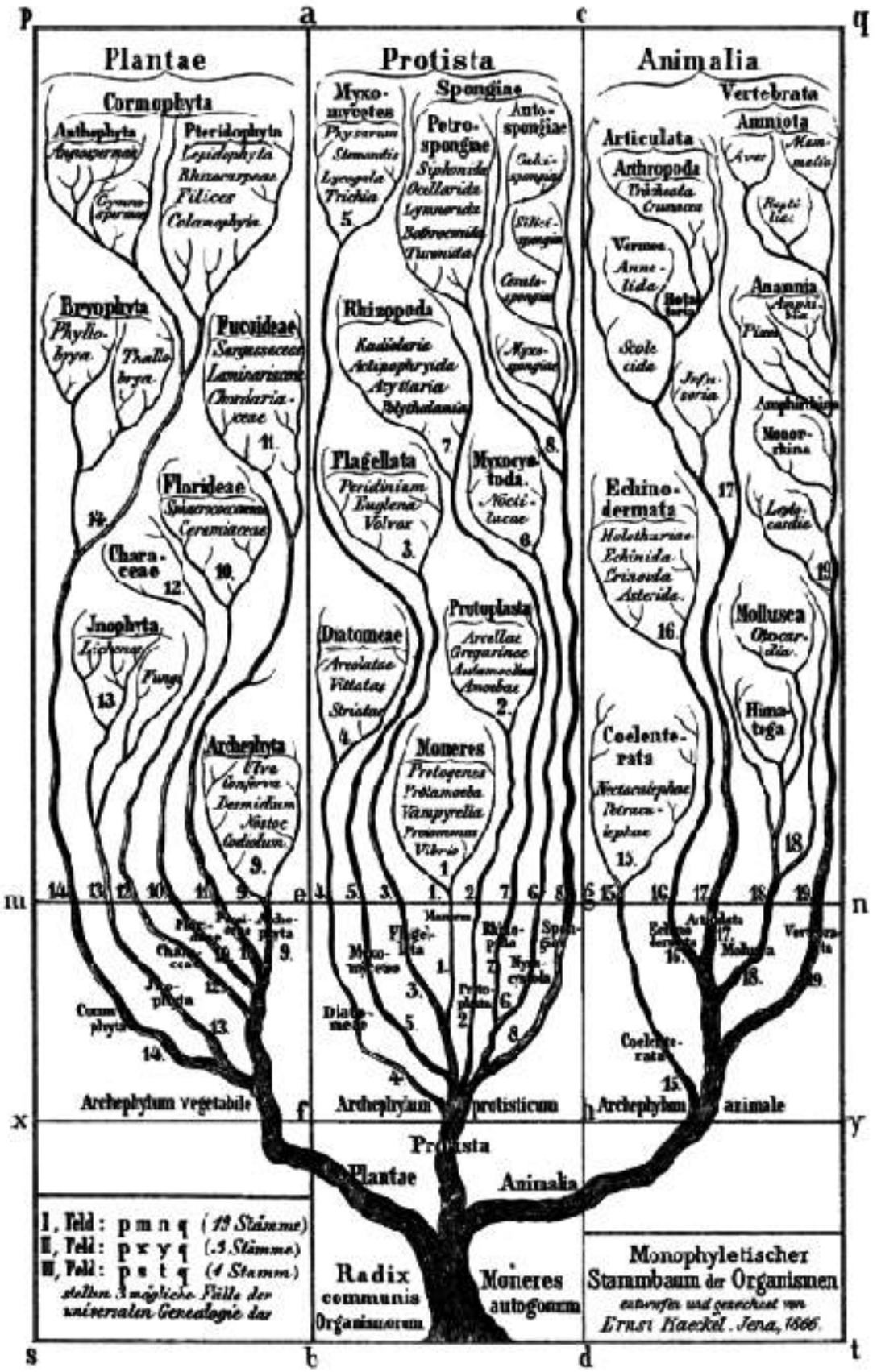
ゲノム情報解析入門

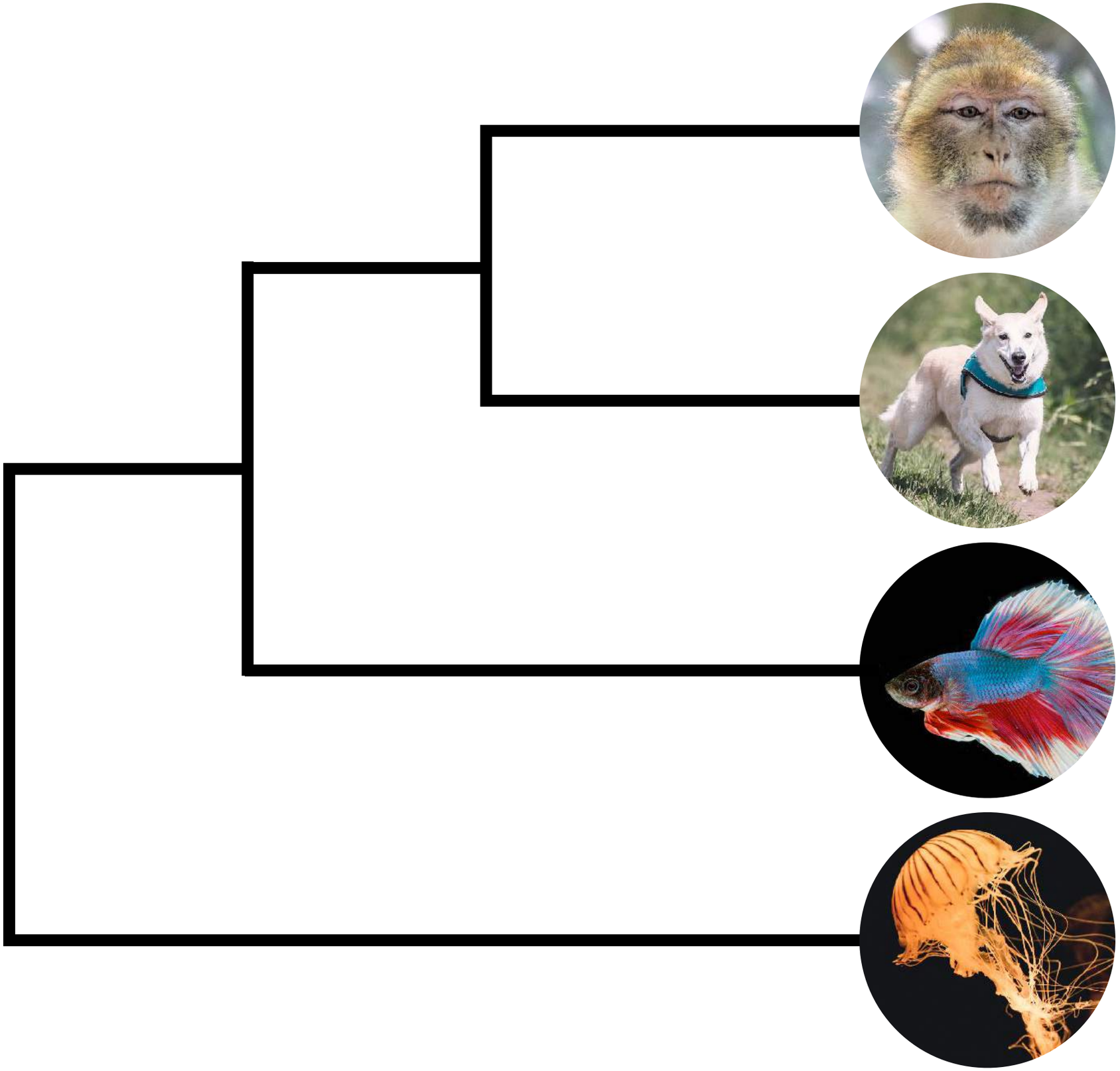
Lab. Crop Evolution

Pythonをはじめる - 実践 -

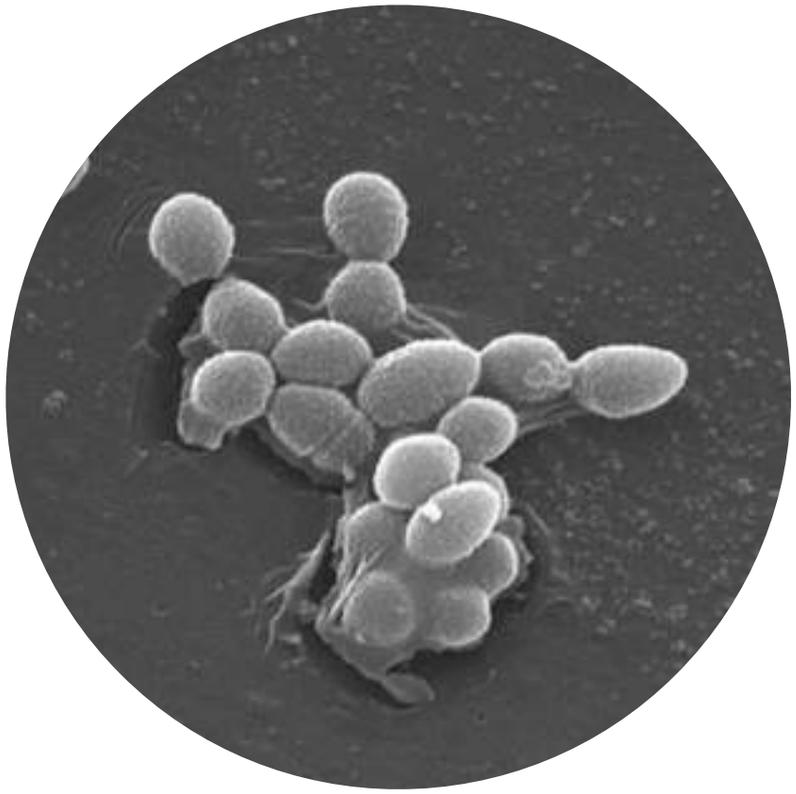


http://ja.wikipedia.org/

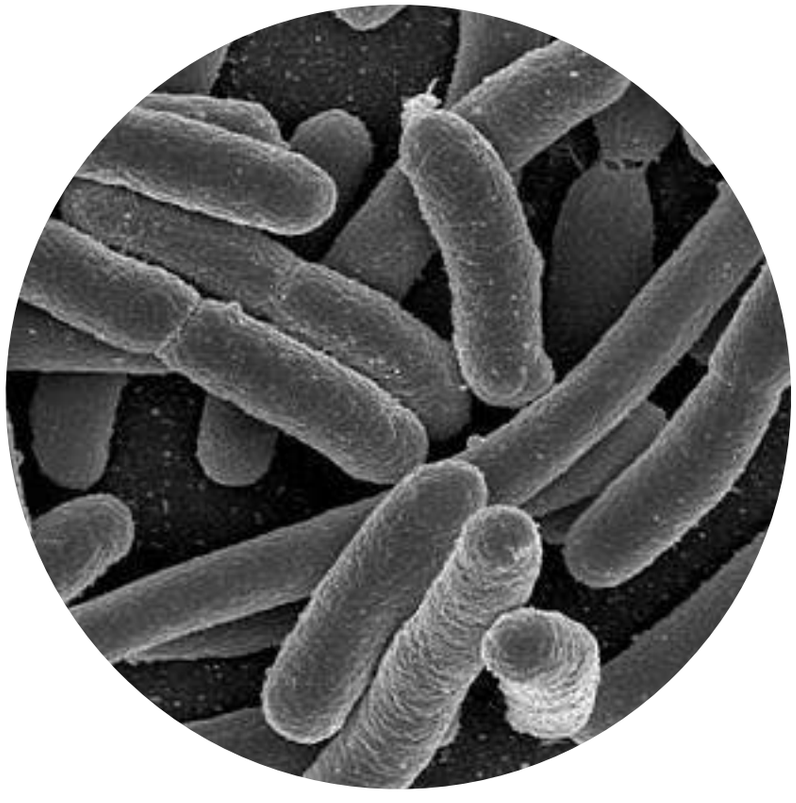
系統関係は？



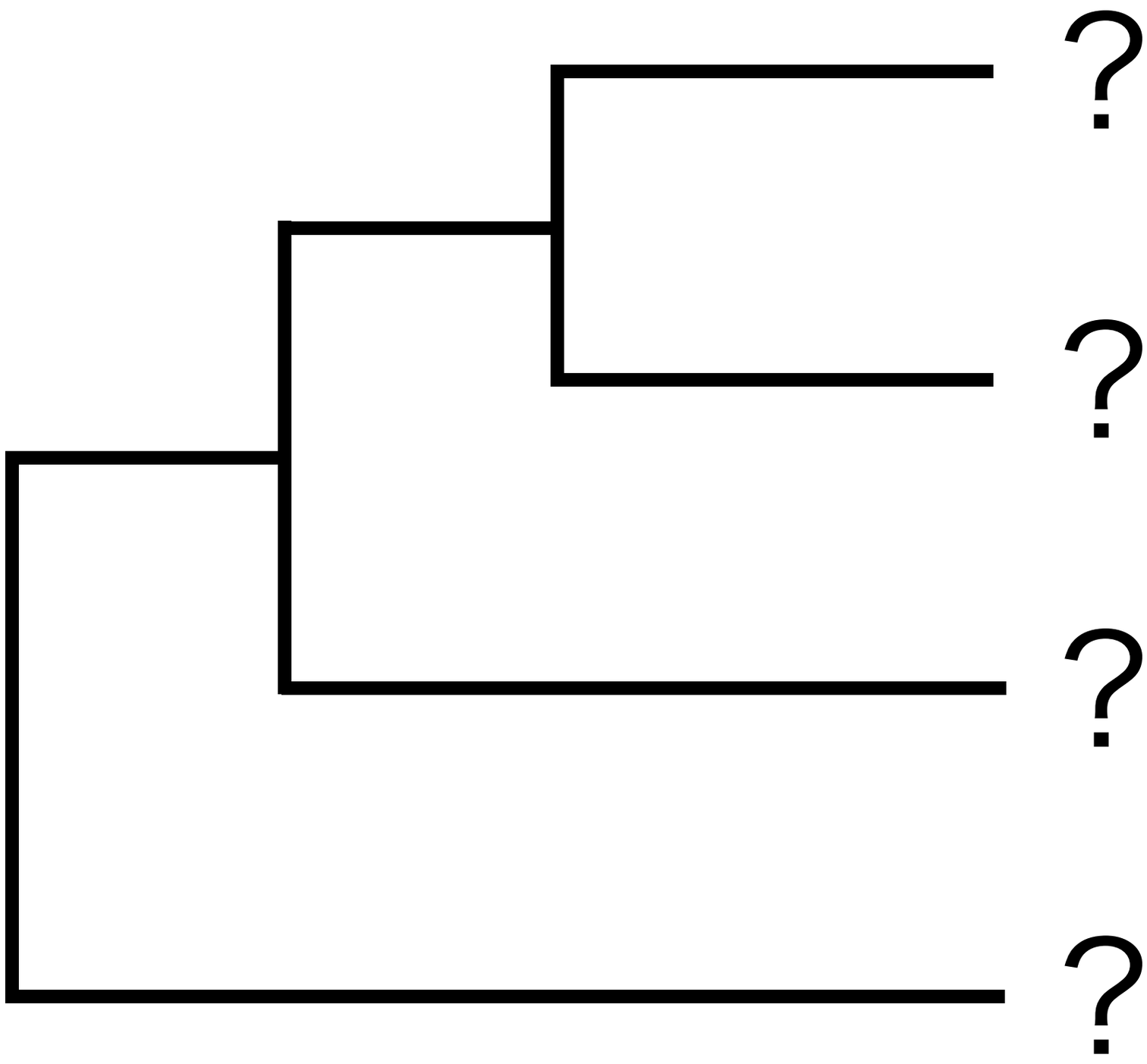
見た目で分かる



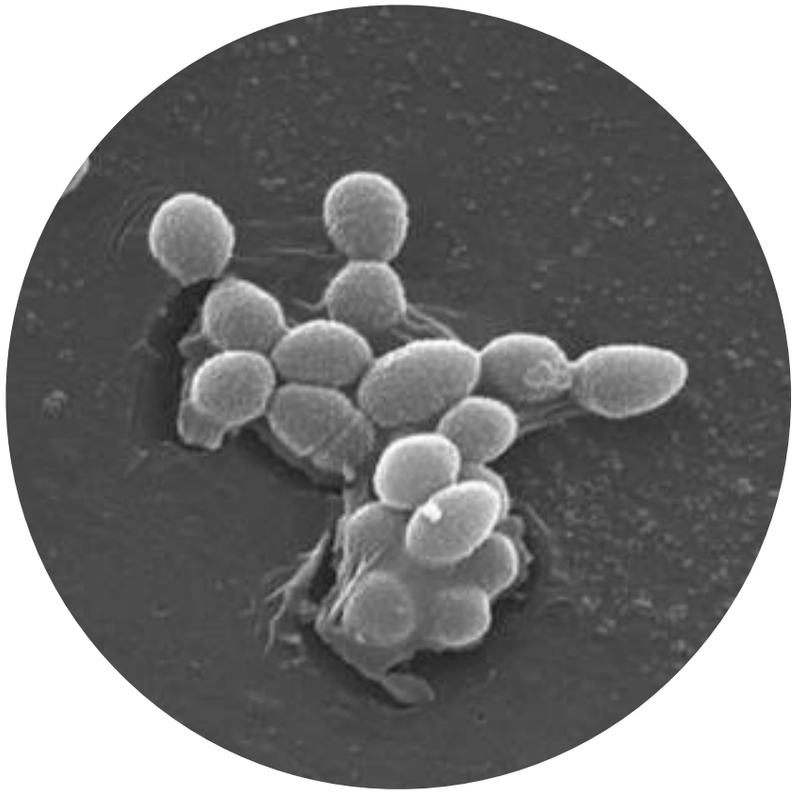
系統関係は？



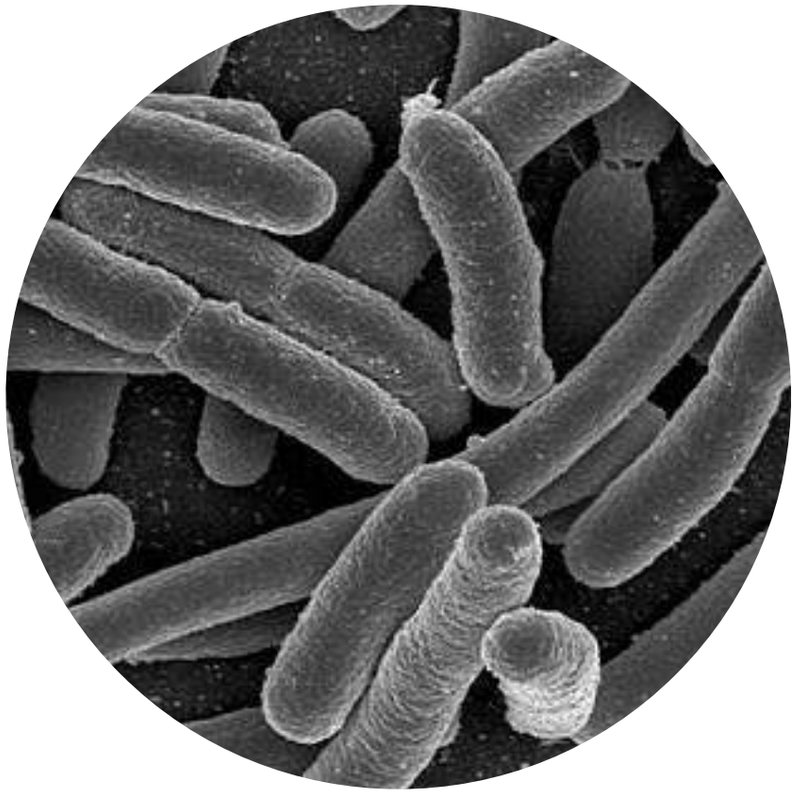
http://ja.wikipedia.org/



見た目で分類できない



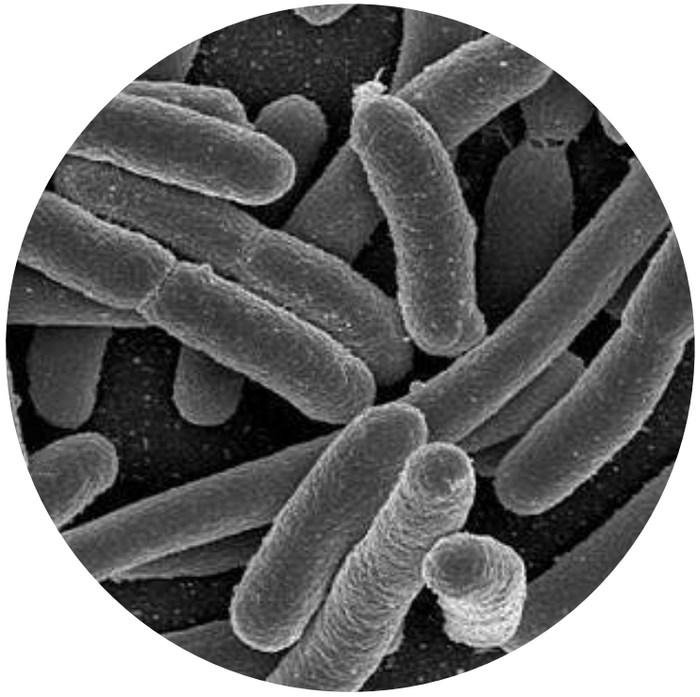
系統関係は？

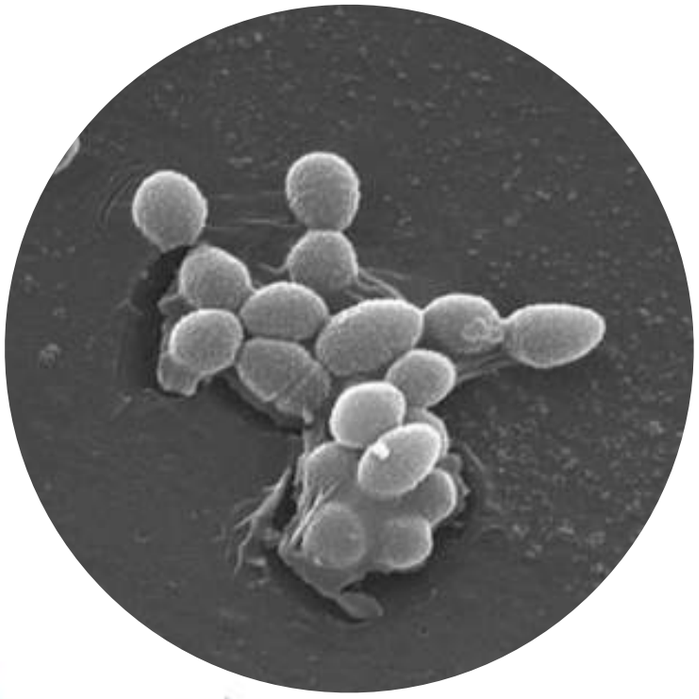


http://ja.wikipedia.org/

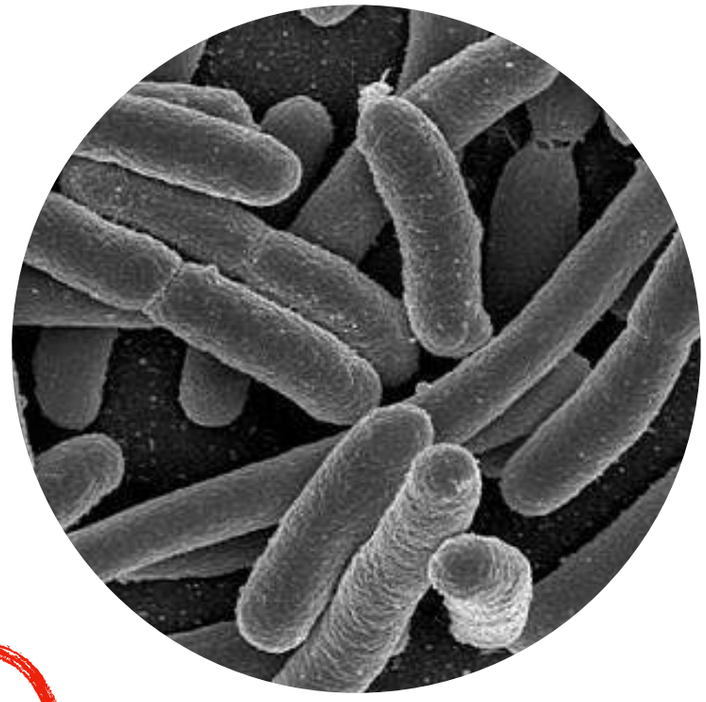
DNA配列

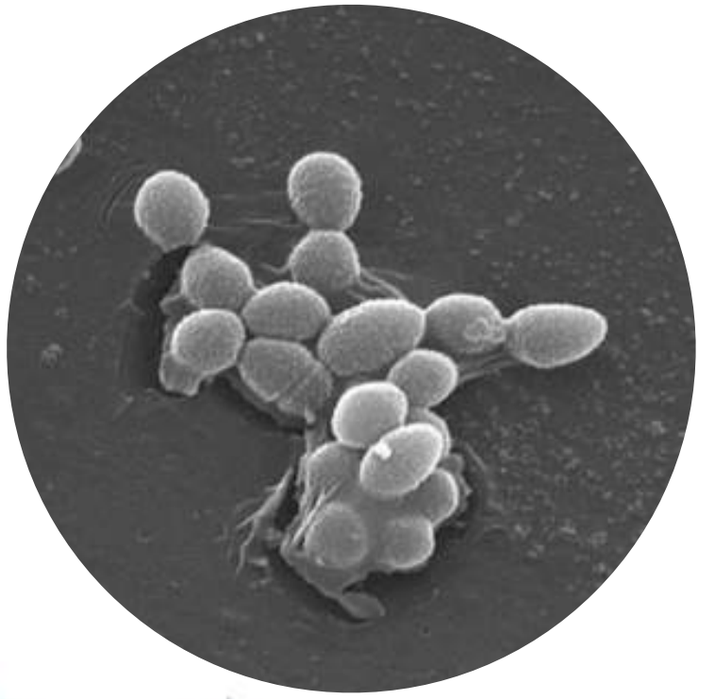
基準

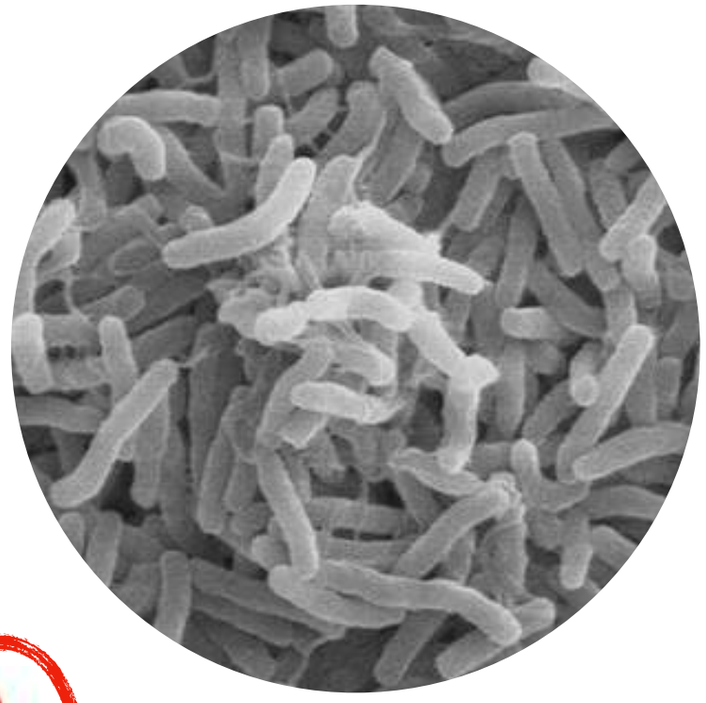




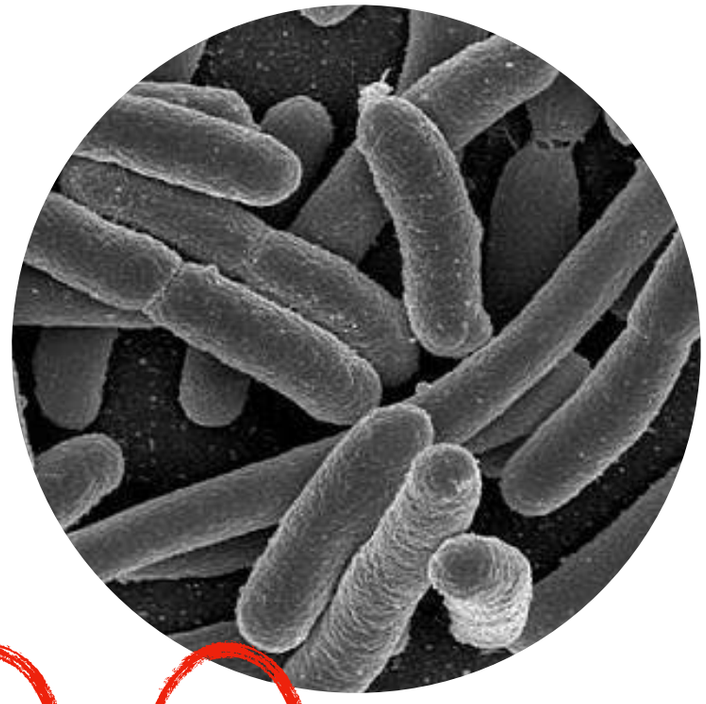
基準

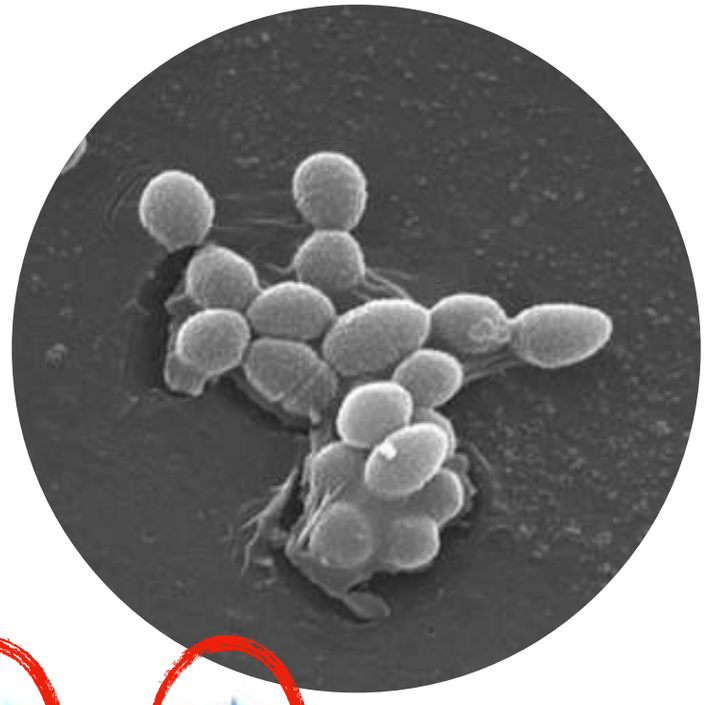




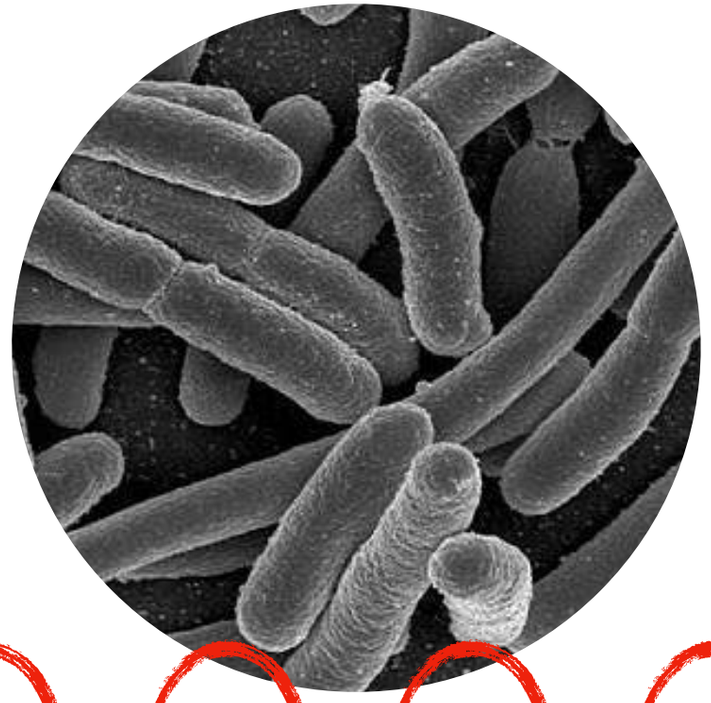


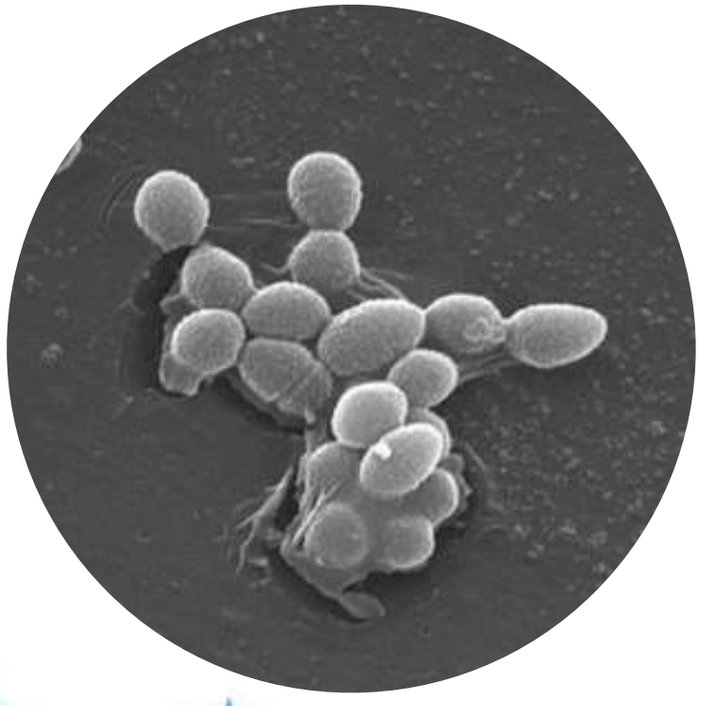
基準

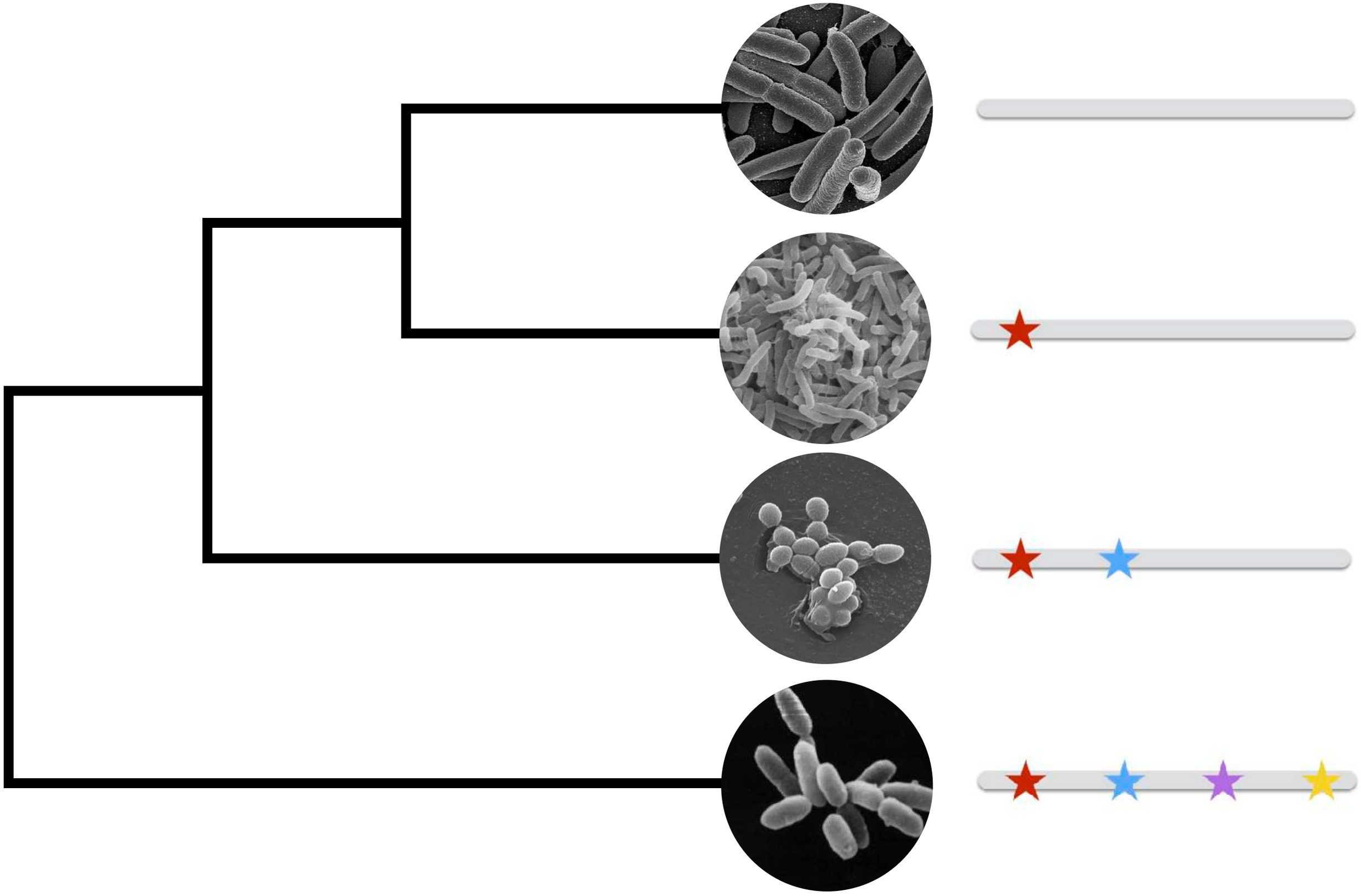




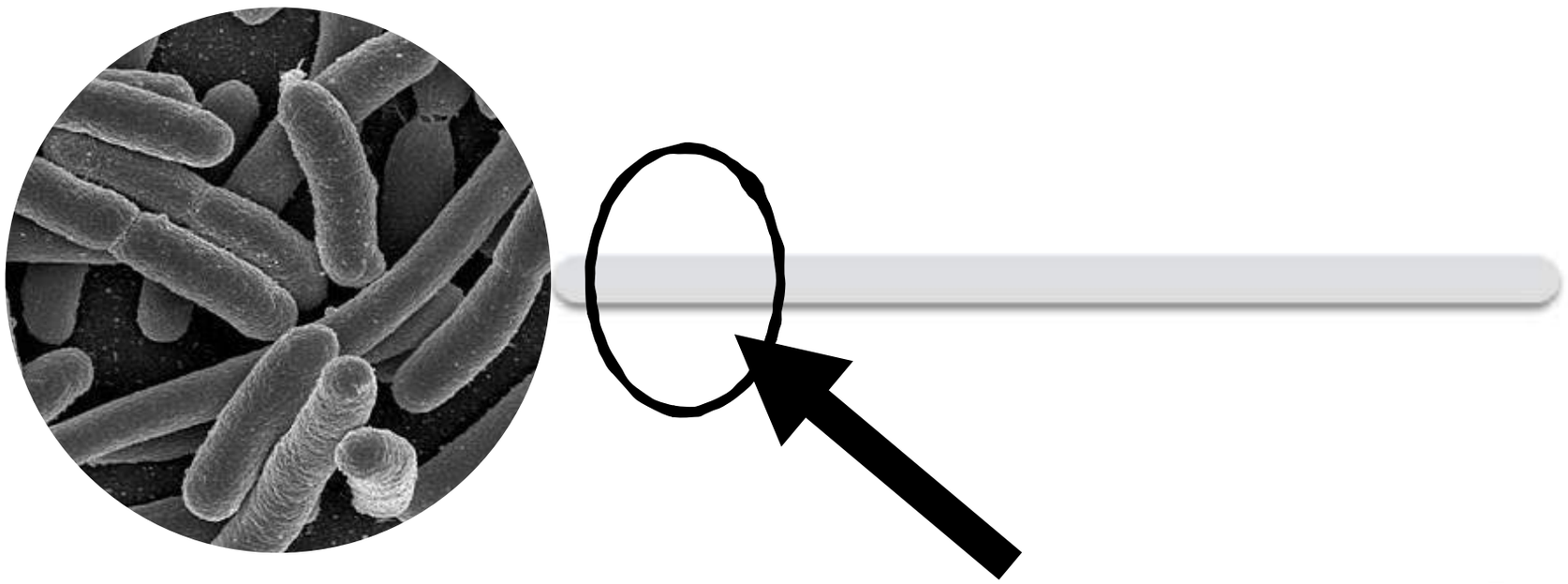
基準



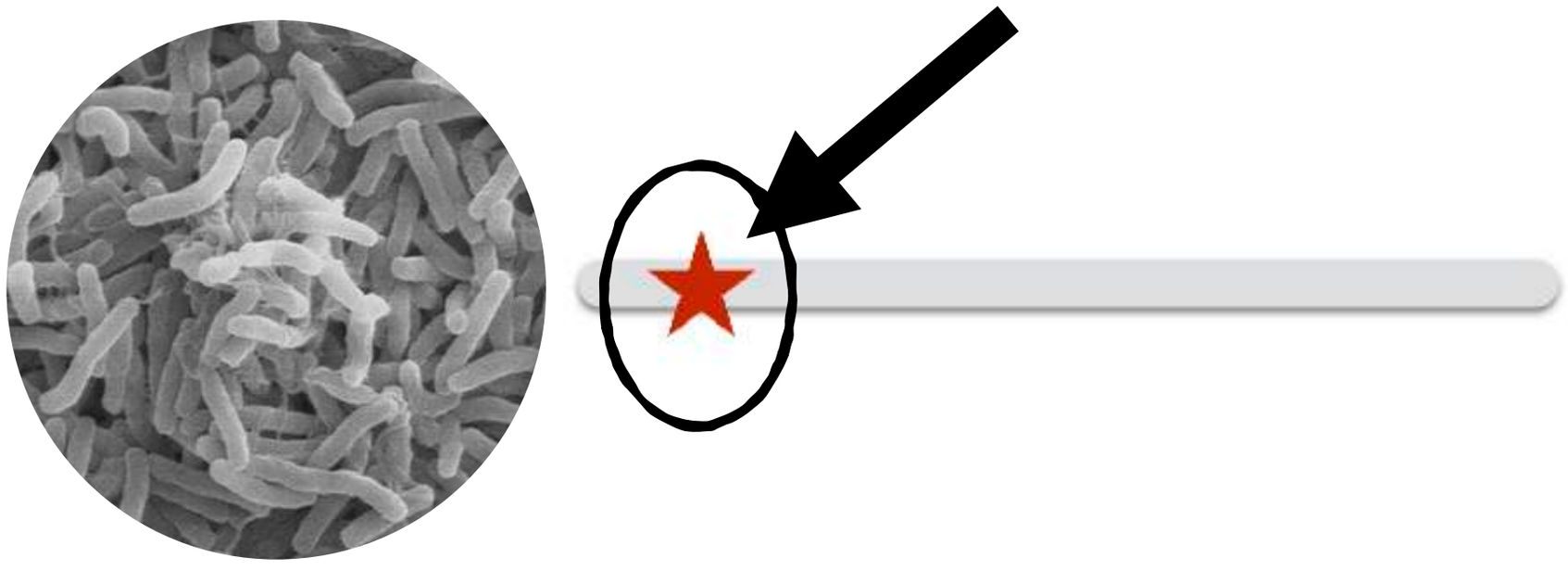




DNA配列で系統関係が分かる

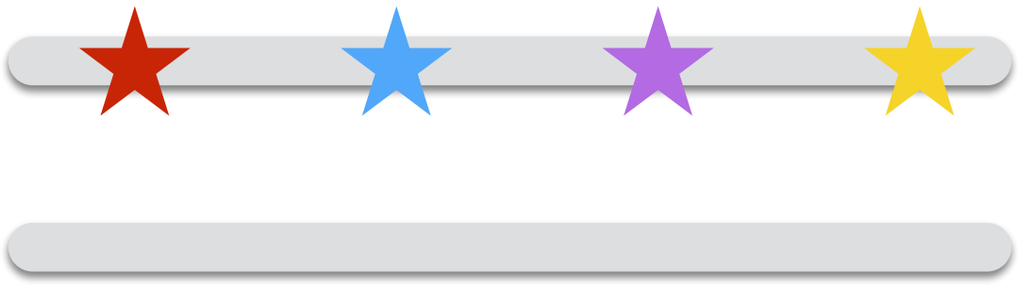


segregating site



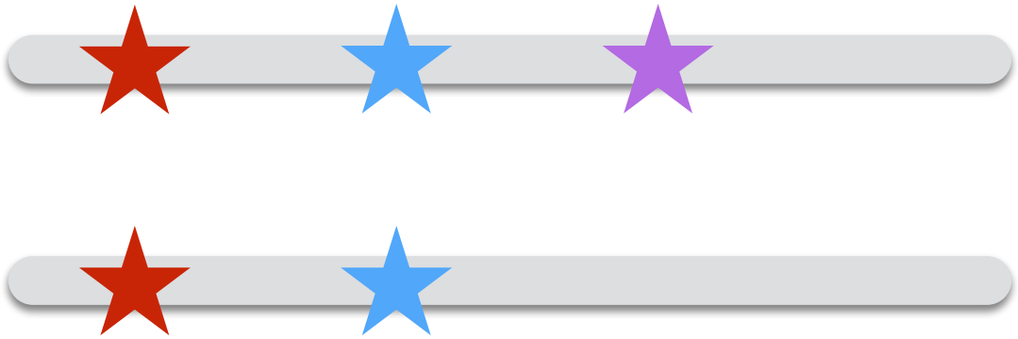
segregating site

例 1)



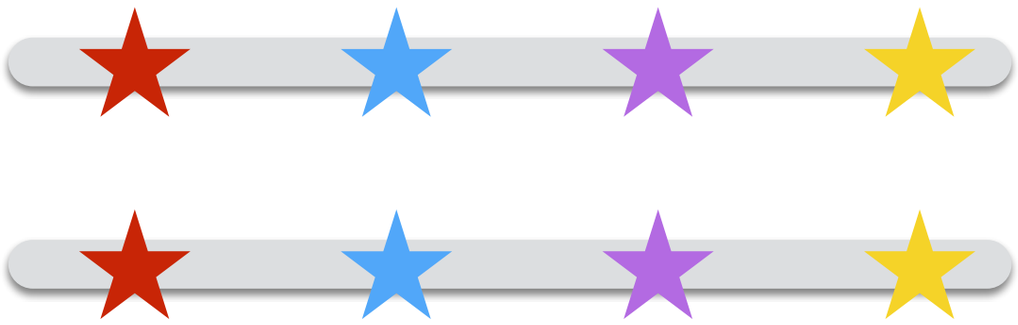
S = 4

例 2)



S = 1

例 3)



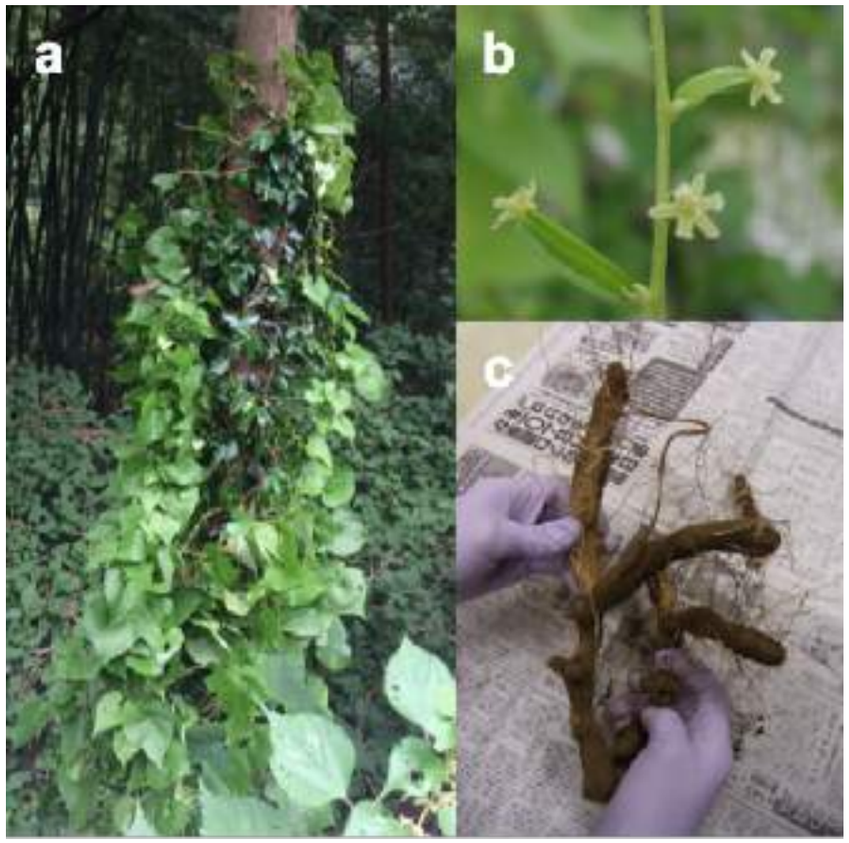
S = 0

Pythonで

計算してみましょう

実際に計算してみましょう

Dioscorea tokoro



S2 = ?

Dioscorea sylvatica

LLIFLE www.llifle.com

S1 = ?

Dioscorea elephantipes

Wikipedia commons

S3 = ?

- segregaticn site
- nuclectide diversity
- Fst

1. segregating site

matKという遺伝子は植物の分類に用いられる遺伝子です。この遺伝子は幅広い植物で保存されています。塩基の欠失や挿入は扱いにくいため今回は除きましたが、一般的な系統樹では含めて計算します。

下のセルから、ヤマノイモ属の3種の配列を読み込んでください。

```
[ ]  #以下の3つの配列を比較してsegregaticn siteを求める
     # matK geneの配列
     Dioscorea_tokoro = "GCAAAGGGACTTCCTATTCCTATATCCACTTCTCTTTCAGGAGTATATTTACACACTCGCTCATGATCATGATTTAAATGGTTCGATTTTTGTGGG
     Dioscorea_sylvatica ="GCAAAGGCACTTCCTATTCCTATATCCACTTCTCTTTCAGGAGTATATTTACACACTTGCTCATGATCATGGTTTAAATGGTTCGATTTTTGTGG
     Dioscorea_elephantipes = "GCAAAGGCACTTCCTATTCCTATATCCACTTCTCTTTCAGGAGTATATTTACACACTTGCTCATGATCATGGTTTAAATGGTTCGATTTTTG
```

まずはDioscorea_tokoroとDioscorea_sylvaticaについて、関数を用いずにsegregation siteの数を求めてみましょう。segregating siteとは、二つの配列を比較した時に塩基が異なっている箇所のことです。

例 1)

segregating site

S = 4

例 2)

S = 1

例 3)

S = 0

- segregaticn site
- nuclectide diversity
- Fst

1. segregating site

matKという遺伝子は植物の分類に用いられる遺伝子です。この遺伝子は幅広い植物で保存されています。塩基の欠失や挿入は扱いにくいため今回は除きましたが、一般的な系統樹では含めて計算します。

下のセルから、ヤマノイモ属の３種の配列を読み込んでください。

```
[ ]  #以下の3つの配列を比較してsegregaticn siteを求める
     # matK geneの配列
     Dioscorea_tokoro = "GCAAAGGGACTTCCTATTCCTATATCCACTTCTCTTTCAGGAGTATATTTACACACTCGCTCATGATCATGATTTAAATGGTTCGATTTTTGTGGG
     Dioscorea_sylvatica ="GCAAAGGCACTTCCTATTCCTATATCCACTTCTCTTTCAGGAGTATATTTACACACTTGCTCATGATCATGGTTTAAATGGTTCGATTTTTGTGG
     Dioscorea_elephantipes = "GCAAAGGCACTTCCTATTCCTATATCCACTTCTCTTTCAGGAGTATATTTACACACTTGCTCATGATCATGGTTTAAATGGTTCGATTTTTG
```

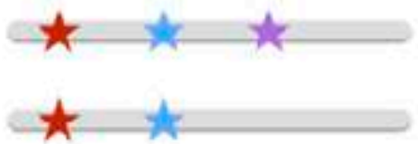
まずはDioscorea_tokoroとDioscorea_sylvaticaについて、関数を用いずにsegregation siteの数を求めてみましょう。segregating siteとは、二つの配列を比較した時に塩基が異なっている箇所のことです。

例 1)

segregating site

S = 4

例 2)



S = 1

例 3)

S = 0

*Dioscorea tokoro*と
*Dioscorea sylvatica*について
Segregating site数を求めましょう

求められたら自作関数を作って
残りも計算しましょう

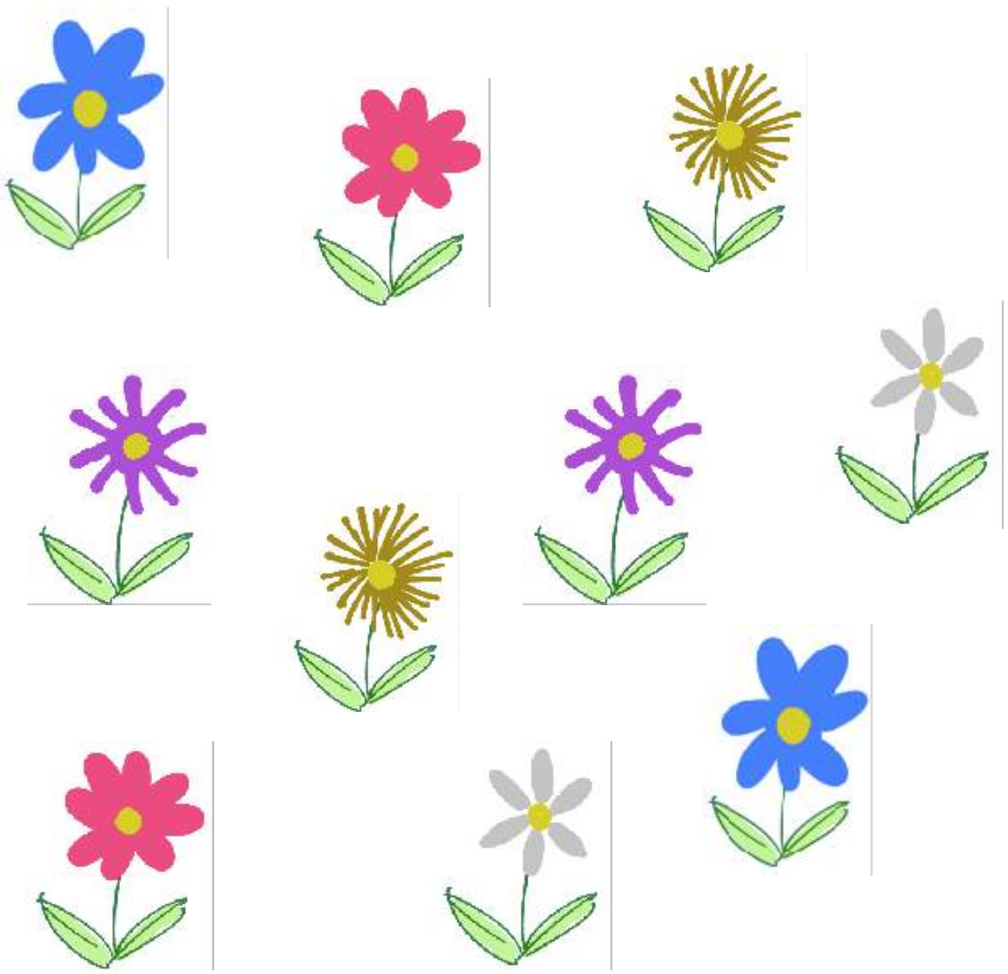
```
# D. tokoro と D. sylvatica の Segregating site数を求める

S = 0      # Segregation siteの初期値

# 2種の塩基配列のループ処理
for nuc1, nuc2 in zip(Dioscorea_tokoro, Dioscorea_sylvatica):
  if nuc1 != nuc2:      # tokoroの塩基とsylvaticaの塩基が異なる場合、
        S = S + 1       # Segregation siteの値を1増やす

print(S)     # 最終的なSegregation siteの値を出力する
```


tokoro
sylvatica
elephantipes

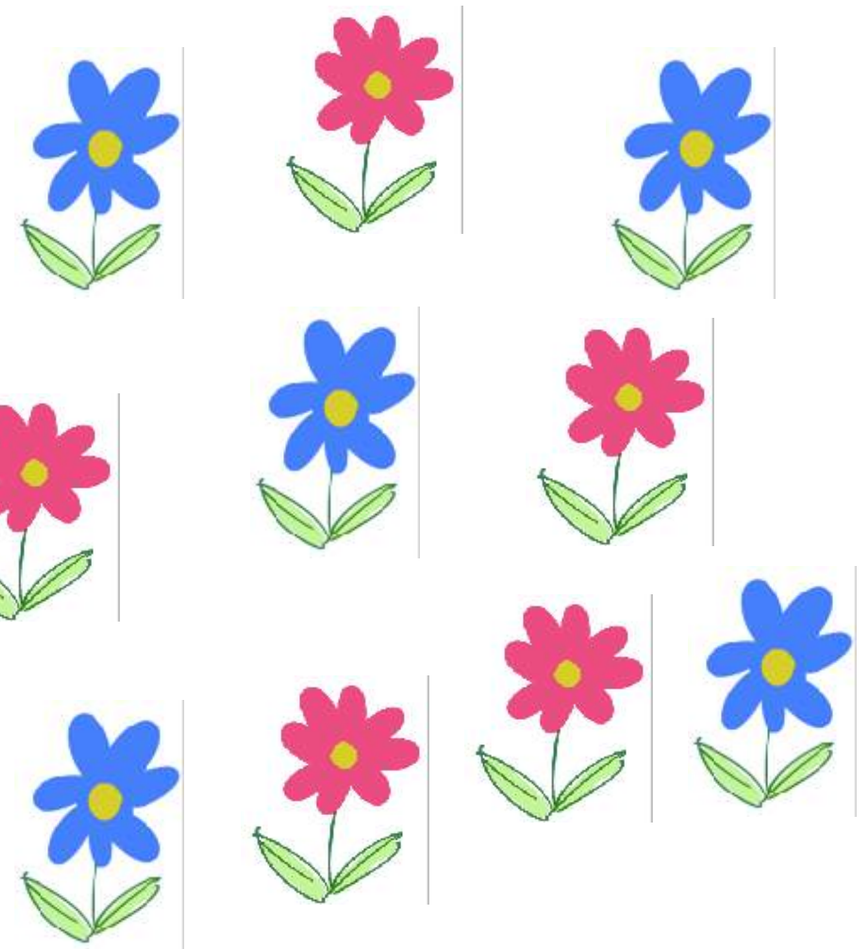
segregating siteで
多様性が分かる

どちらが多様だと思いますか？

A

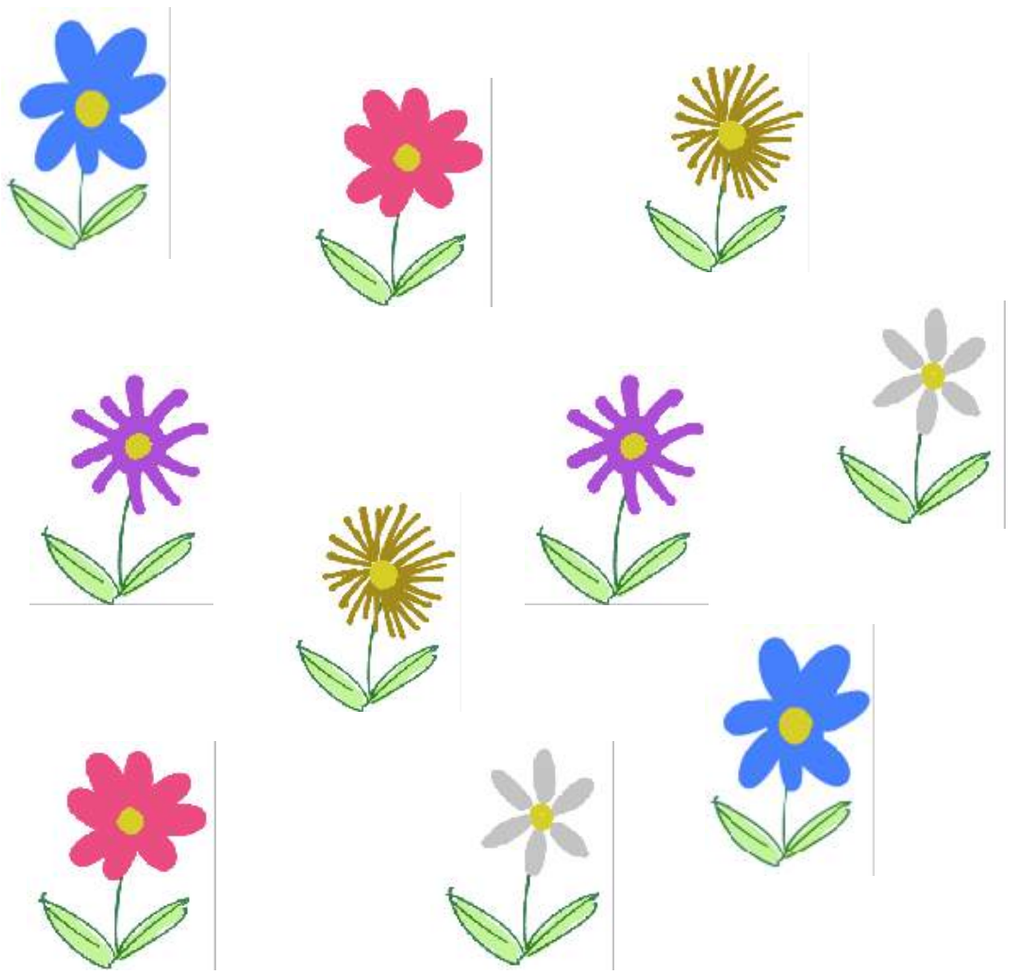


B

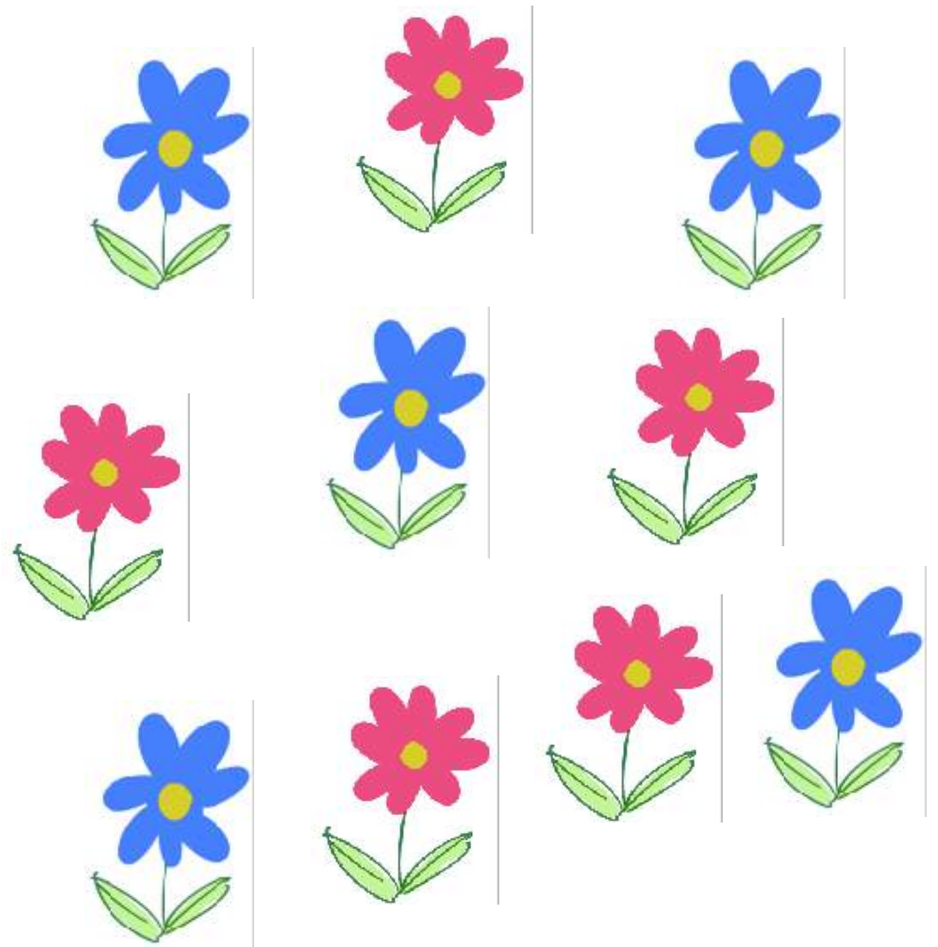


どちらが多様だと思いますか？

A



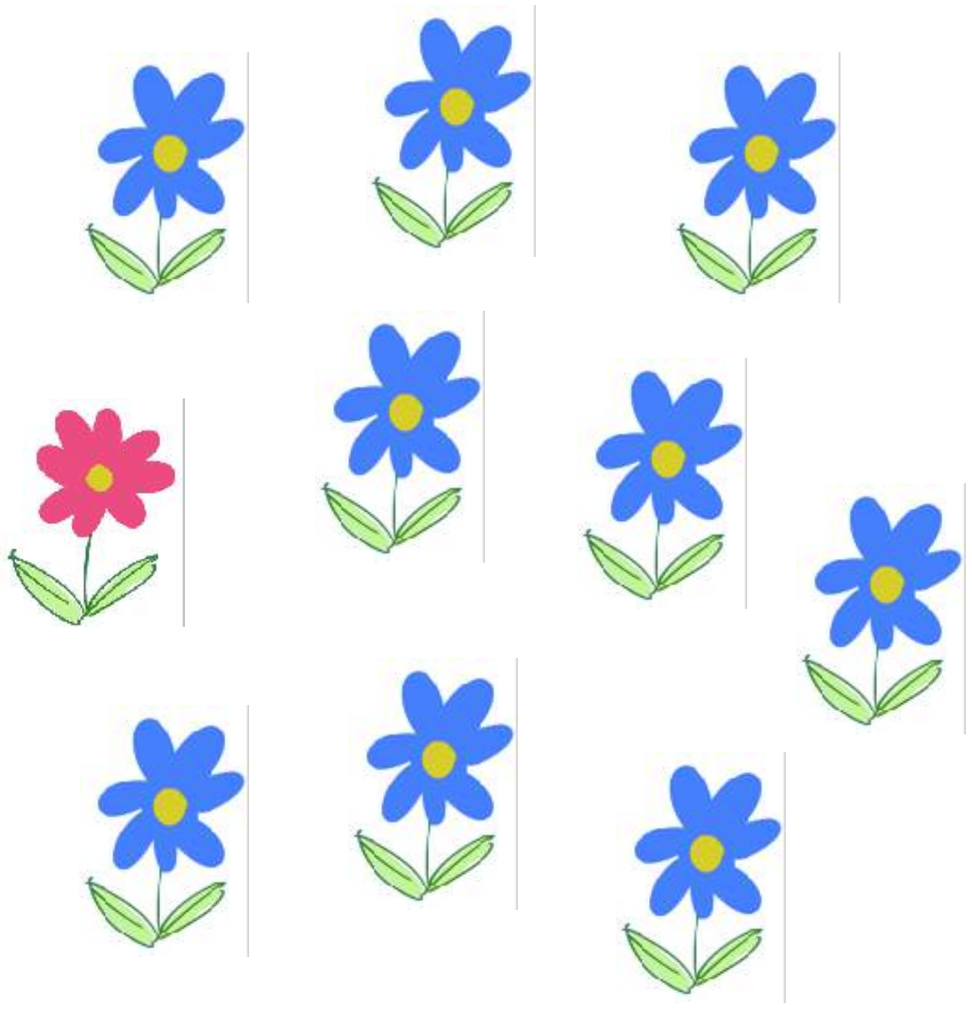
B



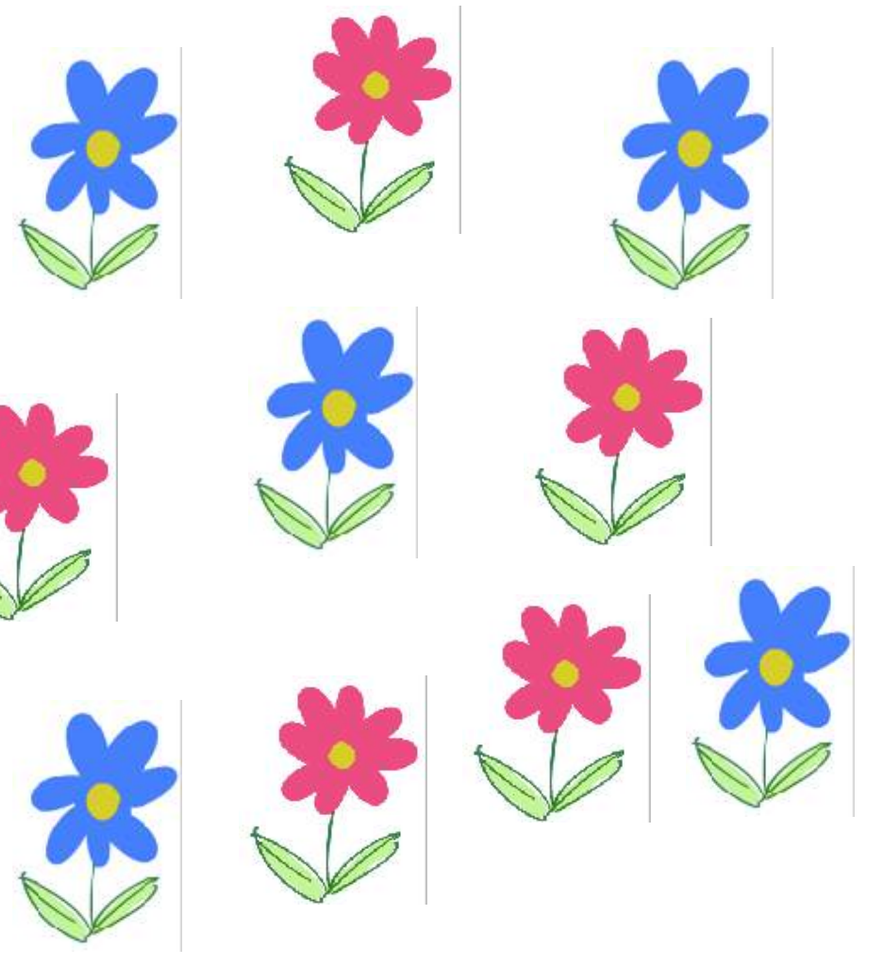
多様性は種類で決まる

どちらが多様だと思いますか？

A

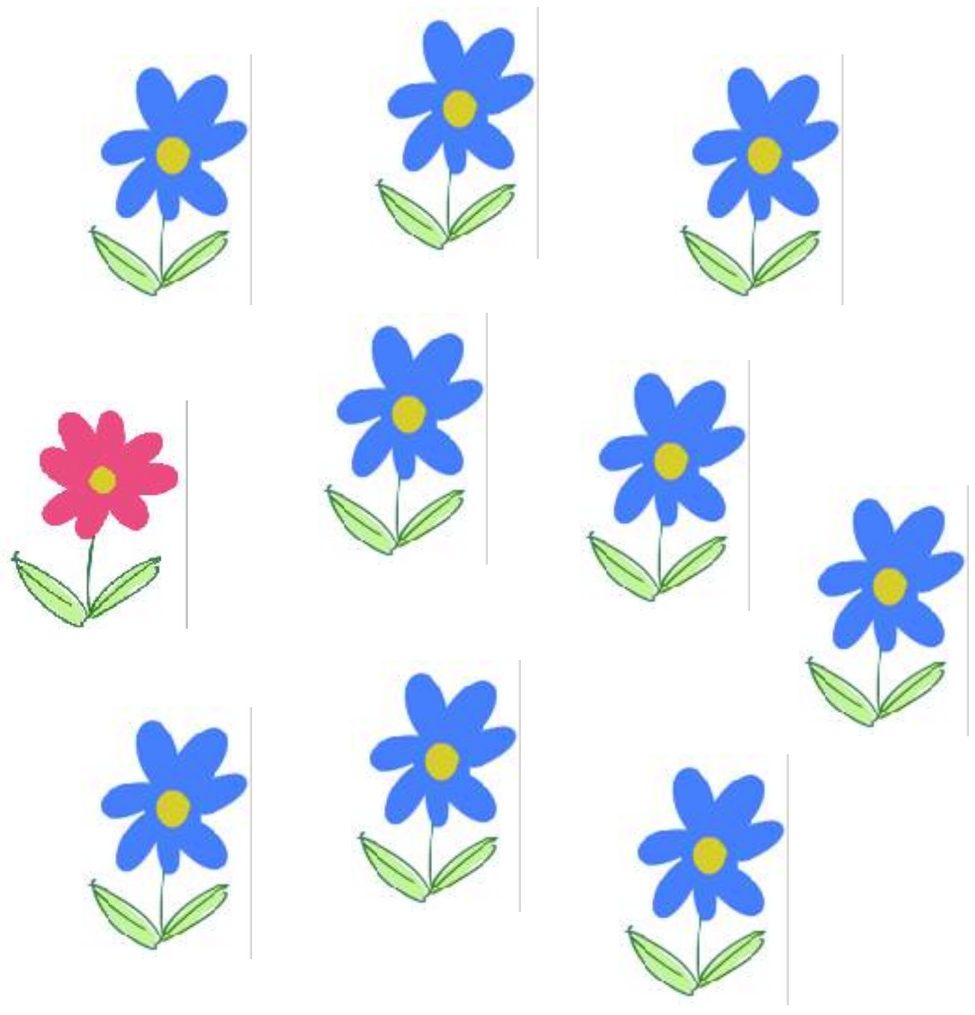


B

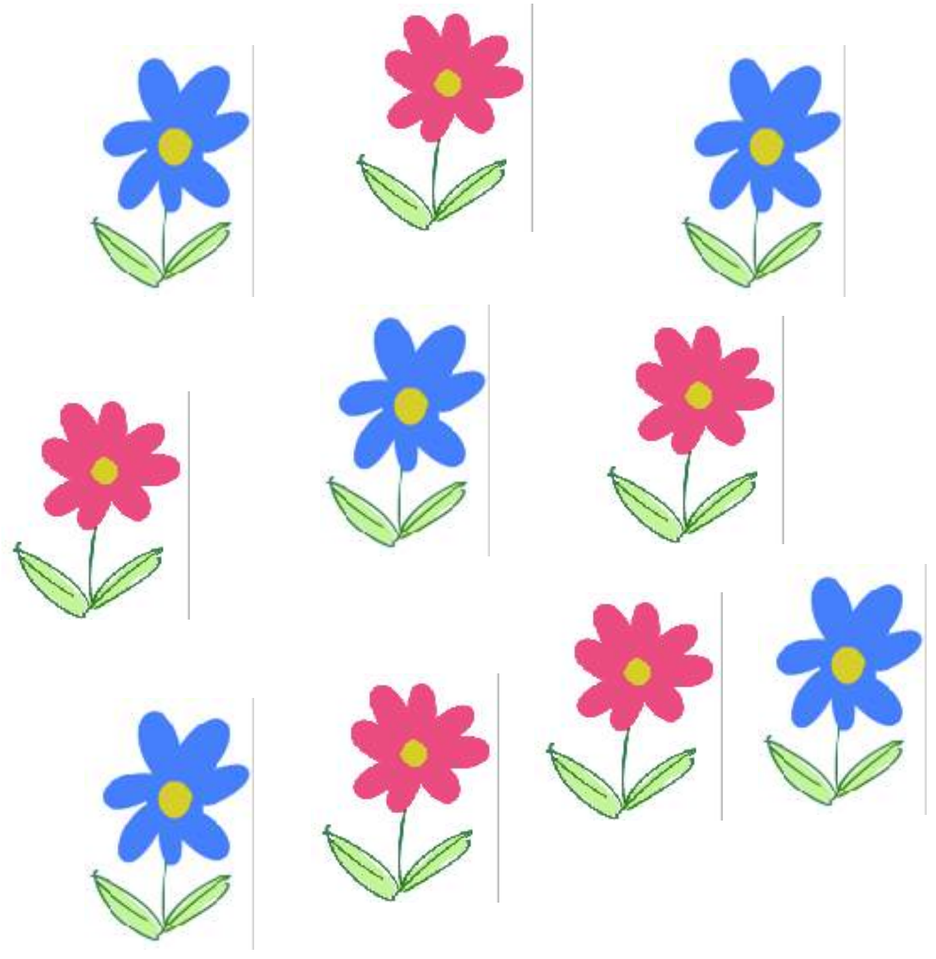


どちらが多様だと思いますか？

A



B



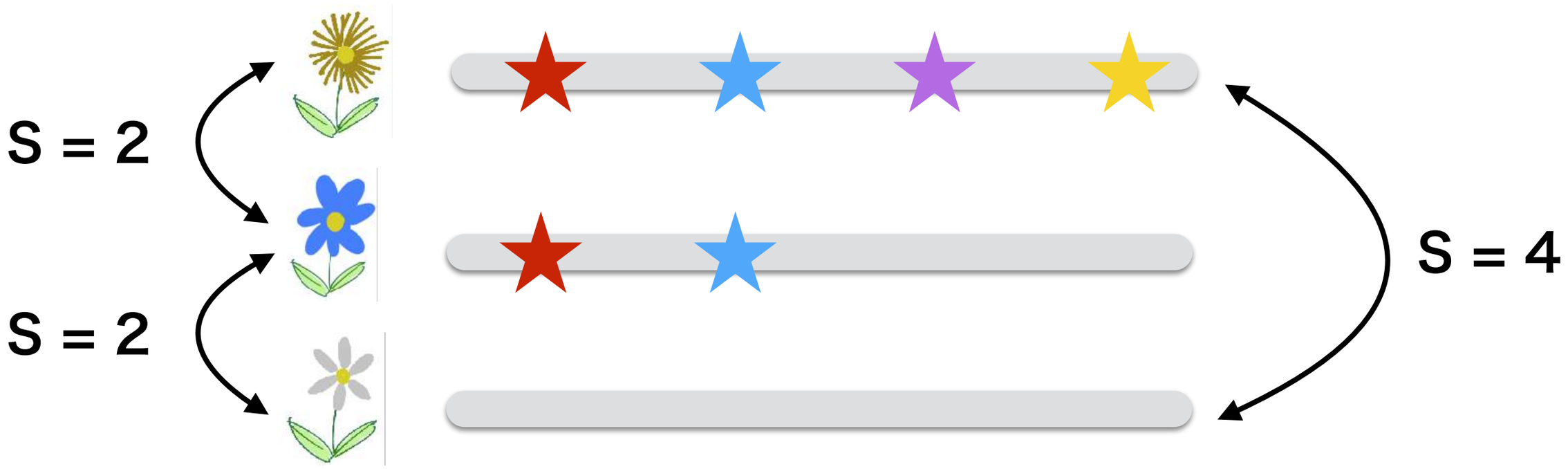
多様性は頻度でも決まる

種類

頻度

多様性

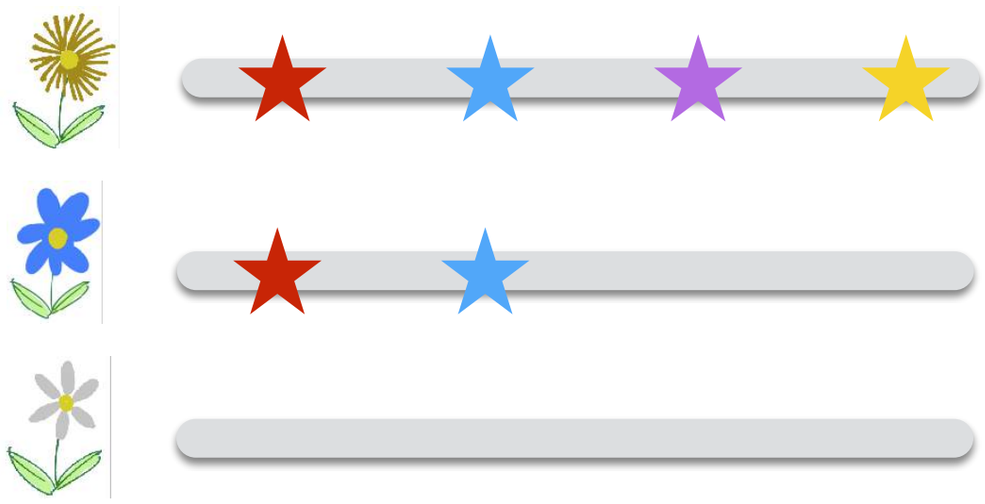
多様性の評価（方法１）



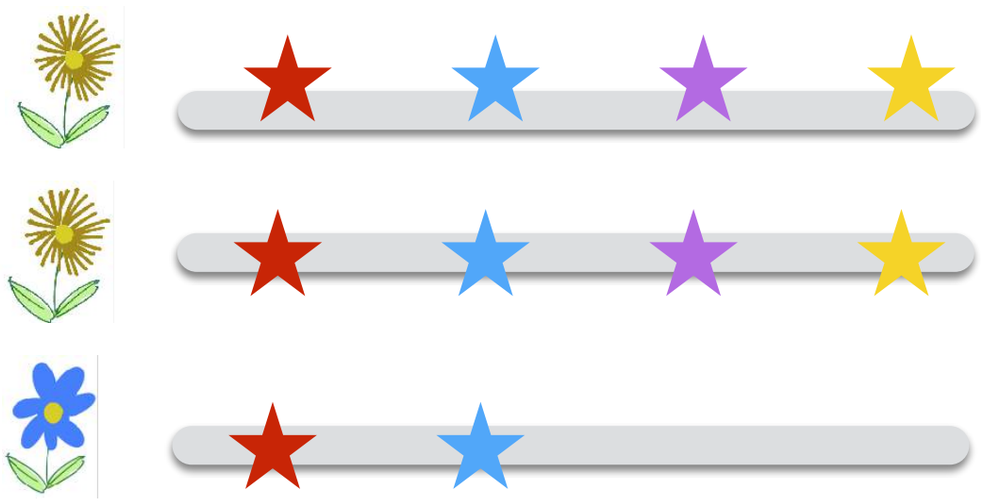
$$\pi = \frac{\text{（全ての組み合わせのSの総和）}}{\text{（配列の組み合わせ）}} = \frac{2 + 2 + 4}{3} = \frac{8}{3}$$

どちらが多様でしょうか？

A

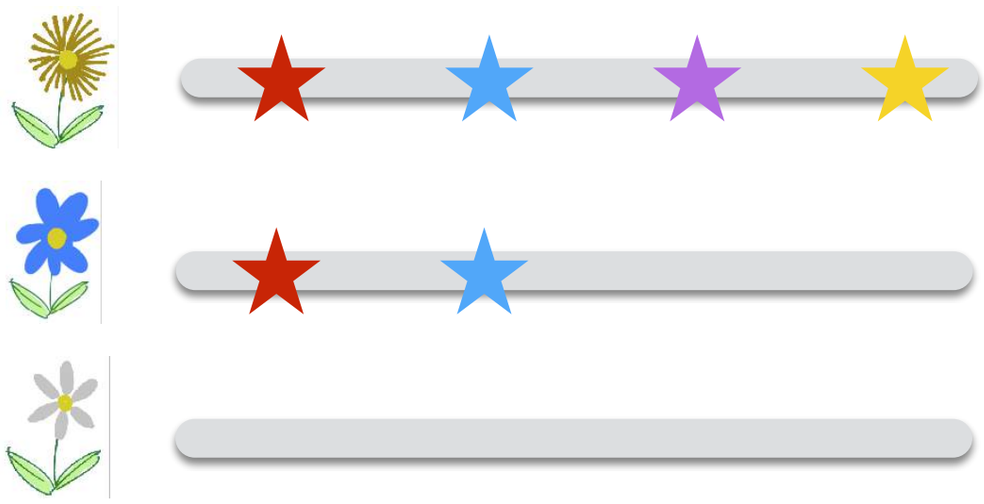


B

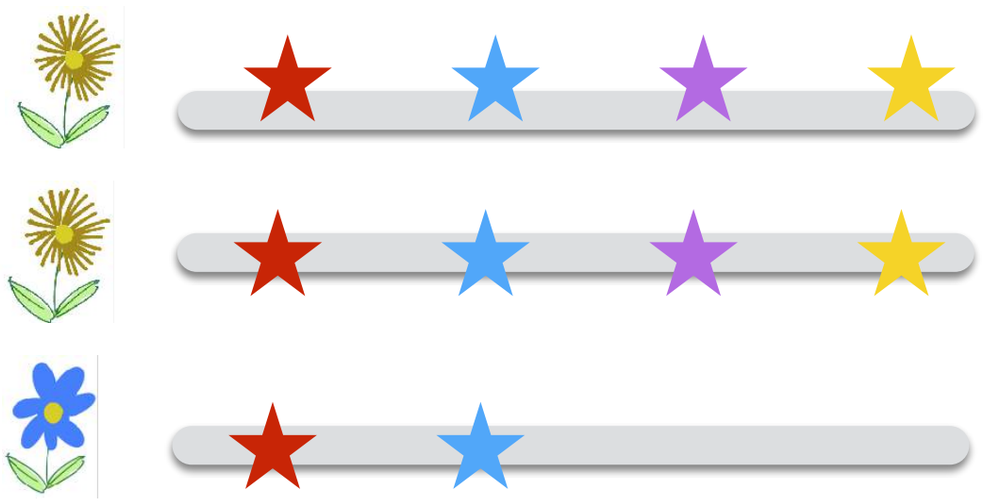


どちらが多様でしょうか？

A



B

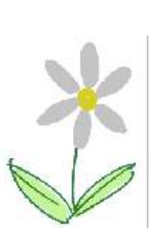


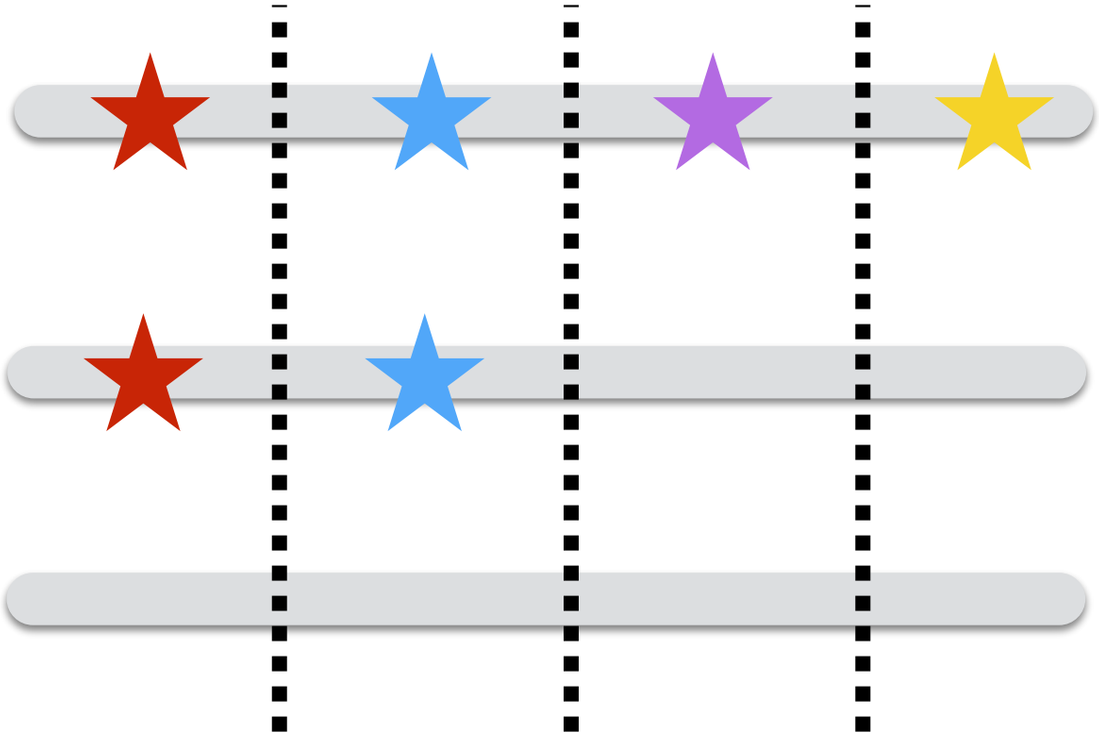
$$\frac{2+2+4}{3} = \frac{8}{3} > \frac{0+2+2}{3} = \frac{4}{3}$$

多様性の評価（方法２）

※ 方法１と方法２は同じ結果を示します。

基準配列 →





(1) 基準配列も含めて

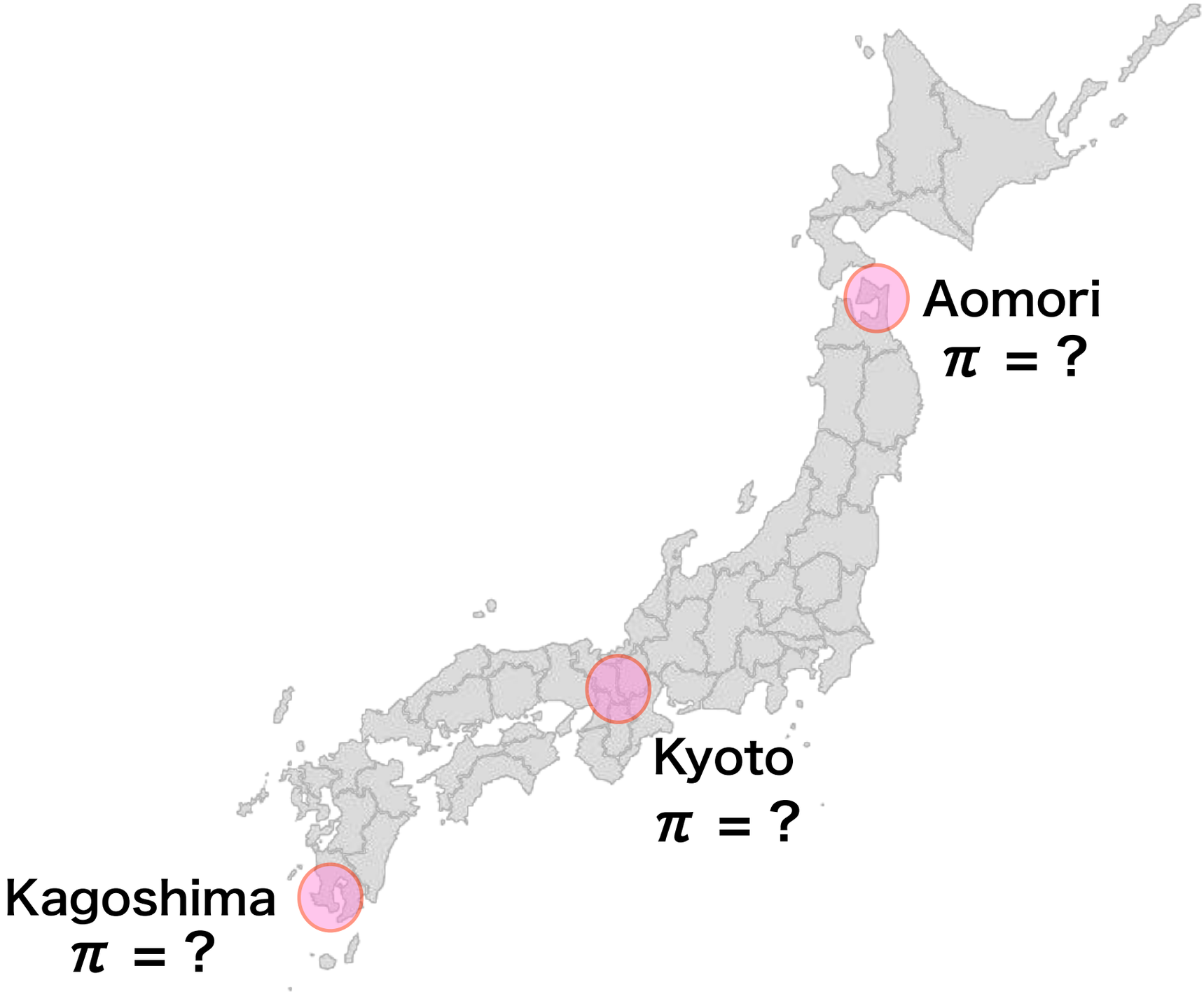
基準と同じ　2

基準と違う　1

(2)（基準と同じ）×（基準と違う）　2 × 1 = 2 ＋ 2 × 1 = 2 ＋ 1 × 2 = 2 ＋ 1 × 2 = 2　＝ 8

(3) 配列の総組み合わせで割る　8 / ${}_3C_2$ = 8/3

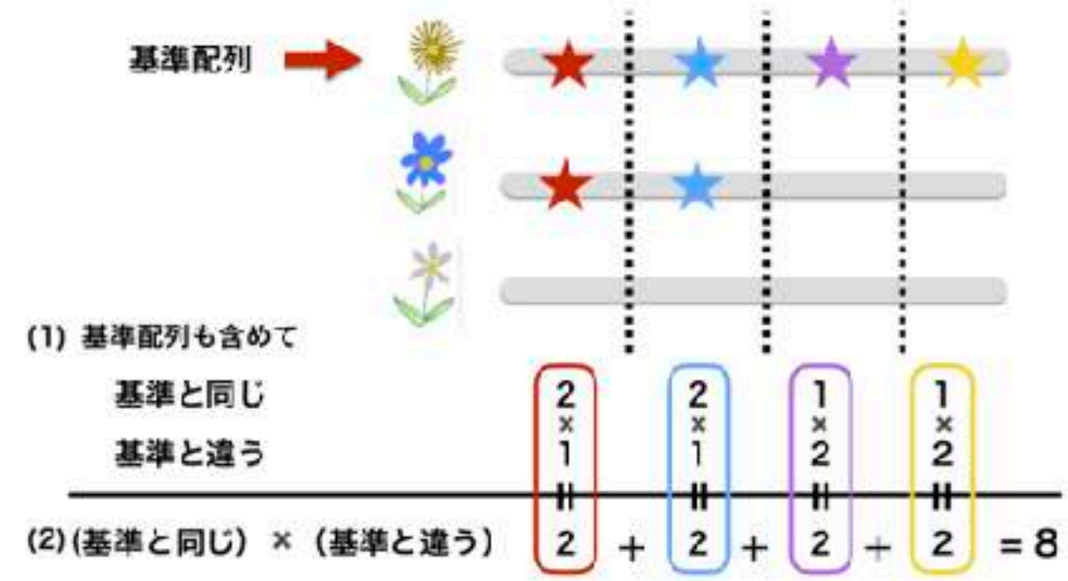
実際に計算してみましょう



2. nucleotide diversity

実際のデータを使って多様性(π, nucleotide diversity)を計算しましょう。

nucleotide diversityとは、**複数の個体(配列)において塩基配列に違いがある確率を示す値** で、下の図のように、一塩基ずつ基準配列と違いがあるか見ていくことで計算することができます。また、nucleotide diversityはπという記号で表されることが一般的です。



以下は、青森県、京都府、鹿児島県でのオニドコロの配列の一部を切り取ったデータです。それぞれの地域で１０個体（二倍体なので、計２０本の染色体）をサンプルしたとしましょう。

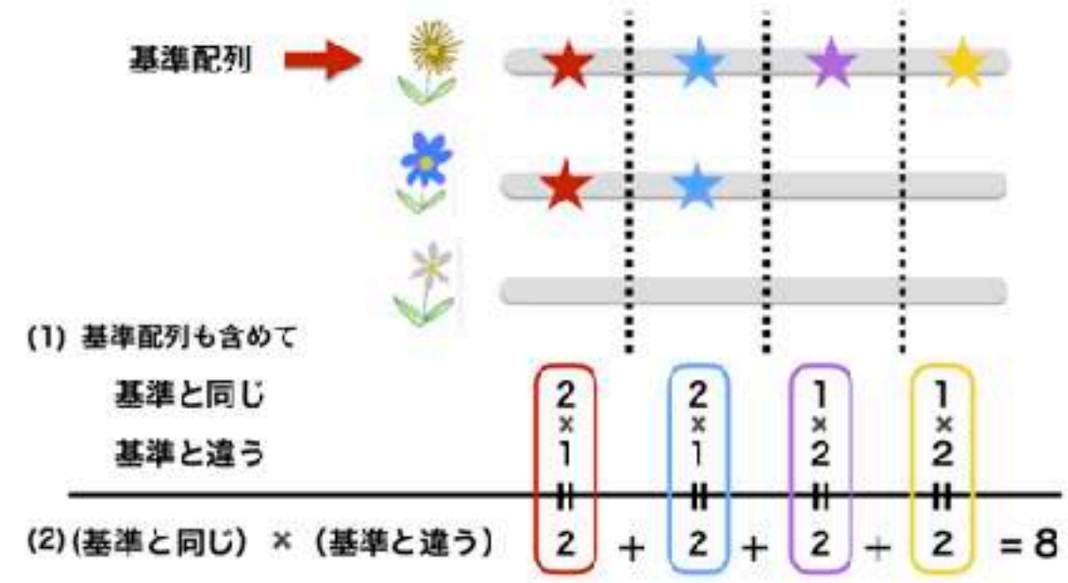
まずは以下のデータを読み込みましょう。

```
[ ]  #nucleotide diversity
     #基準と同じ塩基の数(ref)と、基準と違う塩基の数(alt)がそれぞれわかっていて、リストになっている。

     ref_aomori = [20, 20, 9, 20, 19, 11, 20, 16, 20, 20, 0, 20, 20, 20, 20, 20, 20, 15, 20, 20, 15, 20, 20, 10, 20, 14, 20, 20, 10, 20, 18, 11, 20, 0, 5, 20, 3, 9, 1, 20,
     alt_aomori = [0, 0, 11, 0, 1, 9, 0, 4, 0, 0, 20, 0, 0, 0, 0, 0, 0, 5, 0, 0, 5, 0, 0, 10, 0, 6, 0, 0, 10, 0, 2, 9, 0, 20, 15, 0, 17, 11, 19, 0, 1, 17, 20, 19, 20, 0, 14, 0, 0, 20

     ref_kyoto = [18, 16, 8, 19, 14, 8, 16, 20, 19, 18, 11, 13, 11, 19, 7, 6, 20, 20, 15, 7, 20, 18, 20, 5, 16, 20, 15, 16, 13, 19, 18, 9, 20, 2, 10, 12, 14, 13, 13, 11, 15
     alt_kyoto = [2, 4, 12, 1, 6, 12, 4, 0, 1, 2, 9, 7, 9, 1, 13, 14, 0, 0, 5, 13, 0, 2, 0, 15, 4, 0, 5, 4, 7, 1, 2, 11, 0, 18, 10, 8, 6, 7, 7, 9, 5, 13, 12, 20, 12, 1, 7, 2, 4, 15, 0

     ref_kagoshima = [15, 20, 20, 20, 20, 20, 20, 20, 20, 20, 0, 20, 19, 20, 20, 20, 16, 20, 20, 20, 20, 20, 19, 20, 20, 20, 20, 20, 0, 20, 16, 0, 16, 0, 0, 20, 0, 0, 0, 2
     alt_kagoshima = [5, 0, 0, 0, 0, 0, 0, 0, 0, 0, 20, 0, 1, 0, 0, 0, 4, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 20, 0, 4, 20, 4, 20, 20, 0, 20, 20, 20, 0, 2, 20, 20, 20, 20, 0, 0, 0, 0, 2
```

2. nucleotide diversity

実際のデータを使って多様性(π, nucleotide diversity)を計算しましょう。

nucleotide diversityとは、**複数の個体(配列)において塩基配列に違いがある確率を示す値** で、下の図のように、一塩基ずつ基準配列と違いがあるか見ていくことで計算することができます。また、nucleotide diversityはπという記号で表されることが一般的です。



以下は、青森県、京都府、鹿児島県でのオニドコロの配列の一部を切り取ったデータです。それぞれの地域で１０個体（二倍体なので、計２０本の染色体）をサンプルしたとしましょう。

まずは以下のデータを読み込みましょう。

```
[ ]  #nucleotide diversity
     #基準と同じ塩基の数(ref)と、基準と違う塩基の数(alt)がそれぞれわかっていて、リストになっている。

     ref_aomori = [20, 20, 9, 20, 19, 11, 20, 16, 20, 20, 0, 20, 20, 20, 20, 20, 20, 15, 20, 20, 15, 20, 20, 10, 20, 14, 20, 20, 10, 20, 18, 11, 20, 0, 5, 20, 3, 9, 1, 20,
     alt_aomori = [0, 0, 11, 0, 1, 9, 0, 4, 0, 0, 20, 0, 0, 0, 0, 0, 0, 5, 0, 0, 5, 0, 0, 10, 0, 6, 0, 0, 10, 0, 2, 9, 0, 20, 15, 0, 17, 11, 19, 0, 1, 17, 20, 19, 20, 0, 14, 0, 0, 20

     ref_kyoto = [18, 16, 8, 19, 14, 8, 16, 20, 19, 18, 11, 13, 11, 19, 7, 6, 20, 20, 15, 7, 20, 18, 20, 5, 16, 20, 15, 16, 13, 19, 18, 9, 20, 2, 10, 12, 14, 13, 13, 11, 15
     alt_kyoto = [2, 4, 12, 1, 6, 12, 4, 0, 1, 2, 9, 7, 9, 1, 13, 14, 0, 0, 5, 13, 0, 2, 0, 15, 4, 0, 5, 4, 7, 1, 2, 11, 0, 18, 10, 8, 6, 7, 7, 9, 5, 13, 12, 20, 12, 1, 7, 2, 4, 15, 0

     ref_kagoshima = [15, 20, 20, 20, 20, 20, 20, 20, 20, 20, 0, 20, 19, 20, 20, 20, 16, 20, 20, 20, 20, 20, 19, 20, 20, 20, 20, 20, 0, 20, 16, 0, 16, 0, 0, 20, 0, 0, 0, 2
     alt_kagoshima = [5, 0, 0, 0, 0, 0, 0, 0, 0, 0, 20, 0, 1, 0, 0, 0, 4, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 20, 0, 4, 20, 4, 20, 20, 0, 20, 20, 20, 0, 2, 20, 20, 20, 20, 0, 0, 0, 0, 2
```

ref_aomoriとalt_aomoriの多様性(π)を求めるコードを参考にして自作関数を作りましょう

```python
# Nucleotide diversityを求める関数
def diversity(ref, alt):
    nume=0  # 分子

    # (基準と同じ)*(基準と違う)の結果を足していく
    for ref_n, alt_n in zip(ref, alt):
        nume = nume + ref_n*alt_n

    n = ref[0] + alt[0]  # refとaltの最初の要素を足したものが染色体の本数n
    #n = 20                      # この場合、染色体の本数nはどれも同じなので、これでも良い
    pi = nume/(n * (n-1)/2) # πの計算
    return pi                           # πの値を呼び出し元に返す
```

```python
# diversity関数を使って、各集団のNucleotide diversityを求める
pi1 = diversity(ref_aomori, alt_aomori)                 # 青森
pi2 = diversity(ref_kyoto, alt_kyoto)                   # 京都
pi3 = diversity(ref_kagoshima, alt_kagoshima) # 鹿児島
print(pi1)
print(pi2)
print(pi3)
```

３地点の関係(遺伝的な距離)は？

集団間の遺伝的距離の評価方法

$$F_{ST} = \frac{\pi_{12} - (\pi_1 + \pi_2) / 2}{\pi_{12}}$$

$$F_{ST} = \frac{\pi_{12} - (\pi_1 + \pi_2)/2}{\pi_{12}}$$

集団１

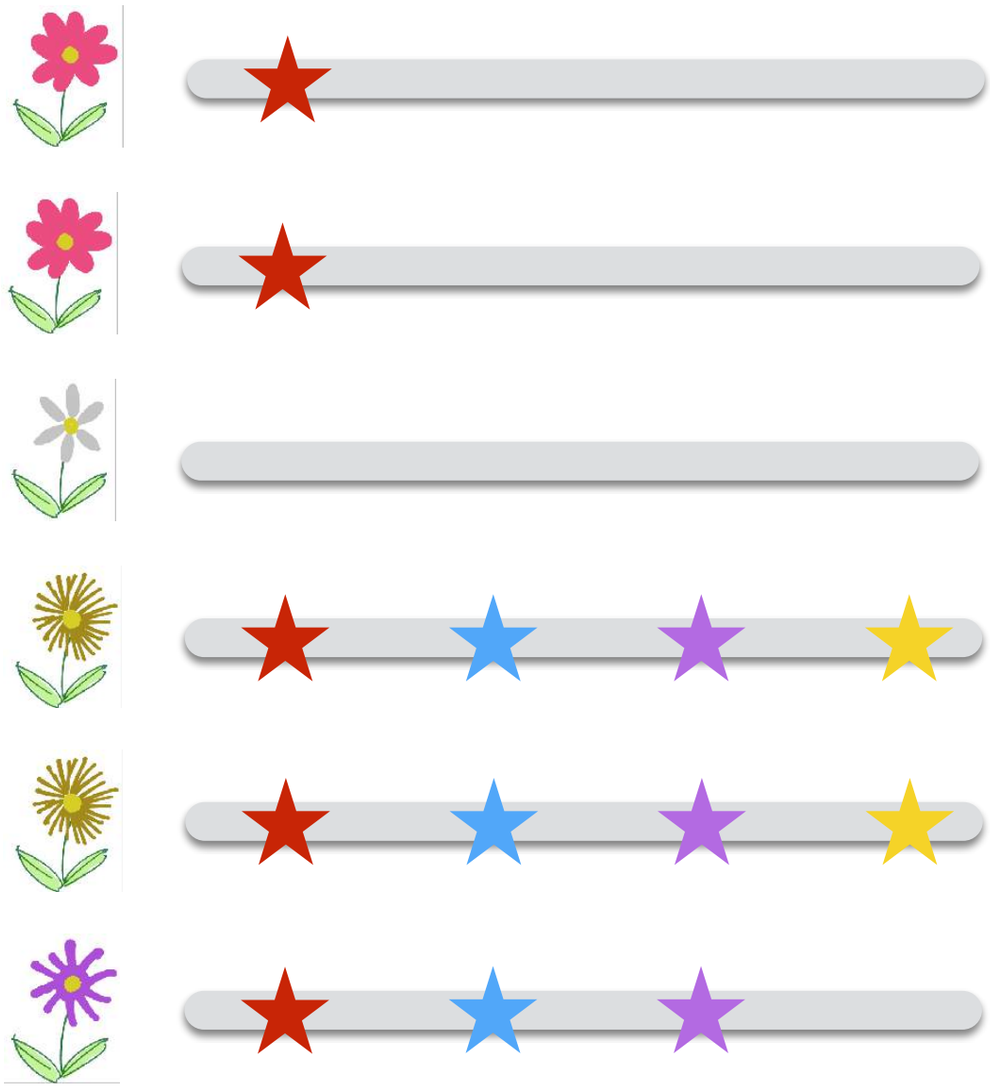
$$\pi_1 = \frac{2 \times 1}{3} = \frac{2}{3}$$

集団２

$$\pi_2 = \frac{2 \times 1}{3} = \frac{2}{3}$$

$$F_{ST} = \frac{\pi_{12} - (\pi_1 + \pi_2)/2}{\pi_{12}}$$

集団１と集団２が同じ集団だと仮定する



$$\pi_{12} = \frac{5\times1 + 3\times3 + 3\times3 + 4\times2}{{}_6C_2} = \frac{31}{15}$$

集団間の遺伝的距離の評価方法

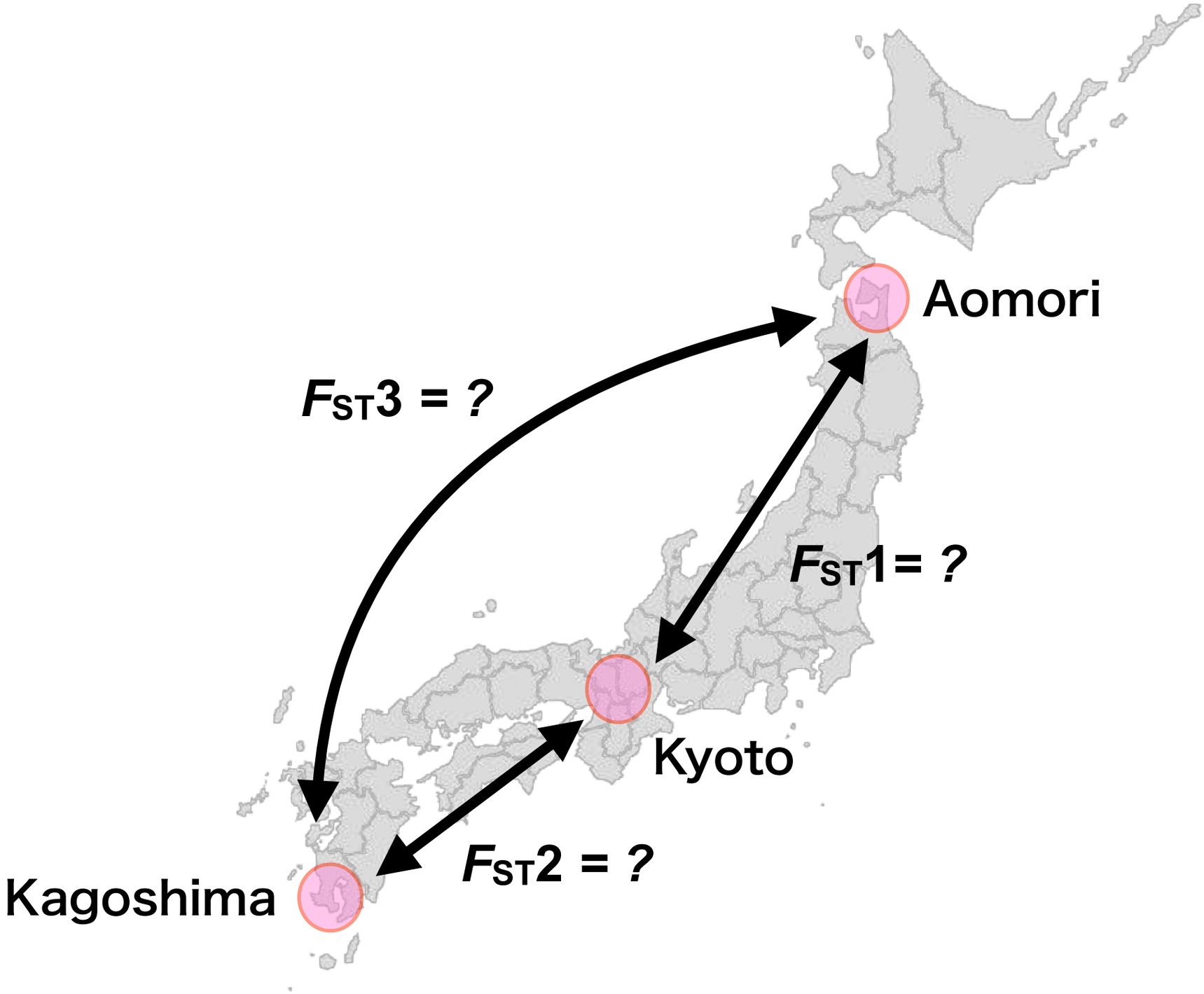
$$F_{ST} = \frac{\pi_{12} - (\pi_1 + \pi_2)/2}{\pi_{12}}$$

$$= \frac{\frac{31}{15} - (\frac{2}{3} + \frac{2}{3})/2}{\frac{31}{15}}$$

$$= \frac{21}{31}$$

F_{ST}は０～１の間の値をとる指標で、およそ0.4より大きいと高いと言われています。

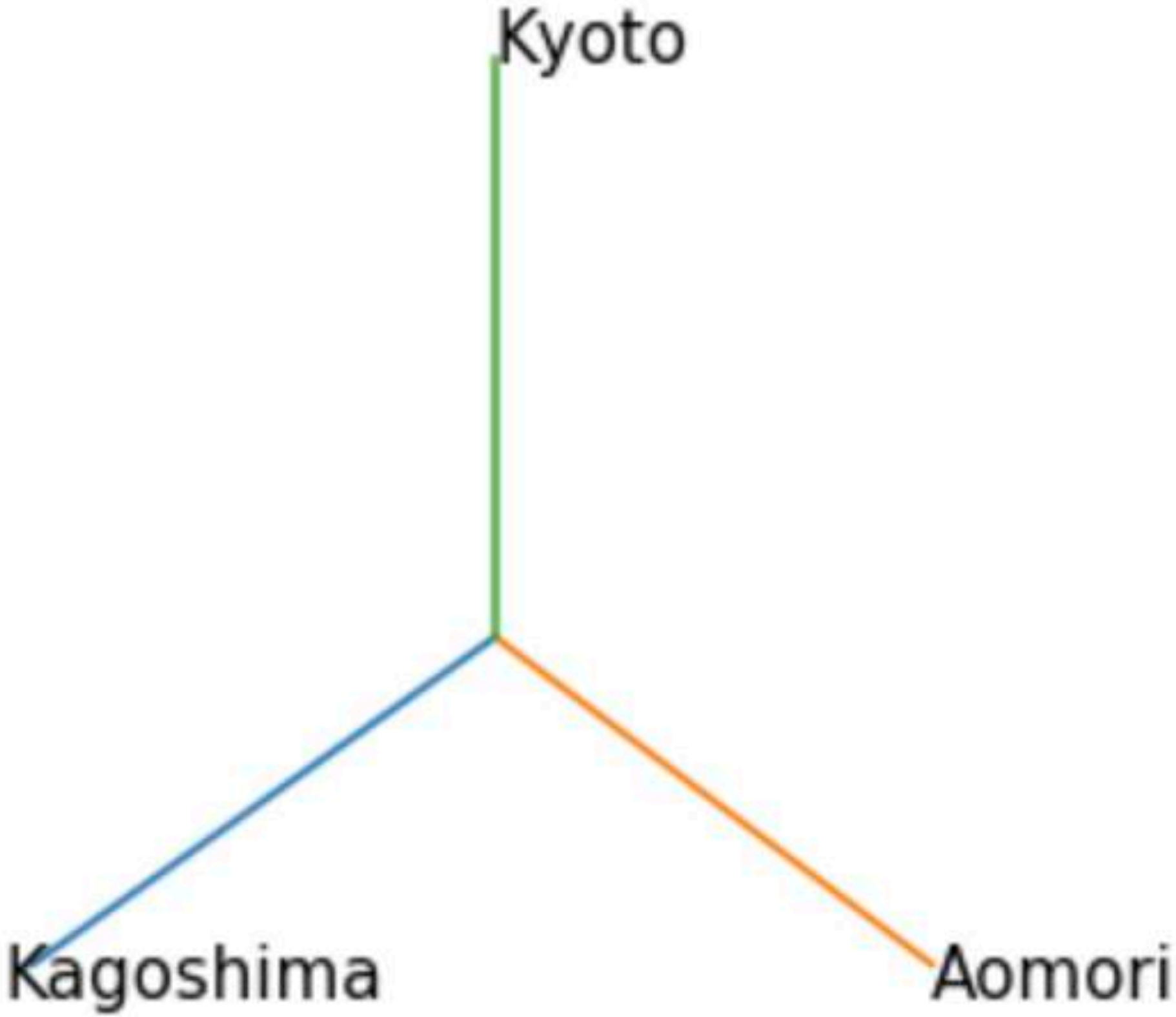
実際に計算してみましょう



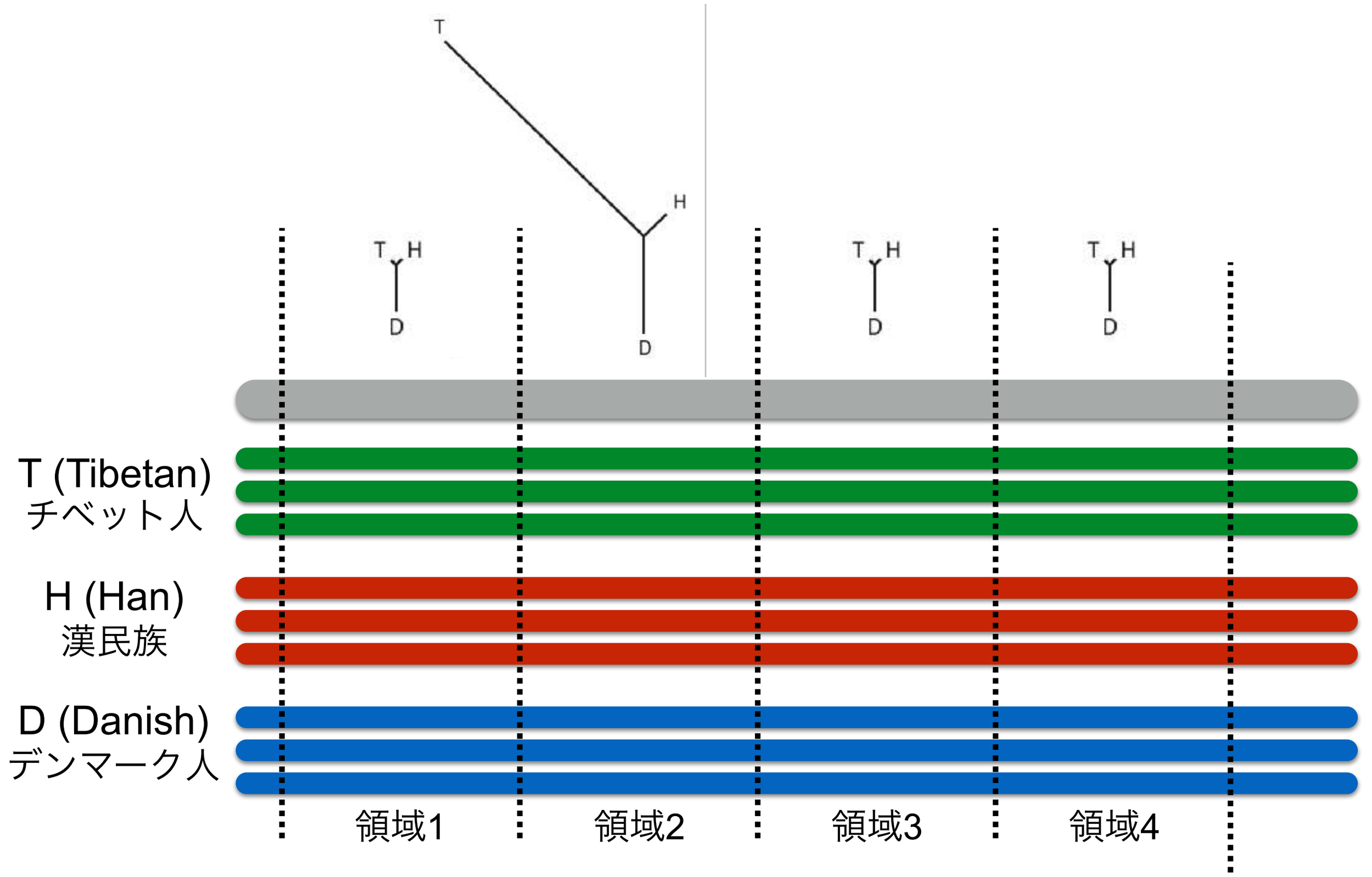
京都と青森の間の
遺伝的距離(Fst)を求めましょう

求められたら関数を作って
残りも計算しましょう

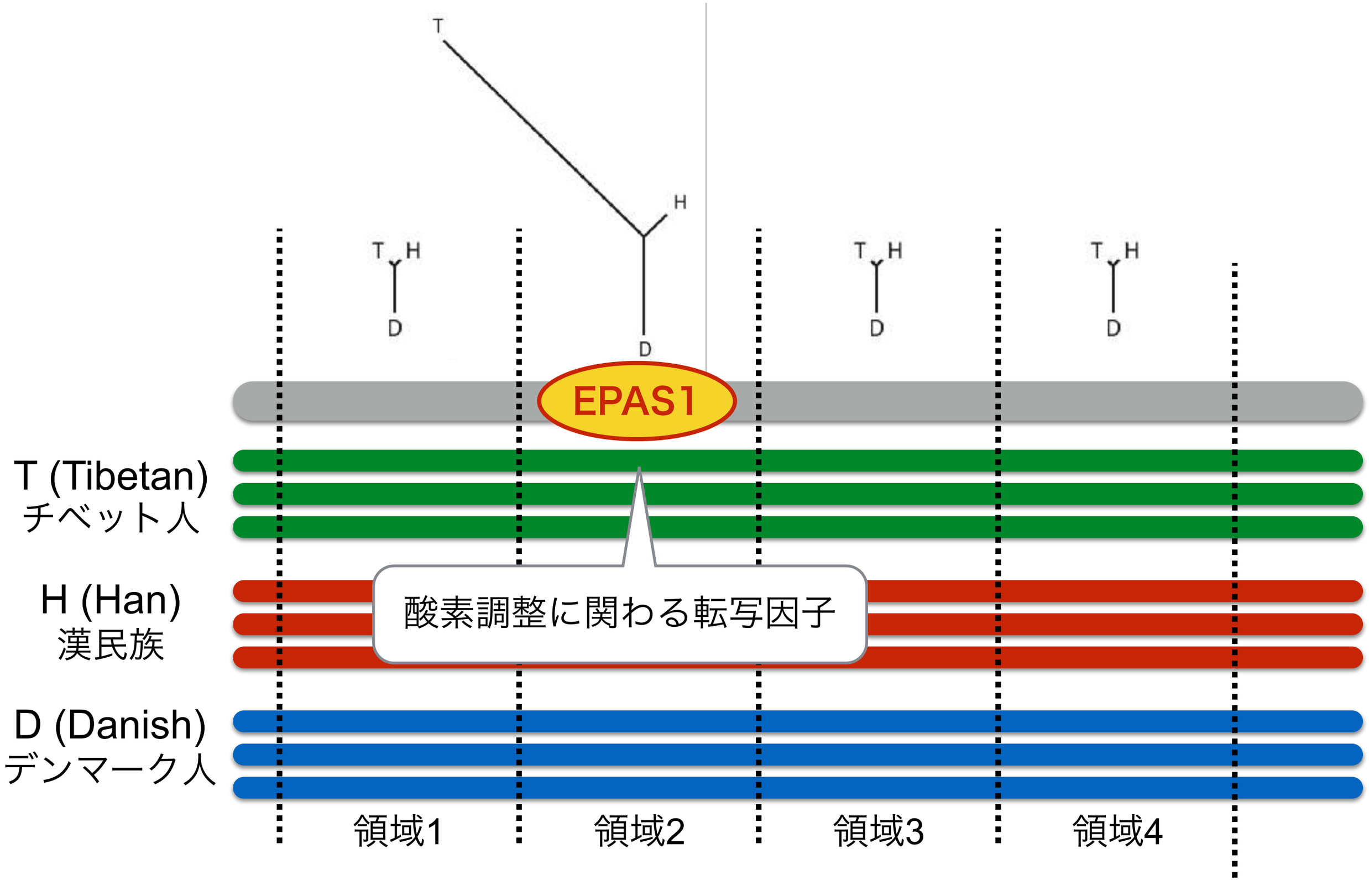
```
# 青森と京都の間のFstを求める

# === ref12を作成する ===
# 2つの集団を合わせた場合のrefを入れるリストを準備する
ref12 = []

# すべての要素（塩基座位）について和を求める
for ref_n1, ref_n2 in zip(ref_aomori, ref_kyoto):
  ref12.append(ref_n1 + ref_n2) # 要素同士の和をリストに追加する

# === alt12を作成する ===
alt12 = []
for alt_n1, alt_n2 in zip(alt_aomori, alt_kyoto):
  alt12.append(alt_n1 + alt_n2)

# === 各πを求める===
# 実習3で作ったdiversity関数を用いて、青森集団のpi1を求める
pi1 = diversity(ref_aomori, alt_aomori)

# 同様に、京都集団のpi2と、青森と京都を合体した集団のpi12を求める
pi2 = diversity(ref_kyoto, alt_kyoto)
pi12 = diversity(ref12, alt12)

# === Fstを計算する ===
fst = (pi12-(pi1+pi2)/2)/pi12
print(fst)
```


Kyoto
Kagoshima
Aomori

F_{ST}を用いてのゲノムスキャン



F_{ST}を用いてのゲノムスキャン



課題

Exercise1, 2, 3

期限：2019年11月11日(月)