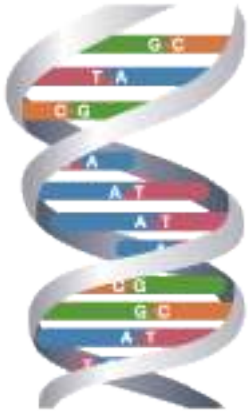
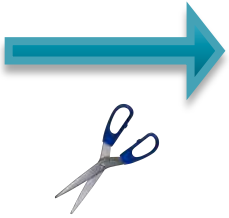
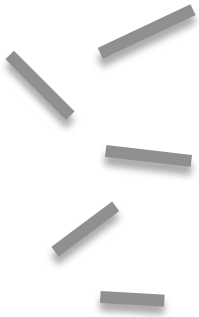
# NGSって何をやってくれる機械なの？







細分化

数100base分の塩基配列を
文字列で出力

×

大量の本数

Seq1 : ACTGANCCTGCTTCTGGTGTGGTTGATATT…
Seq2 : GATGTNCCCATCTGAATGCAATGAAGAAAA…
Seq3 : CTACGNCCAGCAGCAGTGGGGAATTTTCCG…
Seq4 : CTACGNCTCGCAGCAGTGGGGAATCTTGGA…
⋮

# NGS解析ってなに？

**大量の配列データを高速で取得するのが目的**

計算によってこねくり回し、目的のデータを得る

**大量データのため、計算には高速化が求められる**

計算の正確性と速度がちょうど良いアルゴリズムが選ばれる

**大量のデータの計算のため、メモリやCPUの使用量が少ないものが求められる**

ビックデータ解析に近い

# NGS解析ってどうやるの？

## インフォの知識有

方法の調査と検討※1
↓
解析
↓
バイオ独自フォーマットの調査
↓
結果表示

## バイオインフォの知識有

（方法の調査と検討※1）
↓
解析
↓
結果表示
↓
結果の判断

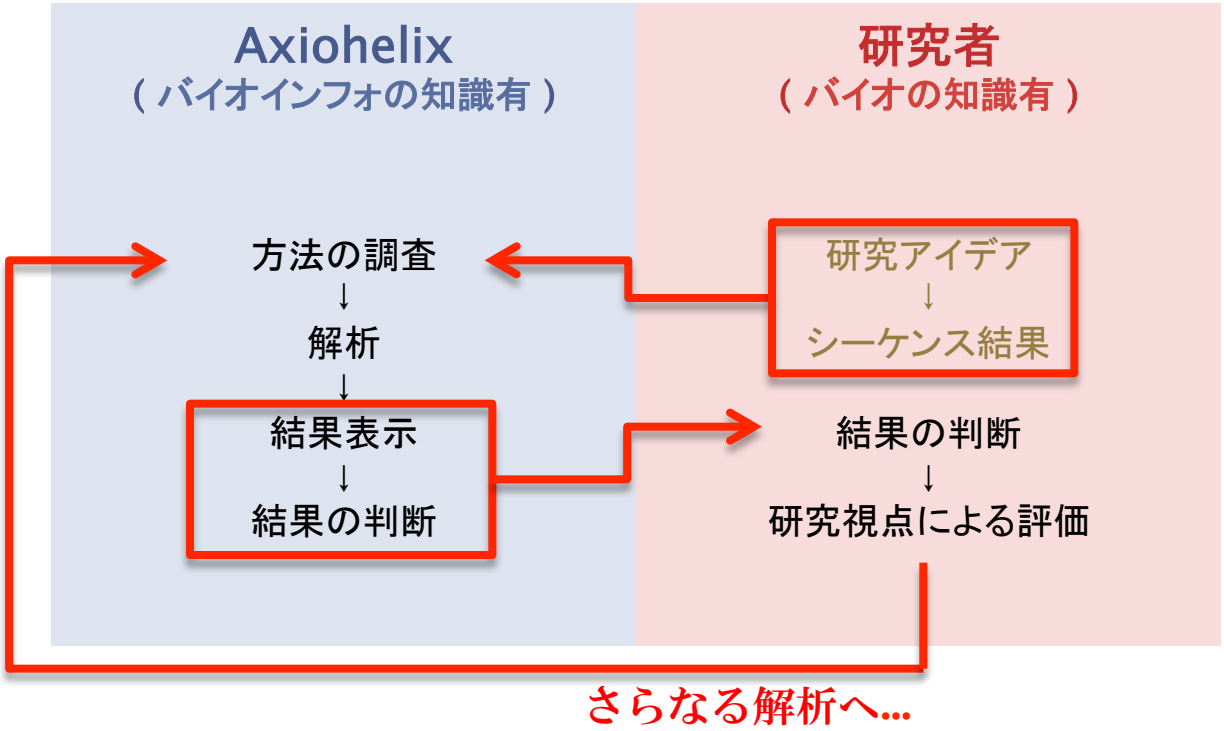
## バイオの知識有

インフォマティクスの習得
↓
方法の調査と検討※1
↓
解析
↓
バイオ独自フォーマットの調査
↓
結果表示
↓
結果の判断
↓
研究視点による評価

マウス操作でできるソフトが開発されている
スタンダードではない解析はできない

※1：バイオインフォマティクス独自の解析方法があり、経験により調査時間に差がある。また、方法の選択の幅や最適なものの選出ができるかは経験が大きい。

# PictBioのサービスは…




## 特徴

- 解析作業時間で契約することにより、カスタマイズ性の高い自由な解析
- 結果の判断により随時レポートし、余計な解析をカット
- 独自な解析アイデアであっても実現可能

# 短い配列から得られるものは？

DNA配列が既知の生物 (単一)

取得した各配列が
どこから取ってきたものか
計算によって求めることができる

Seq1 : ACTGANCCTGCTTCTGGTGTGGTTGATATT…
Seq2 : GATGTNCCCATCTGAATGCAATGAAGAAAA…
Seq3 : CTACGNCCAGCAGCAGTGGGGAATTTTCCG…
Seq4 : CTACGNCTCGCAGCAGTGGGGAATCTTGGA…
⋮

既知DNA塩基配列マップ

……..AGCAAAAGCAGGGGAAAATAAAAACAACCAAAATGAAGGCAAAACTAC……..

ブレイク時間に見る程度の内容

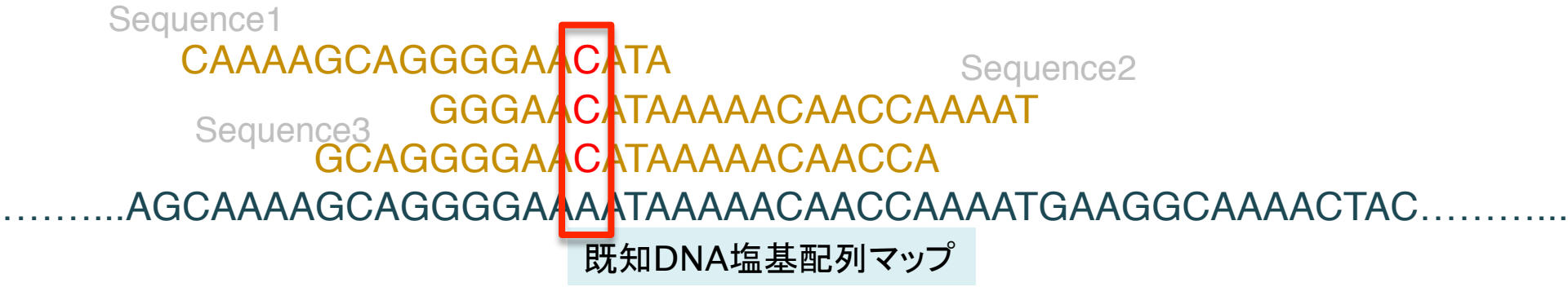
以下の考えのもとに短い配列のマッピングが正しいとの考え
塩基は4種類 -> 100baseのある1つの並び順になる確率は
1 / 4の100乗 = 1 / 1.60694E+60
1Mの長さのDNA配列中にこの並び順が出現する確率は
4の100乗 / 4の1M乗 ← 計算ソフトでは計算できないくらい低確率
よってこの短い配列がこの位置から取得した配列である確率はかなり高い

とは言っても出現しやすい配列が生物にはあります
柔軟性が大事

# 由来がわかったら？

## 続 DNA配列が既知の生物 (単一)

配列の位置がわかるとできること
① 既知配列と取得配列の比較
② どのくらいデータが取得できたか




取得配列ではこの位置は
A が C になっており ( ① )、
3配列のデータが取れた ( ② )

有名な解析
- 変異解析
- 発現解析 ( RNA )
- ChIP-Seq
- メチル化解析

# DNA配列がわかってない場合は？

DNA配列が未知の生物 (単一)

元のDNA配列を
計算によって求めることができる

Sequence1
CAAAAGCAGGGGAACATA

Sequence2
GGGAACATAAAAACAACCAAAAT

Sequence3
GCAGGGGAACATAAAAACAACCA

例えば、配列が同じ領域をくっつけて…

CAAAAGCAGGGGAAAATAAAAACAACCAAAAT

有名なアルゴリズム

**OLC ( Overlap-Layout-Consensus )**
重なった領域をマージして伸ばしていく

**De Bruijn graph**
数base単位で領域をマージし、ネットワークグラフを作成

有名な解析
- De novo
- De novo RNA

# どれくらいの結果が得られるのか？

## 配列のデータが多ければ多いほど得られる

DNA全体を見た場合、人が確認できない程膨大

## DNA全体を見ても部分的に結果が得られないところがある

シーケンサーが苦手、計算アルゴリズムが苦手な配列がある

## 配列取得時のエラーが含まれる

複数データを比較するなど、実験デザインによりある程度解決

既存データとの比較、クオリティ値等によるフィルタリング
ターゲット遺伝子の把握 etc. 絞込みが必要

単純なシーケンスではデータが得られないことも
実験デザインが重要