAC ACG CGA GAT

ATGCCGT なら ATG TGC GCC CCG CGT

GCGATGA なら GCG CGA GAT ATG TGA

ゲノムを構築を体験してみる

オイラーパス問題としてゲノムアセンブル問題を解いてみる

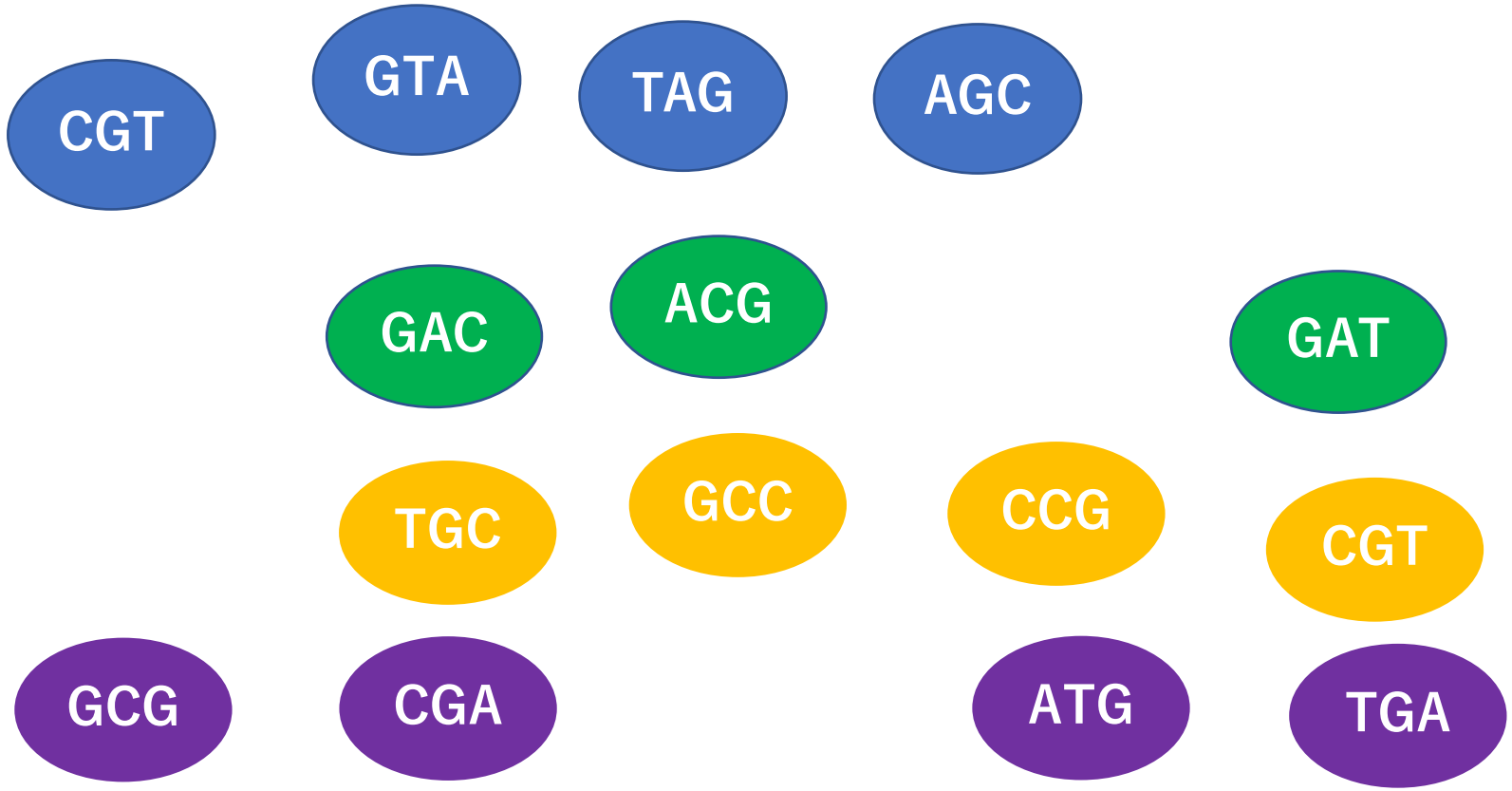


各因子として
分けて表示すると
こんな感じになる。

共通項をまとめて、
各因子を繋いでみる。

ゲノムを構築を体験してみる

オイラーパス問題としてゲノムアセンブル問題を解いてみる

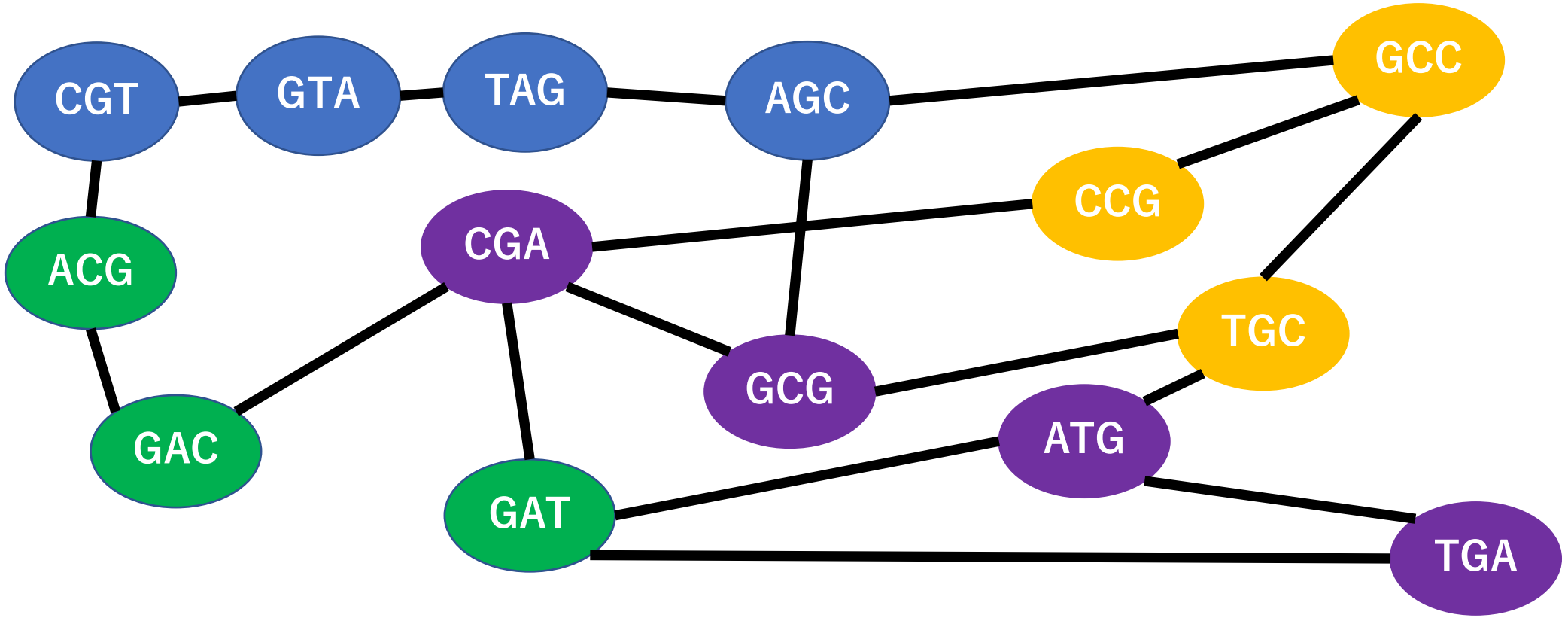


各因子として
分けて表示すると
こんな感じになる。

共通項をまとめて、
各因子を繋いでみる。

ゲノムを構築を体験してみる

オイラーパス問題としてゲノムアセンブル問題を解いてみる



ゲノムを構築を体験してみる

オイラーパス問題としてゲノムアセンブル問題を解いてみる

回答

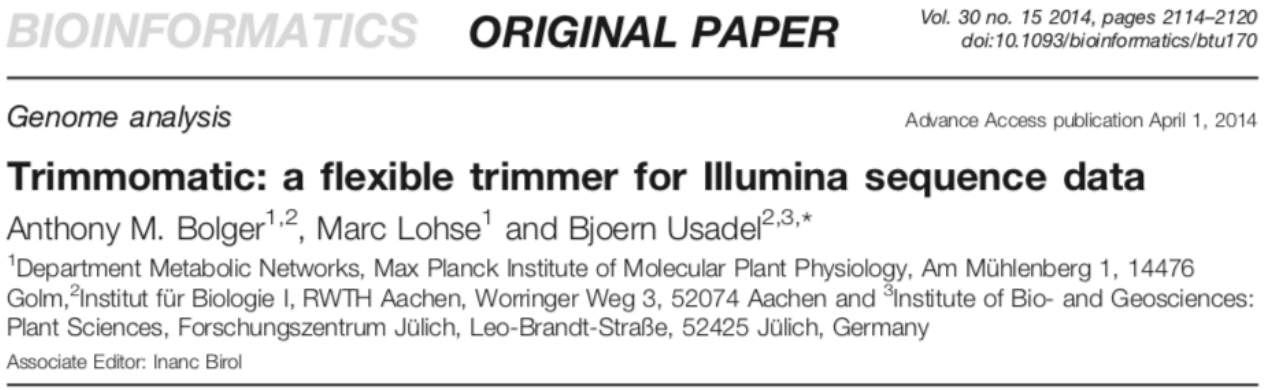
最も長く通るパスを考えると
ATGCCGTAGCGATGACGAT
が答えになる。

手計算だとDe Brujin Graphの作成は非常にめんどくさいが、コンピュータ上では都合のいいアルゴリズムが存在する。

ゲノムアセンブリ前のクオリティートリム

16S rRNA アンプリコンシーケンスと同様に、ゲノムシーケンスの際もクオリティートリミングが必要。

当方はTrimmomaticによるクオリティートリムとアダプタートリムを行っている。
Trimmomaticはスライディングウィンドウ法を利用したクオリティートリミングツールである。



BIOINFORMATICS ORIGINAL PAPER

Vol. 30 no. 15 2014, pages 2114–2120
doi:10.1093/bioinformatics/btu170

Genome analysis

Advance Access publication April 1, 2014

Trimmomatic: a flexible trimmer for Illumina sequence data

Anthony M. Bolger[1,2], Marc Lohse[1] and Bjoern Usadel[2,3,*]

[1]Department Metabolic Networks, Max Planck Institute of Molecular Plant Physiology, Am Mühlenberg 1, 14476 Golm,[2]Institut für Biologie I, RWTH Aachen, Worringer Weg 3, 52074 Aachen and [3]Institute of Bio- and Geosciences: Plant Sciences, Forschungszentrum Jülich, Leo-Brandt-Straße, 52425 Jülich, Germany

Associate Editor: Inanc Birol

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina Sequence Data. *Bioinformatics*, btu170.

Sliding Window法によるクオリティートリム

Sliding Window法は仮想の枠を決めて、この枠の中の平均値がThresholdよりも上かどうかを考える計算方法である。

例えばATGCATという配列があるとして、仮想の枠を4塩基とする。
すると、得られる配列はATGC, TGCA, GCATの3種類。

そして各枠において、クオリティーの平均値を算出してThreshold分あるかどうかを判定する。

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina Sequence Data. *Bioinformatics*, btu170.

Sliding Window法によるクオリティートリム

```
@1¥1
AATGATCGTAGCGATGCAAGCTAGCCCGATGCCCGATCGCATCG
+
????????????????????????????????????????1234
@2¥1
AATGATCGTAGCGATGCAAGCTAGCCCGATGCCCGATCGCATCG
+
????????????????1234????????????????????????
@3¥1
AATGATCGTAGCGATGCAAGCTAGCCCGATGCCCGATCGCATCG
+
1234????????????????????????????????????????
```

例として左記配列を用意した。
枠を4塩基として存在するクオリティー平均値は以下の通り。

????: 30
???1: 26.5
??12: 23.25
?123: 20.25
1234: 17.5

Sliding Window法によるクオリティートリム

```
@1¥1
AATGATCGTAGCGATGCAAGCTAGCCCGATGCCCGATCGCATCG
+
????????????????????????????????????????1234
@2¥1
AATGATCGTAGCGATGCAAGCTAGCCCGATGCCCGATCGCATCG
+
????????????????1234????????????????????
@3¥1
AATGATCGTAGCGATGCAAGCTAGCCCGATGCCCGATCGCATCG
+
1234????????????????????????????????????
```

左記配列を平均クオリティー20以下でトリミングしてみる。

なお、Trimmomaticでは Thresholdを下回ると枠を含む全ての3’側の塩基が排除される。

Sliding Window法によるクオリティートリム

```
@1¥1
AATGATCGTAGCGATGCAAGCTAGCCCGATGCCCGATCGC
+
????????????????????????????????????????
@2¥1
AATGATCGTAGCGATGCA
+
??????????????????
```

左記配列を平均クオリティー20以下でトリミングしてみる。

なお、Trimmomaticでは Thresholdを下回ると枠を含む全ての3’側の塩基が排除される。

Sliding Window法によるクオリティートリム

```
@1¥1
AATGATCGTAGCGATGCAAGCTAGCCCGATGCCCGATCGC
+
????????????????????????????????????????
@2¥1
AATGATCGTAGCGATGCA
+
??????????????????
```

クオリティートリム後は変な配列がないかをFastQCというソフトウェアで確認した方が良い。

これをクリアした配列を次のアセンブリに利用する。

ゲノムアセンブリ

De Brujin Graph法を利用した有名なソフトウェアにSPAdesがある。

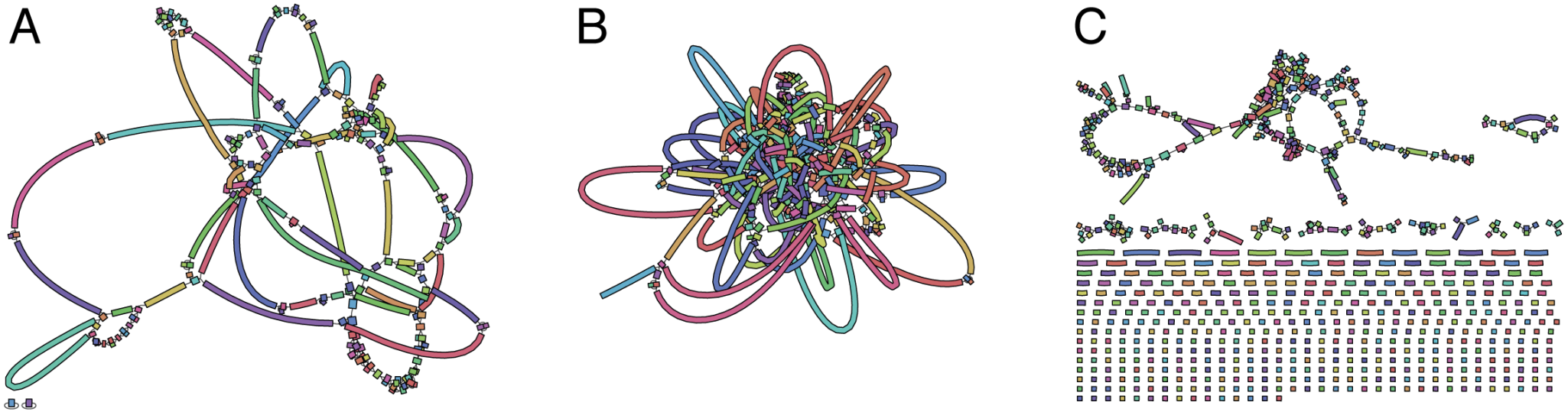
データをPolishする際、様々なアルゴリズムを利用するが、ベースとなる計算はDe Brujin Graph法である。

SPAdesはProkaryotic cellのゲノムを専門とするゲノムアセンブラになる。Eukaryotic cellをやる場合は別のアセンブラを利用する必要がある。

＊ 当方は普段はUnicyclerというSPAdes依存のアセンブラを利用している。

Bankevich A, Nurk S, Antipov D, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol*. 2012;19(5):455–477. doi:10.1089/cmb.2012.0021

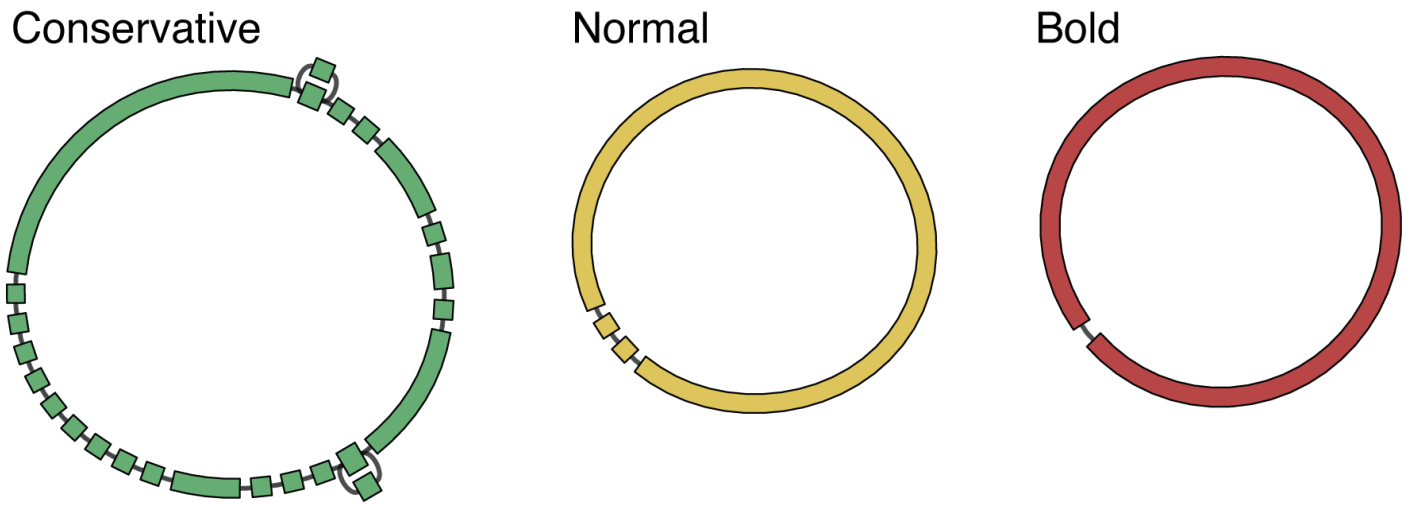
アセンブリの段階



シーケンスデータの状態は用意したサンプルの状態などによって異なる。うまくくっつけばAのような状態となり、ゲノム合成まで後一歩という感じになる。シーケンスの調子が悪いとDe Brujin Graphがうまく構築できずCのようになる。

https://github.com/rrwick/Unicycler

アセンブリの実際



同じデータをわざとくっつきにくいパラメータにして解析した場合の模式図。
このようにコンピュータがコンティグを作る際に悩んだ場合はロングリードを入れて再計算するか、リシーケンスデータを混ぜると解決する。

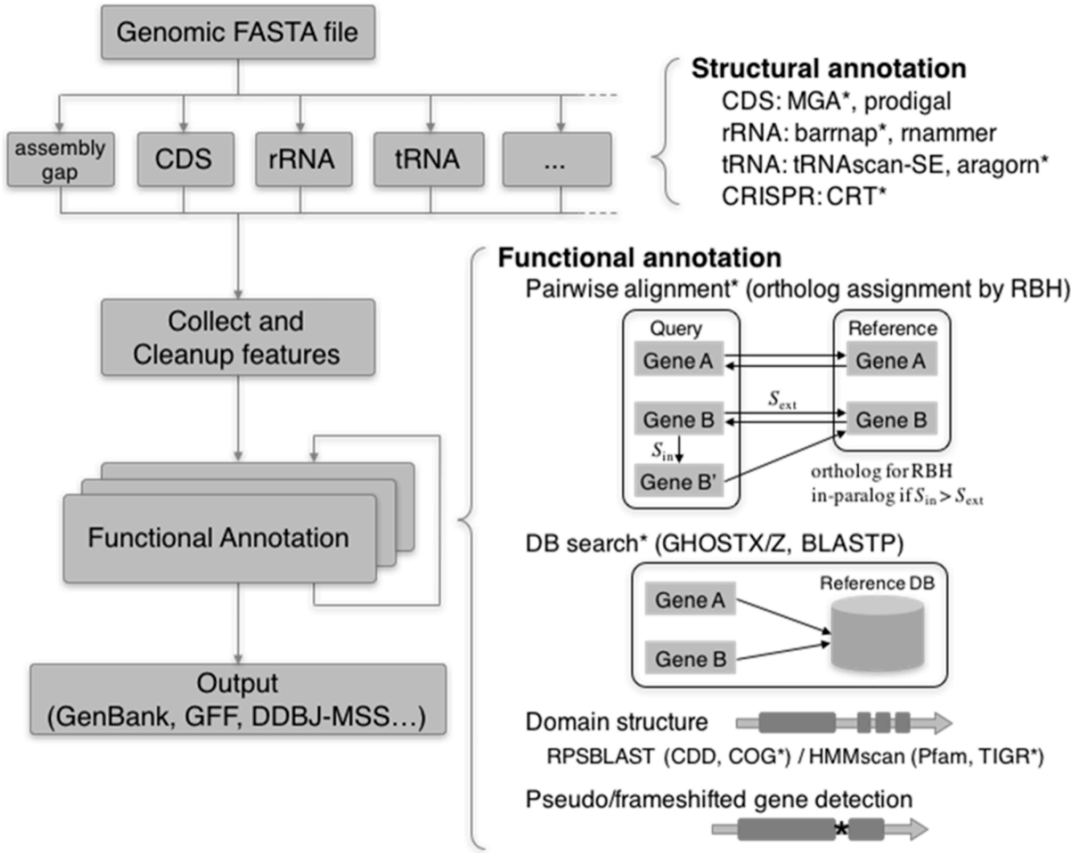
https://github.com/rrwick/Unicycler

アノテーション

ゲノムアセンブリをしても、所詮データは数Mbの塩基配列情報である。これを有効活用するにはアセンブルしたデータをアノテーションし、遺伝的形質を探る必要がある。

基本的にはBLASTの原理を利用して、Homologyによる推定を行っている。

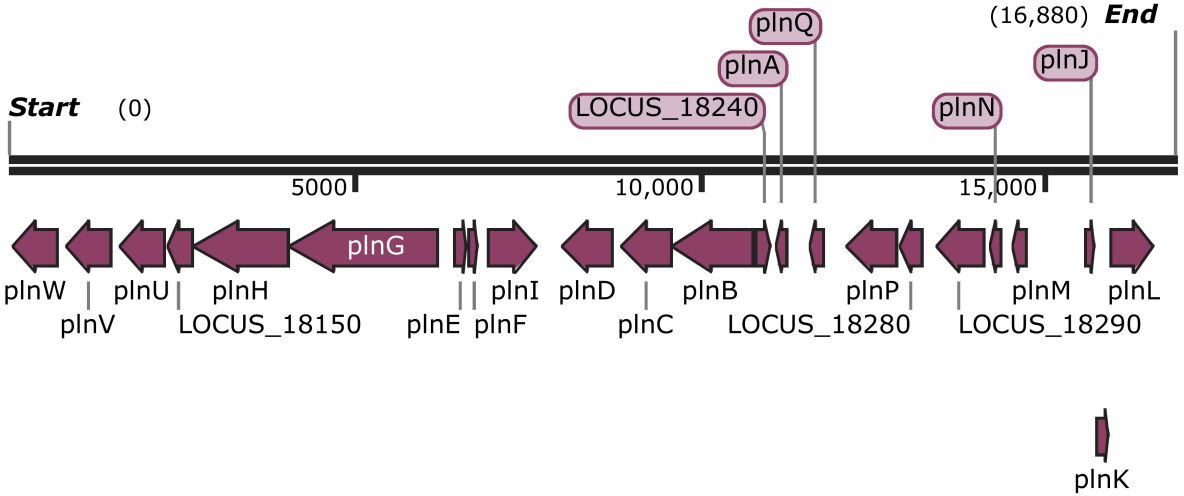
ソフトウェアとしてはProkkaやDFASTがある。



https://dfast.nig.ac.jp

アノテーション

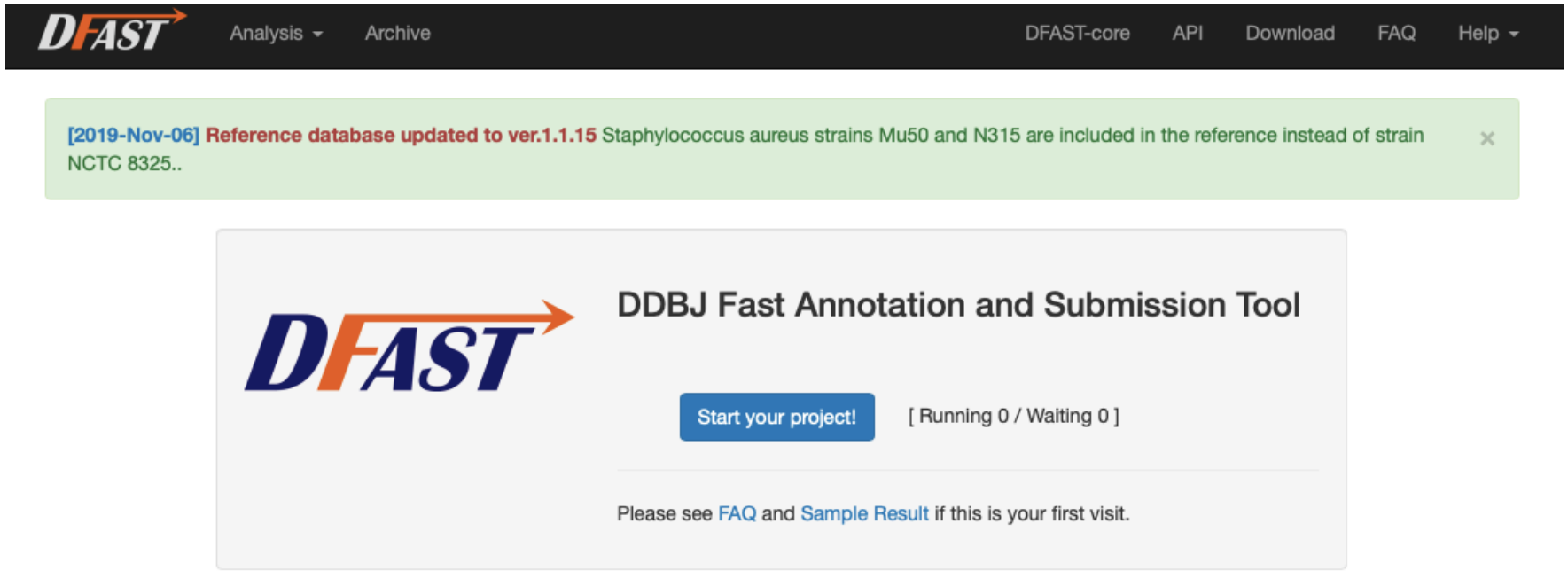
乳酸菌*Lactobacillus plantarum*のゲノムアセンブリを行い、これをDFASTによりアノテーションした。アノテーション結果よりバクテリオシン関連遺伝子クラスターを推定し、ゲノム配列より抽出しマップとした。



unpublished

DFAST

DFASTは遺伝研が発明したオンラインアノテーションパイプラインである。Graphical User Interfaceで構成されており、クリックを行うだけでアノテーションが完了する。



https://dfast.nig.ac.jp

終わりに

冒頭と繰り返しになりますが、本講義で紹介したゲノムアセンブリ法は数ある例のうちの一つです。現在もゲノムアセンブリパイプラインは世界各国で検討されており、発展している分野になります。

また、ゲノムデータ解析は非常に複雑なため、今回は省きました。ただ、ゲノムを知るという観点において「**比較**」は欠かすことのできない部分で、解析する際に最も面白いところになります。

今後ゲノム解析をする機会があれば、さらに面白い解析をして教えてください。