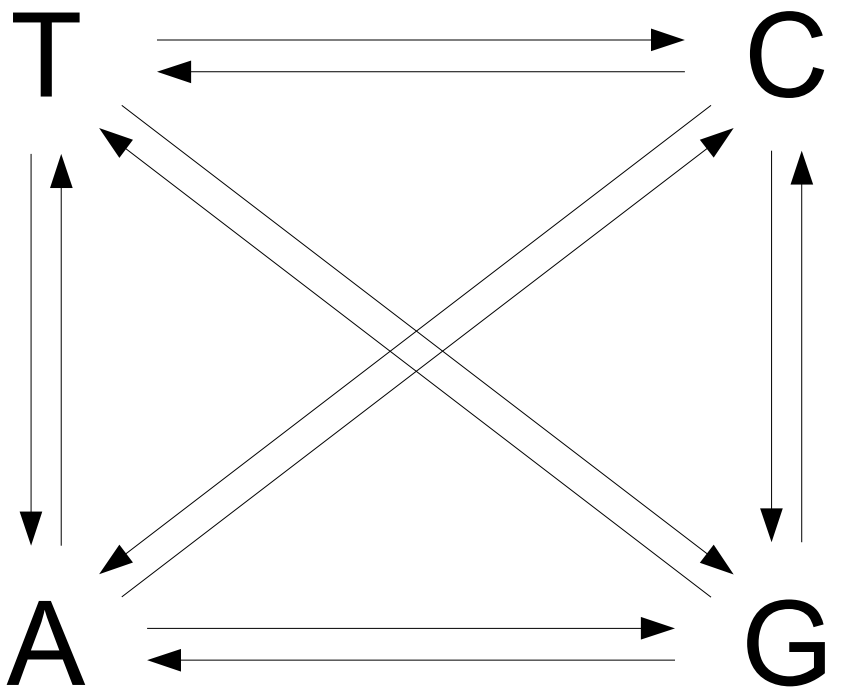
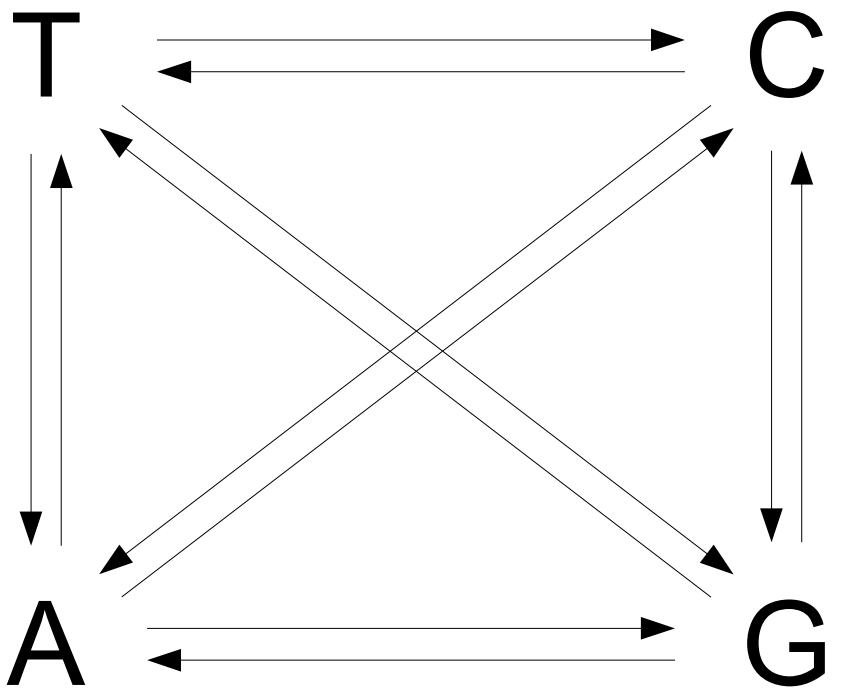
# 分子進化の統計モデリング
# とモデル選択

講義編

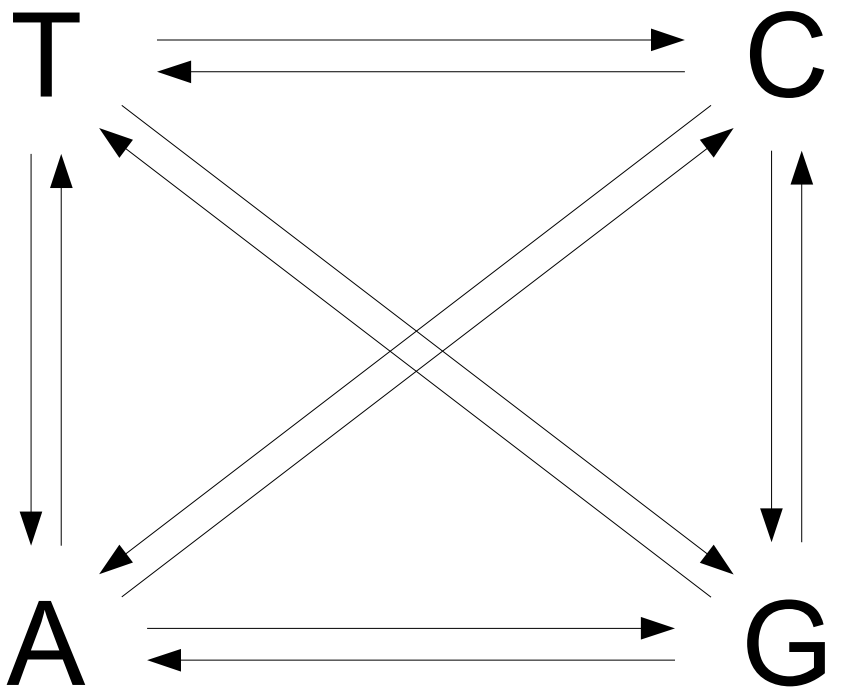
田辺晶史

# 分子進化の統計モデリング


T
C
A
G




- 塩基の置換パターンは 12 通り




- 塩基の置換パターンは 12 通り
- それぞれの起きる速度の違いをモデル化する必要がある

# 塩基置換速度行列

時間反転可能モデル

| From \ To | A | C | G | T |
|---|---|---|---|---|
| A | | $r_{AC}\pi_C$ | $r_{AG}\pi_G$ | $r_{AT}\pi_T$ |
| C | $r_{AC}\pi_A$ | | $r_{CG}\pi_G$ | $r_{CT}\pi_T$ |
| G | $r_{AG}\pi_A$ | $r_{CG}\pi_C$ | | $r_{GT}\pi_G$ |
| T | $r_{AT}\pi_A$ | $r_{CT}\pi_C$ | $r_{GT}\pi_G$ | |

時間反転不能モデル

| From \ To | A | C | G | T |
|---|---|---|---|---|
| A | | $r_{AC}$ | $r_{AG}$ | $r_{AT}$ |
| C | $r_{CA}$ | | $r_{CG}$ | $r_{CT}$ |
| G | $r_{GA}$ | $r_{GC}$ | | $r_{GT}$ |
| T | $r_{TA}$ | $r_{TC}$ | $r_{TG}$ | |

時間反転可能モデル

| From \ To | A | C | G | T |
|---|---|---|---|---|
| A | | $r_{AC}\pi_C$ | $r_{AG}\pi_G$ | $r_{AT}\pi_T$ |
| C | $r_{AC}\pi_A$ | | $r_{CG}\pi_G$ | $r_{CT}\pi_T$ |
| G | $r_{AG}\pi_A$ | $r_{CG}\pi_C$ | | $r_{GT}\pi_G$ |
| T | $r_{AT}\pi_A$ | $r_{CT}\pi_C$ | $r_{GT}\pi_G$ | |

時間反転不能モデル

| From \ To | A | C | G | T |
|---|---|---|---|---|
| A | | $r_{AC}$ | $r_{AG}$ | $r_{AT}$ |
| C | $r_{CA}$ | | $r_{CG}$ | $r_{CT}$ |
| G | $r_{GA}$ | $r_{GC}$ | | $r_{GT}$ |
| T | $r_{TA}$ | $r_{TC}$ | $r_{TG}$ | |

- 各マスは塩基 X から塩基 Y への置換速度

時間反転可能モデル

| From \ To | A | C | G | T |
|---|---|---|---|---|
| A | | $r_{AC}\pi_C$ | $r_{AG}\pi_G$ | $r_{AT}\pi_T$ |
| C | $r_{AC}\pi_A$ | | $r_{CG}\pi_G$ | $r_{CT}\pi_T$ |
| G | $r_{AG}\pi_A$ | $r_{CG}\pi_C$ | | $r_{GT}\pi_G$ |
| T | $r_{AT}\pi_A$ | $r_{CT}\pi_C$ | $r_{GT}\pi_G$ | |

時間反転不能モデル

| From \ To | A | C | G | T |
|---|---|---|---|---|
| A | | $r_{AC}$ | $r_{AG}$ | $r_{AT}$ |
| C | $r_{CA}$ | | $r_{CG}$ | $r_{CT}$ |
| G | $r_{GA}$ | $r_{GC}$ | | $r_{GT}$ |
| T | $r_{TA}$ | $r_{TC}$ | $r_{TG}$ | |

- 各マスは塩基 X から塩基 Y への置換速度
- $\pi_X$ は塩基 X の頻度

時間反転可能モデル

| From \ To | A | C | G | T |
|---|---|---|---|---|
| A | | $r_{AC}\pi_C$ | $r_{AG}\pi_G$ | $r_{AT}\pi_T$ |
| C | $r_{AC}\pi_A$ | | $r_{CG}\pi_G$ | $r_{CT}\pi_T$ |
| G | $r_{AG}\pi_A$ | $r_{CG}\pi_C$ | | $r_{GT}\pi_G$ |
| T | $r_{AT}\pi_A$ | $r_{CT}\pi_C$ | $r_{GT}\pi_G$ | |

時間反転不能モデル

| From \ To | A | C | G | T |
|---|---|---|---|---|
| A | | $r_{AC}$ | $r_{AG}$ | $r_{AT}$ |
| C | $r_{CA}$ | | $r_{CG}$ | $r_{CT}$ |
| G | $r_{GA}$ | $r_{GC}$ | | $r_{GT}$ |
| T | $r_{TA}$ | $r_{TC}$ | $r_{TG}$ | |

- 各マスは塩基 X から塩基 Y への置換速度
- $\pi_X$ は塩基 X の頻度
- $r_{XY}=r_{YX}$ なモデルを時間反転可能という

時間反転可能モデル

| From \ To | A | C | G | T |
|---|---|---|---|---|
| A | | $r_{AC}\pi_C$ | $r_{AG}\pi_G$ | $r_{AT}\pi_T$ |
| C | $r_{AC}\pi_A$ | | $r_{CG}\pi_G$ | $r_{CT}\pi_T$ |
| G | $r_{AG}\pi_A$ | $r_{CG}\pi_C$ | | $r_{GT}\pi_G$ |
| T | $r_{AT}\pi_A$ | $r_{CT}\pi_C$ | $r_{GT}\pi_G$ | |

時間反転不能モデル

| From \ To | A | C | G | T |
|---|---|---|---|---|
| A | | $r_{AC}$ | $r_{AG}$ | $r_{AT}$ |
| C | $r_{CA}$ | | $r_{CG}$ | $r_{CT}$ |
| G | $r_{GA}$ | $r_{GC}$ | | $r_{GT}$ |
| T | $r_{TA}$ | $r_{TC}$ | $r_{TG}$ | |

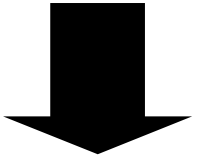
- 各マスは塩基 X から塩基 Y への置換速度
- $\pi_X$ は塩基 X の頻度
- $r_{XY}=r_{YX}$ なモデルを時間反転可能という
- ほとんどの系統樹推定では時間反転可能モデルを用いる

# アミノ酸置換速度行列

- 20x20 の置換速度行列

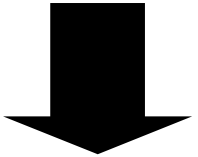
- 20x20 の置換速度行列
- 189 の $r_{XY}$ と 19 の $\pi_X$ をパラメータとして持つ

- 20x20 の置換速度行列
- 189 の $r_{XY}$ と 19 の $\pi_X$ をパラメータとして持つ

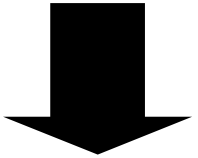


計算困難

- 20x20 の置換速度行列
- 189 の $r_{XY}$ と 19 の $\pi_X$ をパラメータとして持つ

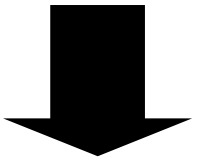


計算困難

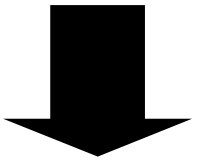


- Empirical Model

- 20x20 の置換速度行列
- 189 の $r_{XY}$ と 19 の $\pi_X$ をパラメータとして持つ

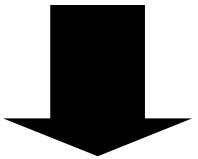


計算困難

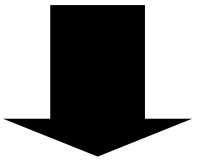


- Empirical Model
  - 既知の系統樹と大量のデータから推定した値に固定

- 20x20 の置換速度行列
- 189 の $r_{XY}$ と 19 の $\pi_X$ をパラメータとして持つ

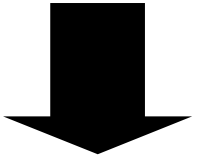


計算困難

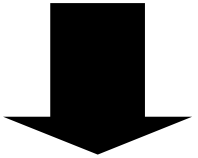


- Empirical Model
  - 既知の系統樹と大量のデータから推定した値に固定
  - データから推定するパラメータ数は 0

- 20x20 の置換速度行列
- 189 の $r_{XY}$ と 19 の $\pi_X$ をパラメータとして持つ

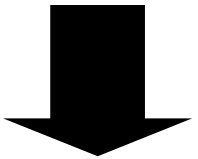


計算困難

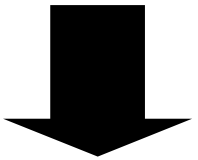


- Empirical Model
  - 既知の系統樹と大量のデータから推定した値に固定
  - データから推定するパラメータ数は 0
- +F モデル

- 20x20 の置換速度行列
- 189 の $r_{XY}$ と 19 の $\pi_X$ をパラメータとして持つ



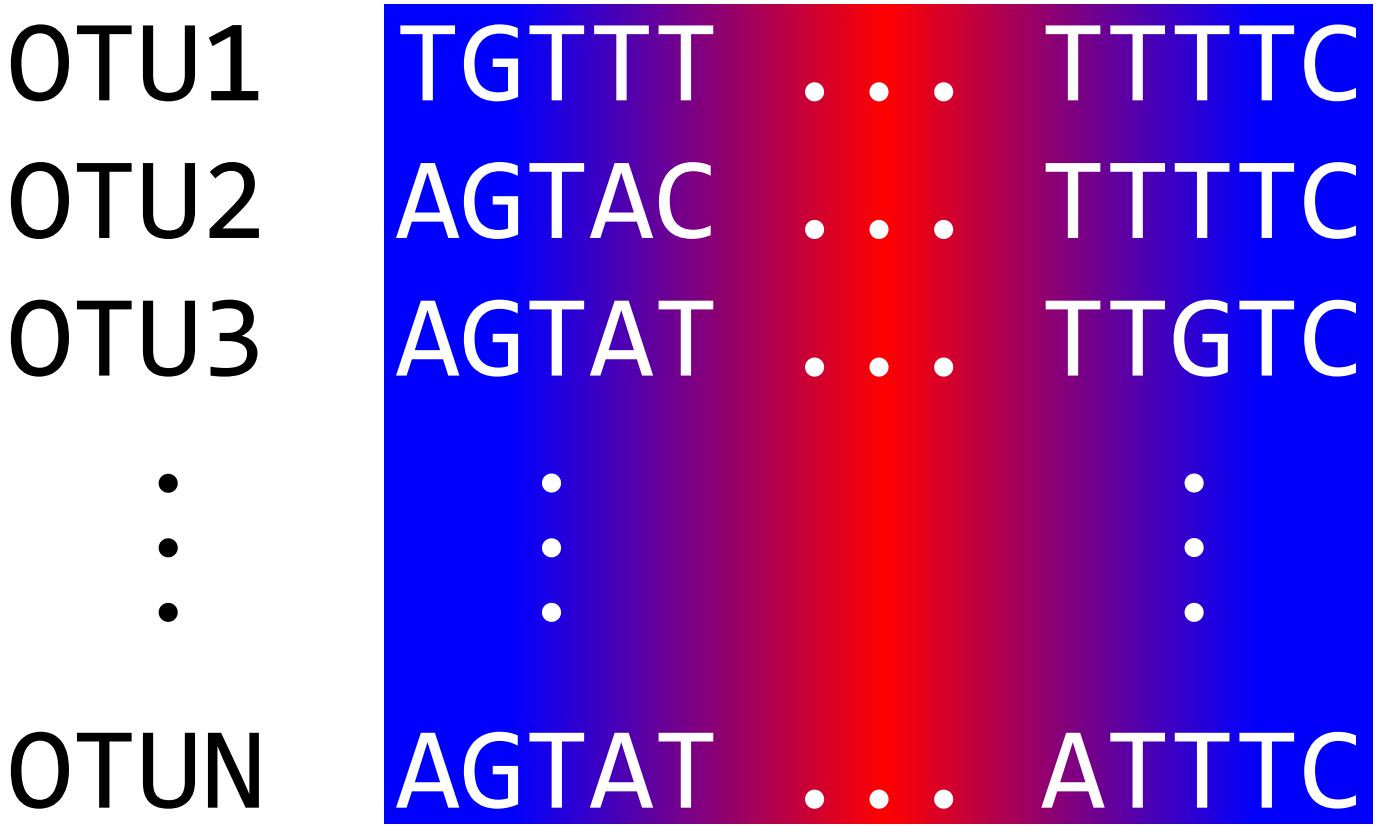
計算困難

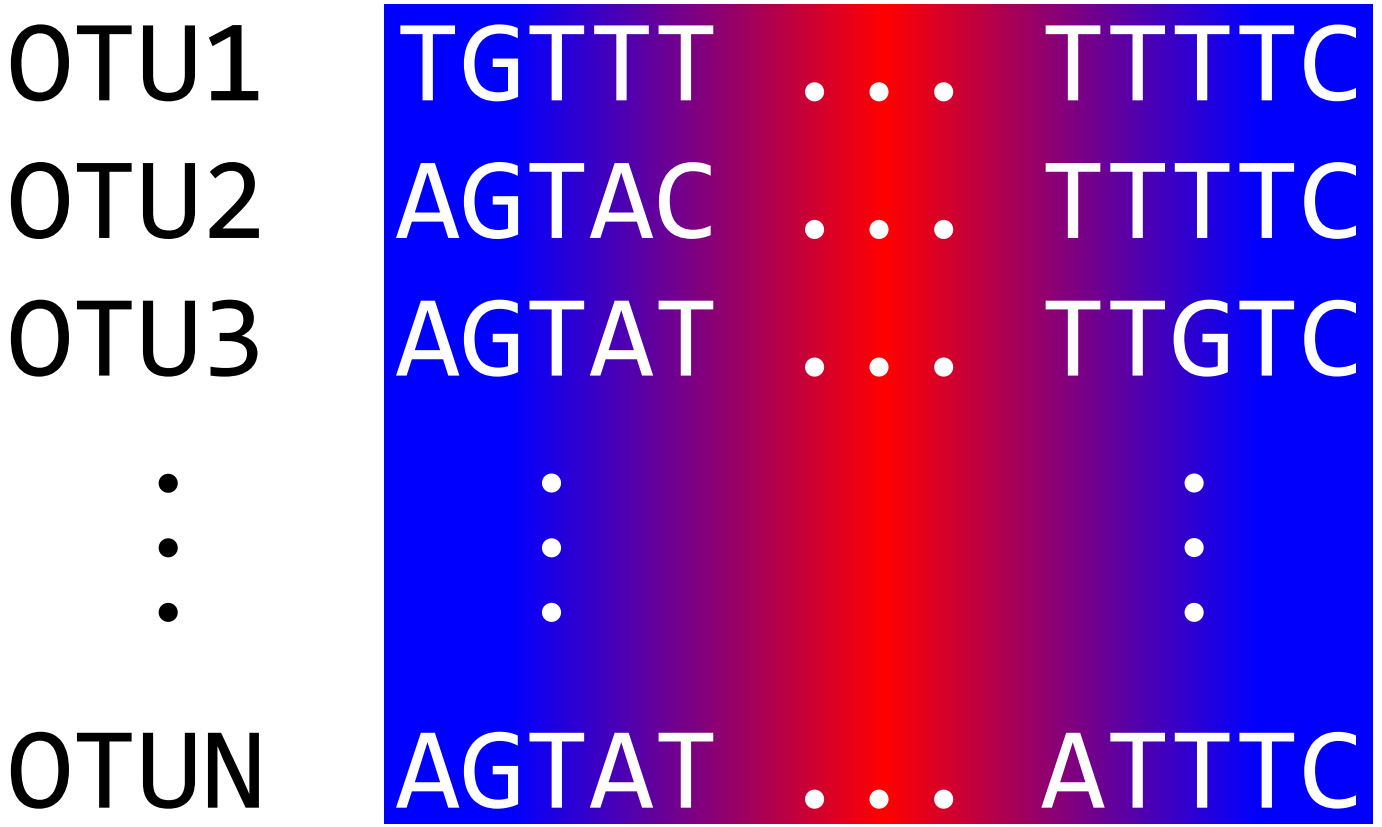


- Empirical Model
  - 既知の系統樹と大量のデータから推定した値に固定
  - データから推定するパラメータ数は 0
- +F モデル
  - $\pi_X$ だけはデータから推定する

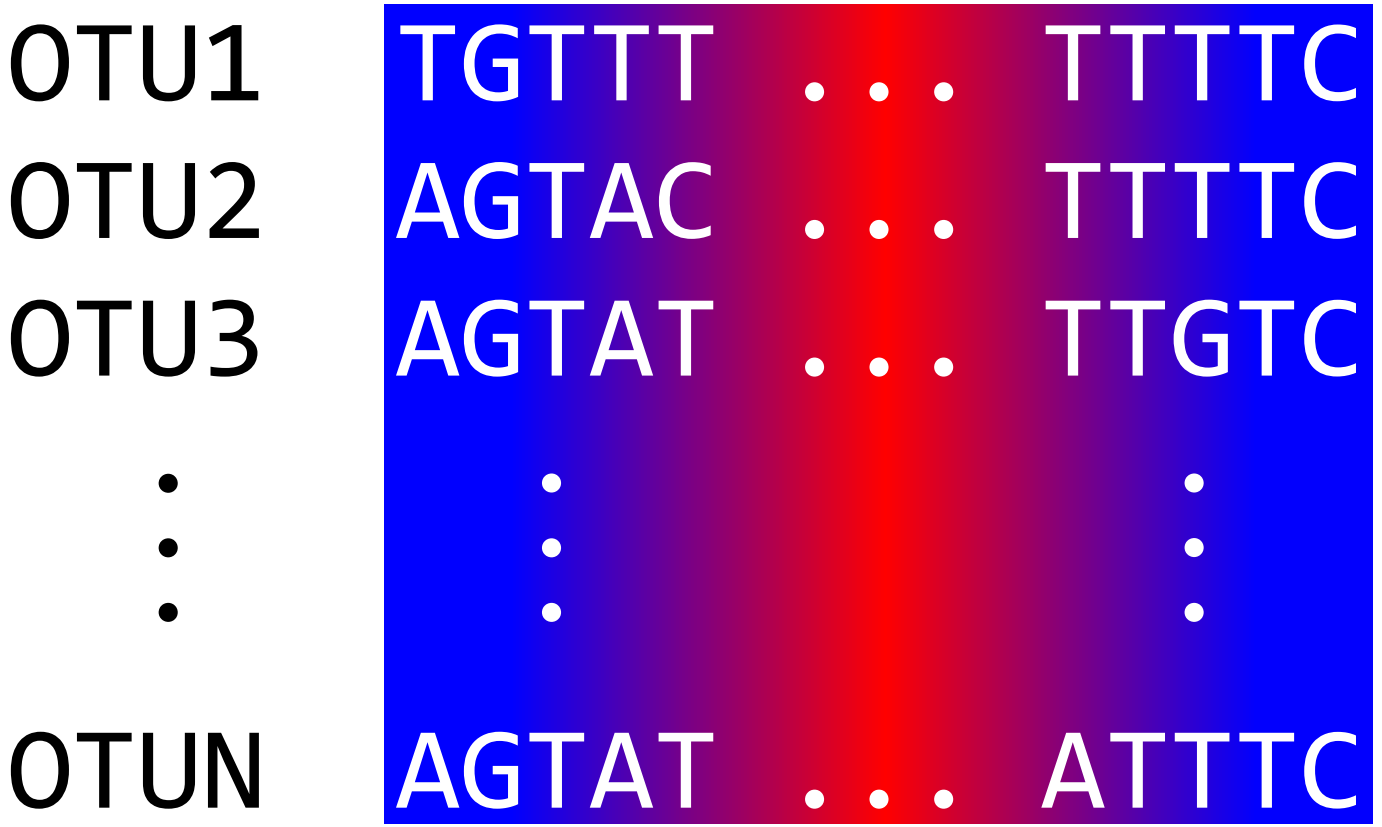
# 座位間の置換速度不均質性

| | | | |
|---|---|---|---|
| OTU1 | TGTTT | ... | TTTTC |
| OTU2 | AGTAC | ... | TTTTC |
| OTU3 | AGTAT | ... | TTGTC |
| ⋮ | ⋮ | | ⋮ |
| OTUN | AGTAT | ... | ATTTC |


OTU1 TGTTT ... TTTTC
OTU2 AGTAC ... TTTTC
OTU3 AGTAT ... TTGTC
OTUN AGTAT ... ATTTC




- 変異の多い＝置換の速い座位とそうでない座位がある

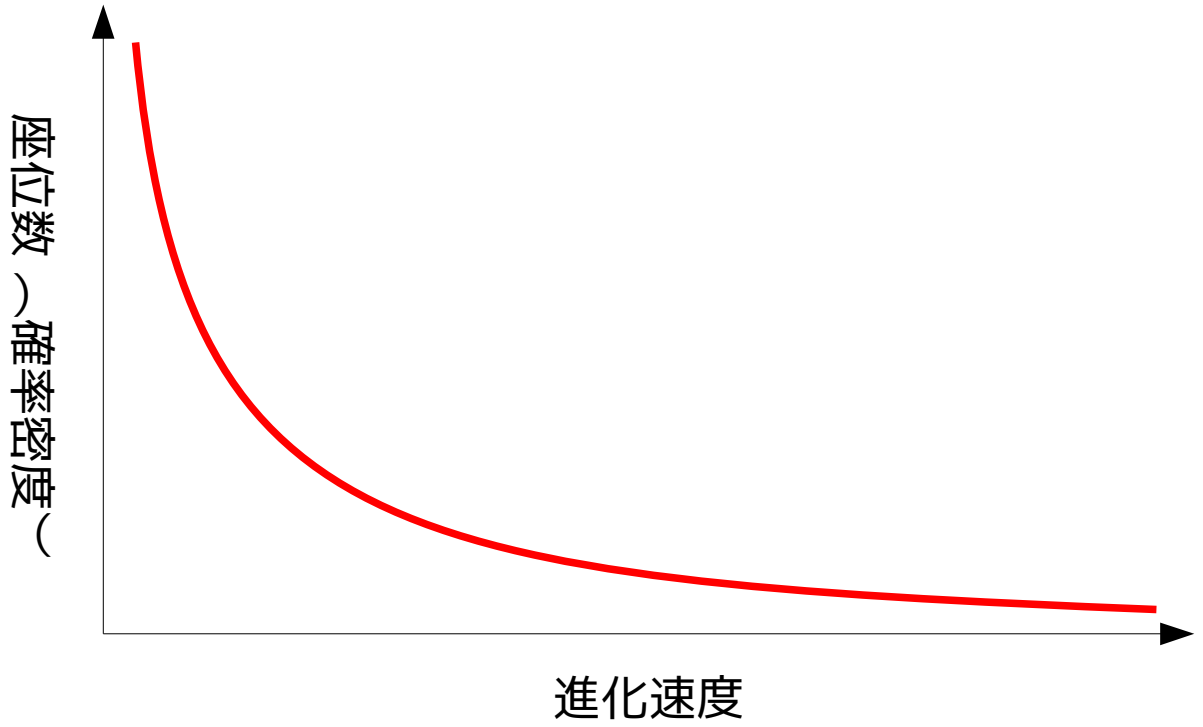



- 変異の多い＝置換の速い座位とそうでない座位がある
- 座位間の置換速度の違いをモデル化する必要がある

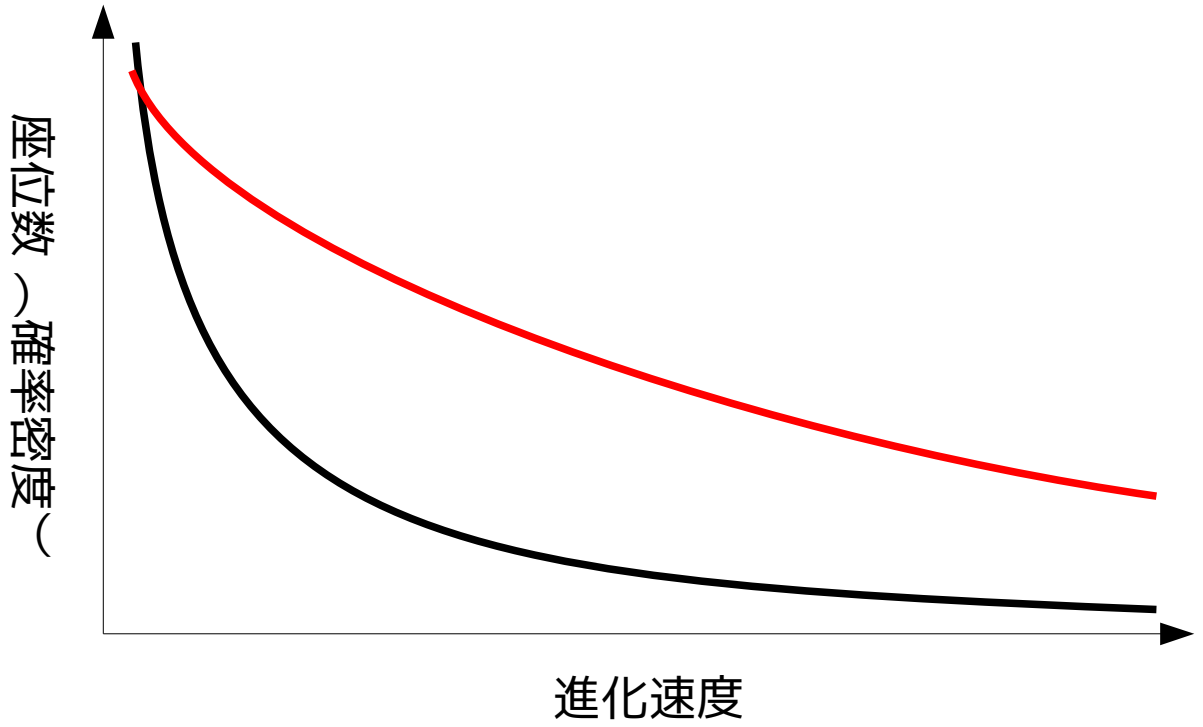
- 置換速度は連続量

- 置換速度は連続量
- 確率密度分布は非対称

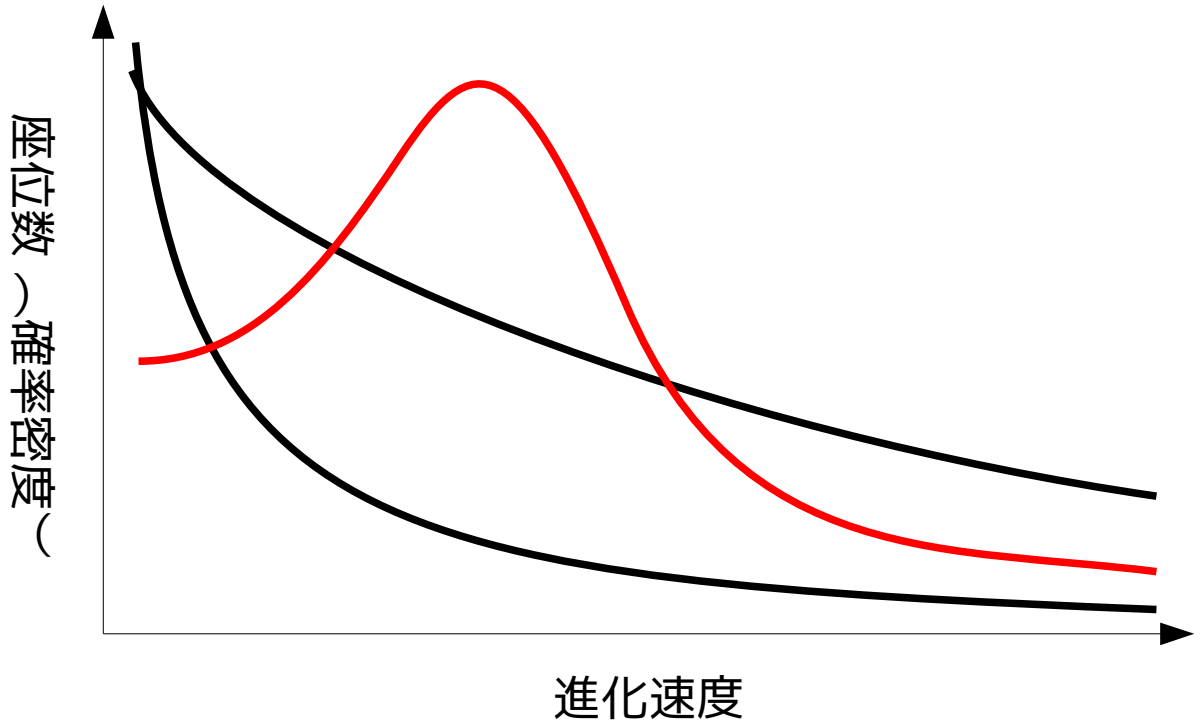

- 進化速度は連続量
- 確率密度分布は非対称



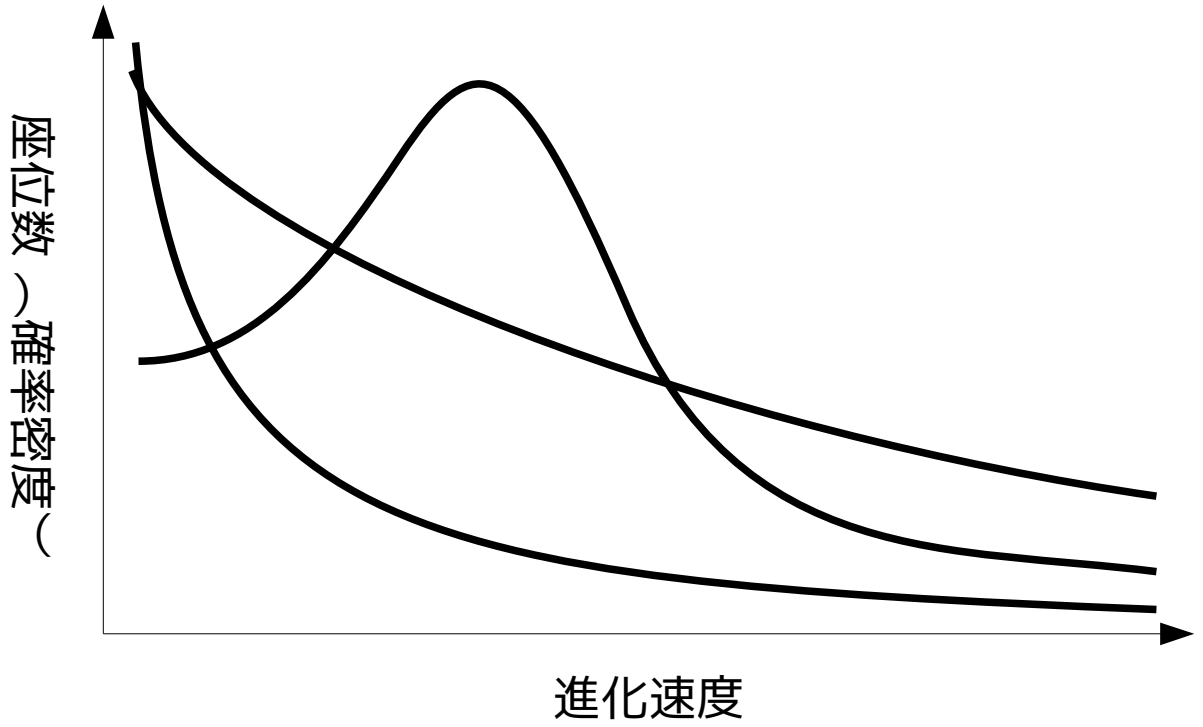

- 進化速度は連続量
- 確率密度分布は非対称




- 進化速度は連続量
- 確率密度分布は非対称




- 進化速度は連続量
- 確率密度分布は非対称




- 進化速度は連続量
- 確率密度分布は非対称

Γ 分布によるモデル化

# 複数遺伝子データの場合

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| OTU1 | TGTTTTCTTT | ... | TTTTC | OTU1 | TCCTCTACTA | ... | AGCTA |
| OTU2 | AGTATTCTTC | ... | TTTTC | OTU2 | TCTCCTACTA | ... | AGCTA |
| OTU3 | AGTATTCTTT | ... | TTTTC | OTU3 | TCTACTACTA | ... | AGCTA |
| OTU4 | AATATTTTTT | ... | TTTTC | OTU4 | TCTTCCACTA | ... | AGTTA |
| OTU5 | AGTATTTTTT | ... | TTTTC | OTU5 | TCTGCCGCTA | ... | AGTTA |
| OTU6 | AGTATTCTCT | ... | TTTCC | OTU6 | TCTTTCACTA | ... | AGTCA |
| OTU7 | AGTATTCTTT | ... | TTTTC | OTU7 | TCTTTTACTA | ... | AGTCA |
| OTU8 | AGTATTTTTT | ... | TTTTC | OTU8 | TTTCCCGCTG | ... | AGCCA |
| OTU9 | AGTATTCTTT | ... | TTTTC | OTU9 | ATTTCCACTG | ... | AGCCA |

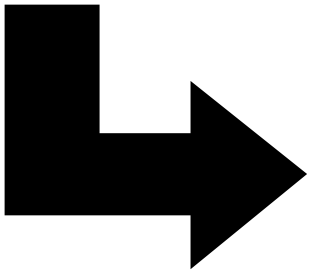
```
OTU1 TGTTTTCTTT ... TTTTC    OTU1 TCCTCTACTA ... AGCTA
OTU2 AGTATTCTTC ... TTTTC    OTU2 TCTCCTACTA ... AGCTA
OTU3 AGTATTCTTT ... TTTTC    OTU3 TCTACTACTA ... AGCTA
OTU4 AATATTTTTT ... TTTTC    OTU4 TCTTCCACTA ... AGTTA
OTU5 AGTATTTTTT ... TTTTC    OTU5 TCTGCCGCTA ... AGTTA
OTU6 AGTATTCTCT ... TTTCC    OTU6 TCTTTCACTA ... AGTCA
OTU7 AGTATTCTTT ... TTTTC    OTU7 TCTTTTACTA ... AGTCA
OTU8 AGTATTTTTT ... TTTTC    OTU8 TTTCCCGCTG ... AGCCA
OTU9 AGTATTCTTT ... TTTTC    OTU9 ATTTCCACTG ... AGCCA
```

- 複数の遺伝子配列のそれぞれに，異なる塩基置換速度行列と座位間の置換速度不均質性を当てはめたい

```
OTU1 TGTTTTCTTT ... TTTTC
OTU2 AGTATTCTTC ... TTTTC
OTU3 AGTATTCTTT ... TTTTC
OTU4 AATATTTTTT ... TTTTC
OTU5 AGTATTTTTT ... TTTTC
OTU6 AGTATTCTCT ... TTTCC
OTU7 AGTATTCTTT ... TTTTC
OTU8 AGTATTTTTT ... TTTTC
OTU9 AGTATTCTTT ... TTTTC
```

```
OTU1 TCCTCTACTA ... AGCTA
OTU2 TCTCCTACTA ... AGCTA
OTU3 TCTACTACTA ... AGCTA
OTU4 TCTTCCACTA ... AGTTA
OTU5 TCTGCCGCTA ... AGTTA
OTU6 TCTTTCACTA ... AGTCA
OTU7 TCTTTTACTA ... AGTCA
OTU8 TTTCCCGCTG ... AGCCA
OTU9 ATTTCCACTG ... AGCCA
```

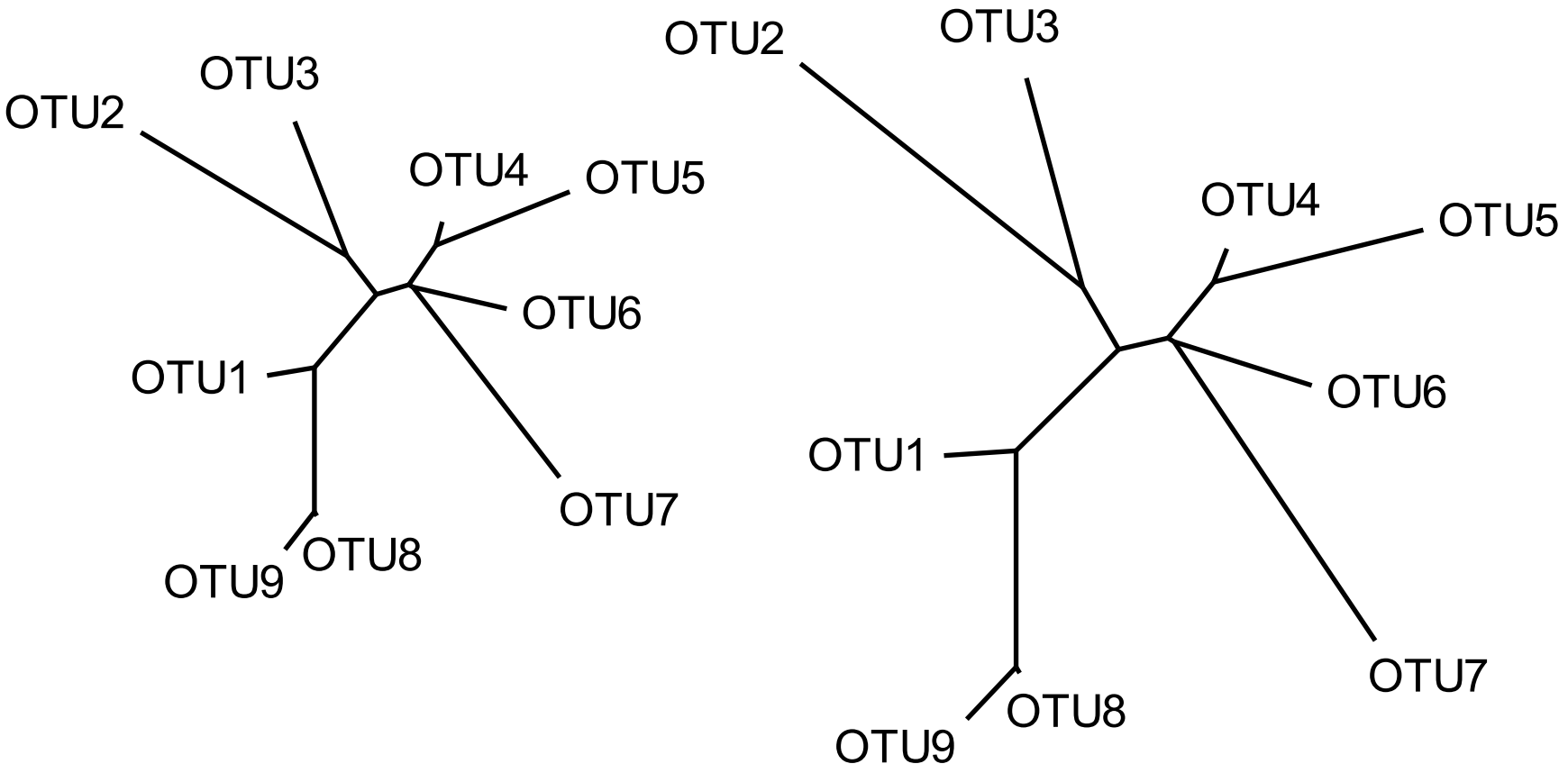
- 複数の遺伝子配列のそれぞれに，異なる塩基置換速度行列と座位間の置換速度不均質性を当てはめたい



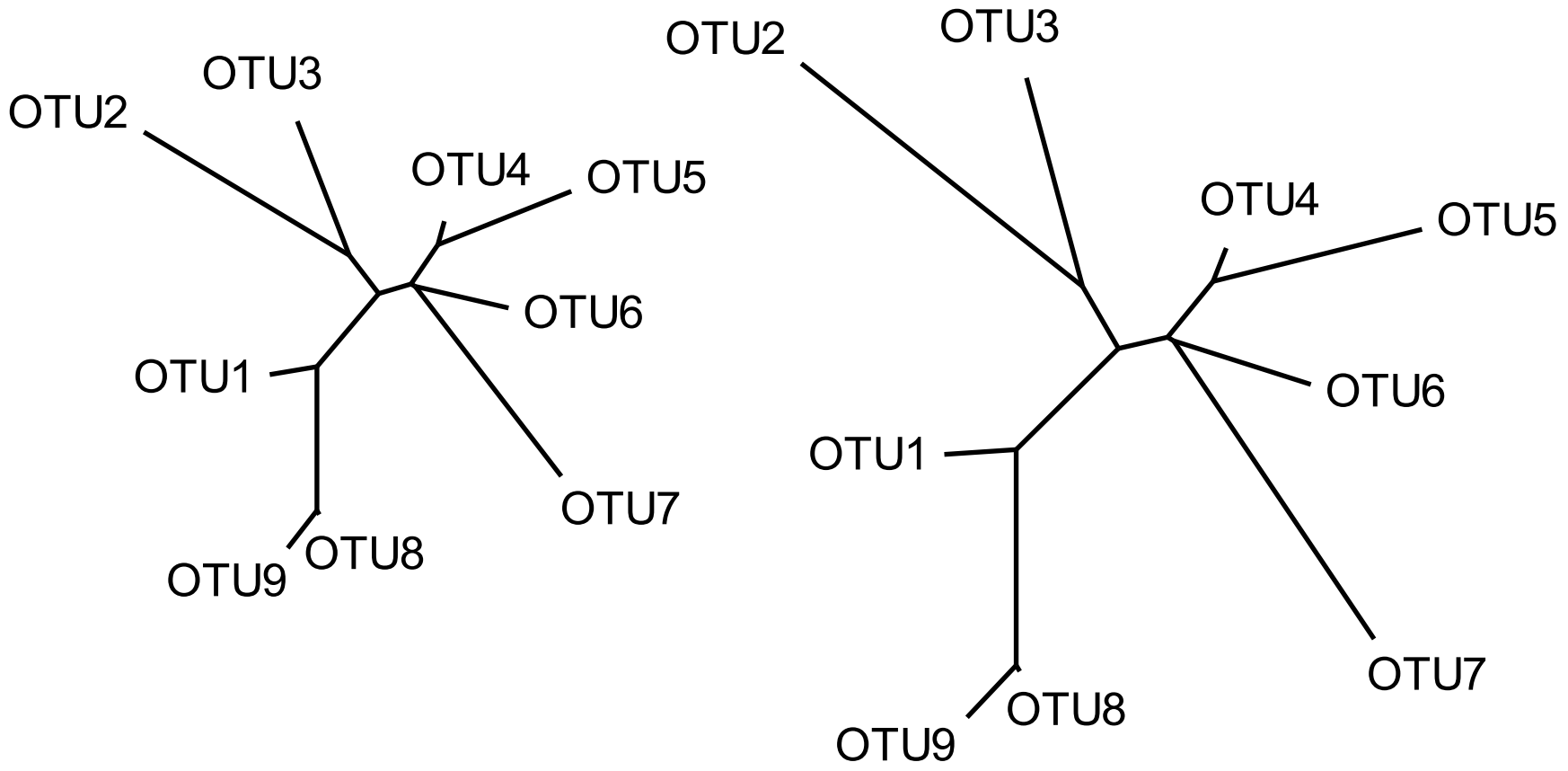
比例モデルと
分離モデル

# 比例モデル

# 比例モデル


OTU2
OTU3
OTU4
OTU5
OTU6
OTU1
OTU7
OTU9
OTU8
OTU2
OTU3
OTU4
OTU5
OTU6
OTU1
OTU7
OTU9
OTU8

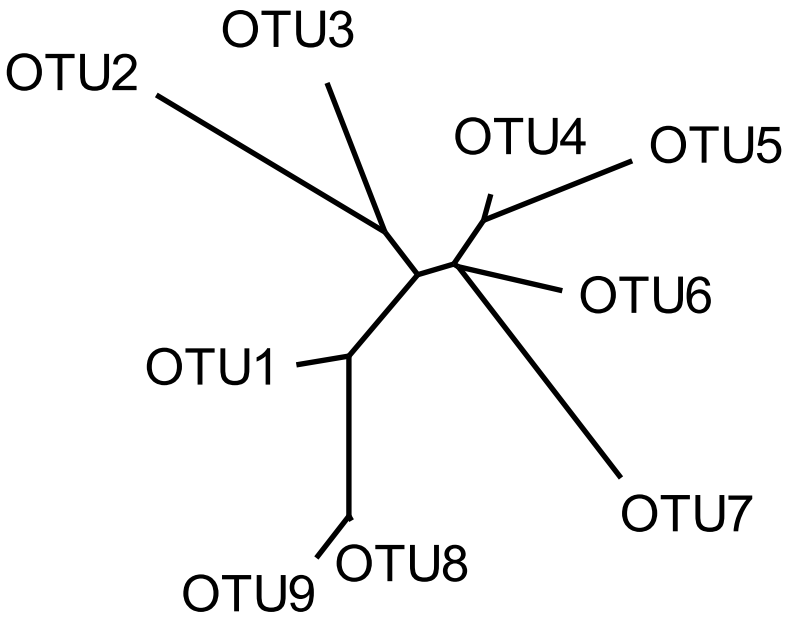
# 比例モデル

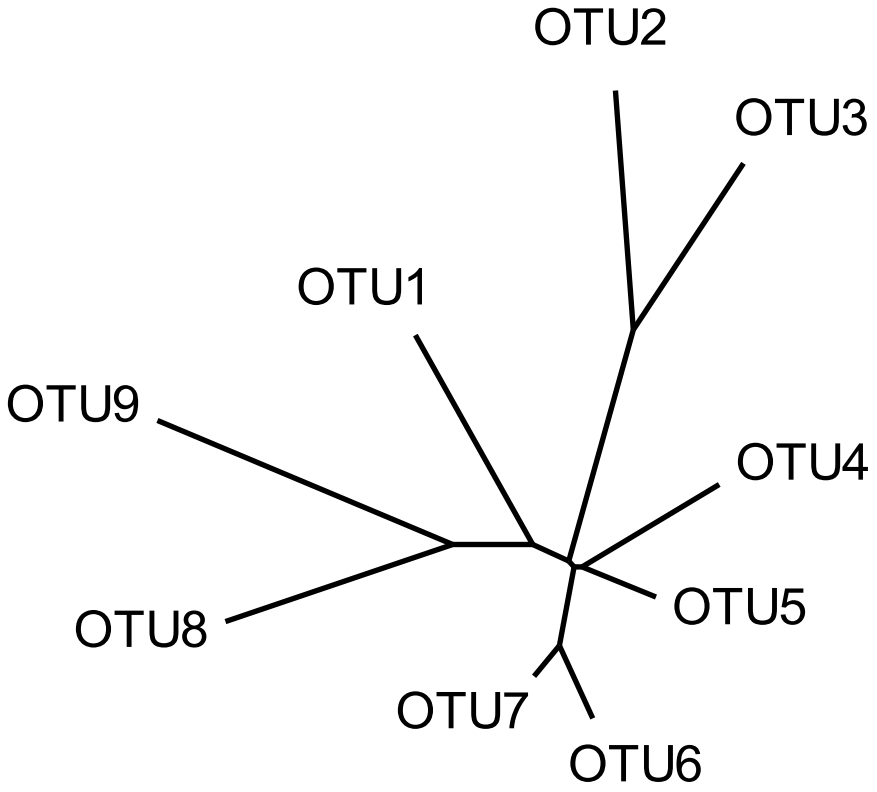



- 系統樹は相似形＝枝長比（置換速度比）が系統樹上で一定と仮定

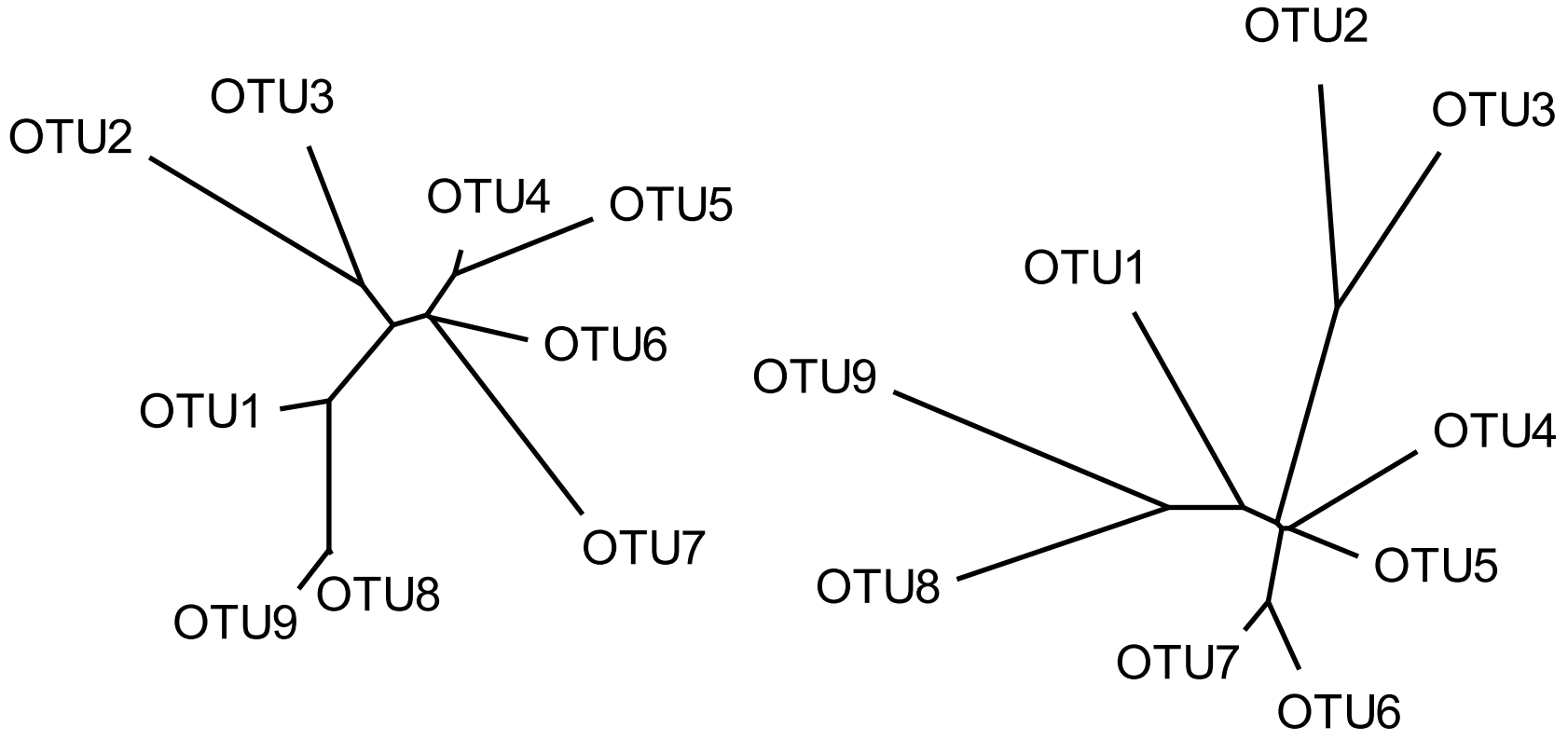
# 分離モデル

# 分離モデル


OTU2
OTU3
OTU4
OTU5
OTU6
OTU1
OTU7
OTU9
OTU8

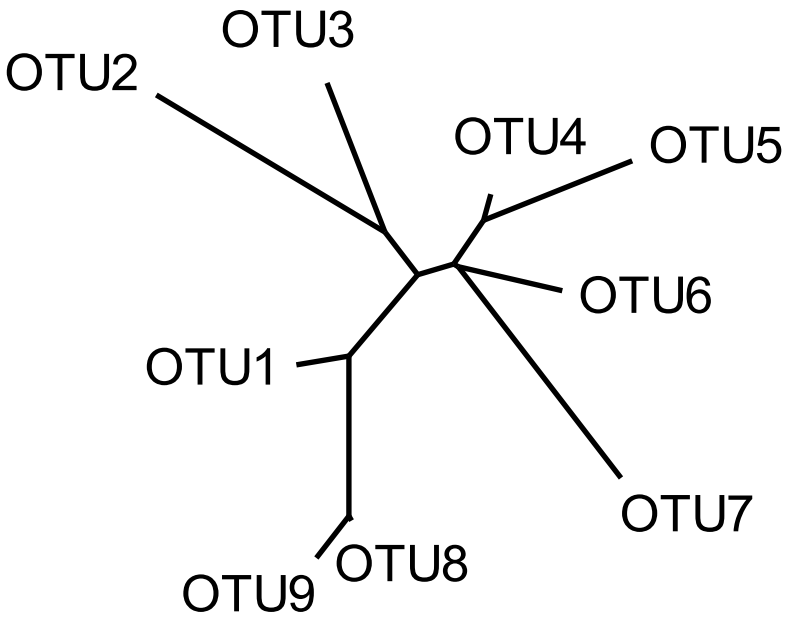

OTU2
OTU3
OTU1
OTU9
OTU4
OTU8
OTU5
OTU7
OTU6

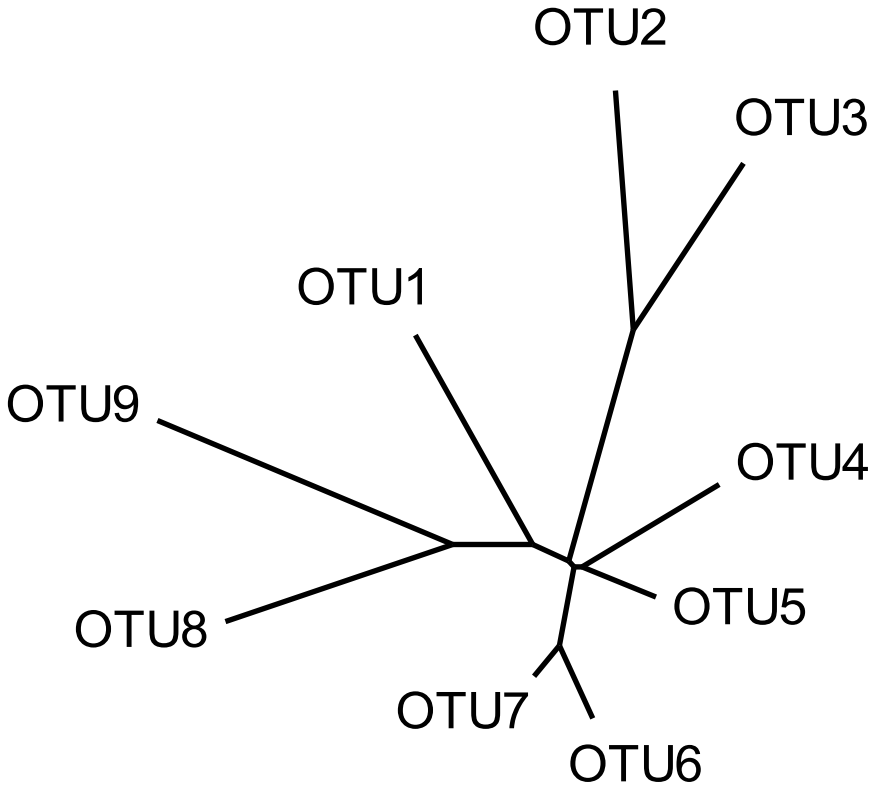
# 分離モデル




- 分離モデルでは系統樹は相似でない＝置換速度比が系統樹上で一定でないと仮定

# 分離モデル

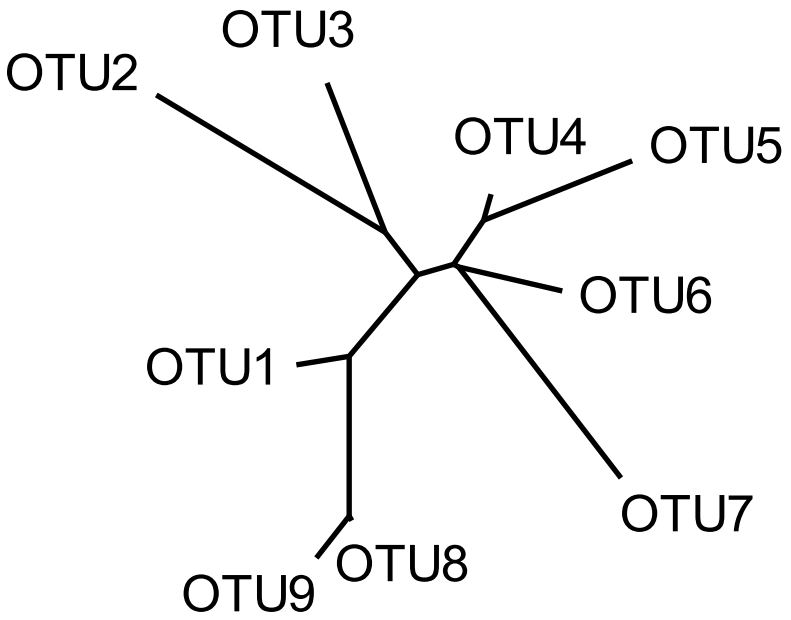


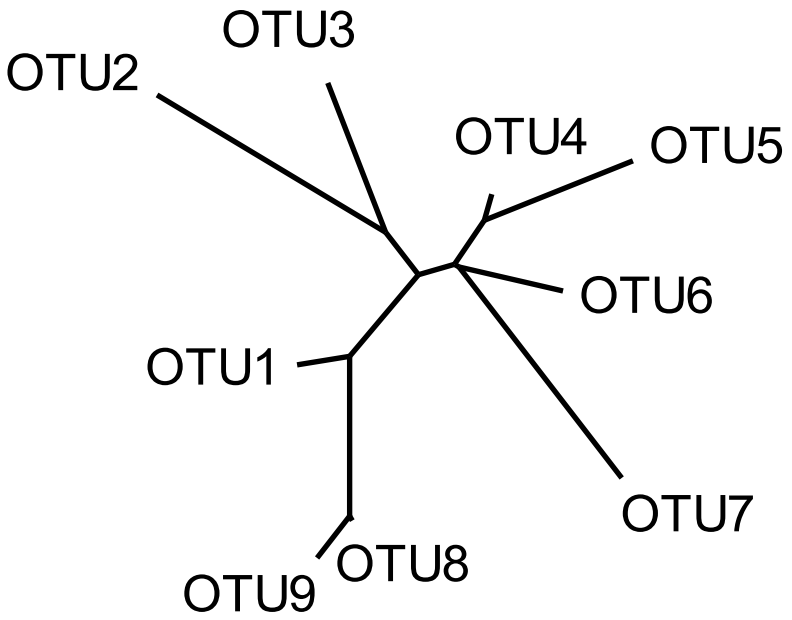




- 分離モデルでは系統樹は相似でない＝置換速度比が系統樹上で一定でないと仮定
  - 推定すべき枝長パラメータが激増

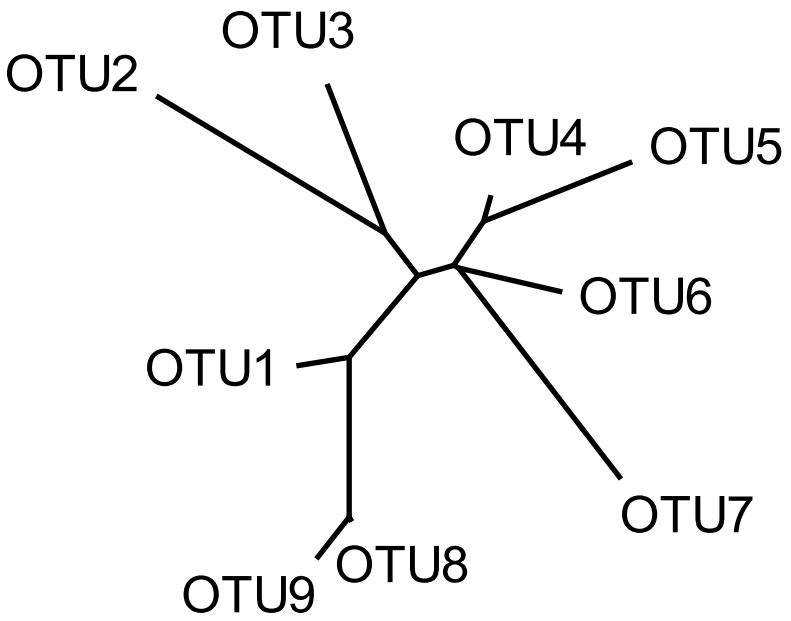
# パーティション間等速度モデル

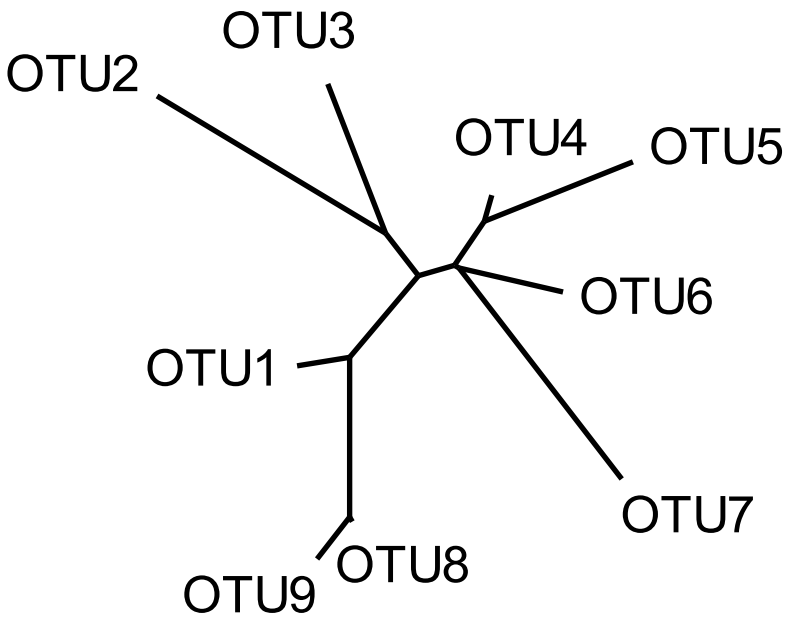
# パーティション間等速度モデル



# パーティション間等速度モデル




- 系統樹は同一＝枝長（置換速度）が等しいと仮定

# データから推定するパラメータ数のまとめ

# データから推定するパラメータ数のまとめ

- 置換速度行列

# データから推定するパラメータ数のまとめ

- 置換速度行列
  - DNA では $r_{XY}$ が 0 ～ 5
  - DNA では $\pi_X$ が 0 ～ 3

# データから推定するパラメータ数のまとめ

- 置換速度行列
  - DNA では $r_{XY}$ が 0 ～ 5
  - DNA では $\pi_X$ が 0 ～ 3
- 座位間の置換速度不均質性

# データから推定するパラメータ数のまとめ

- 置換速度行列
  - DNA では $r_{XY}$ が 0 ～ 5
  - DNA では $\pi_X$ が 0 ～ 3
- 座位間の置換速度不均質性
  - 領域ごとに 0 以上

# データから推定するパラメータ数のまとめ

- 置換速度行列
  - DNA では $r_{XY}$ が 0 ～ 5
  - DNA では $\pi_X$ が 0 ～ 3
- 座位間の置換速度不均質性
  - 領域ごとに 0 以上

- 非区分モデルとパーティション間等速度モデル

# データから推定するパラメータ数のまとめ

- 置換速度行列
  - DNA では $r_{XY}$ が 0 ～ 5
  - DNA では $\pi_X$ が 0 ～ 3
- 座位間の置換速度不均質性
  - 領域ごとに 0 以上

- 非区分モデルとパーティション間等速度モデル
  - 枝長：OTU 数 × 2 − 3

# データから推定するパラメータ数のまとめ

- 置換速度行列
  - DNA では $r_{XY}$ が 0 ～ 5
  - DNA では $\pi_X$ が 0 ～ 3
- 座位間の置換速度不均質性
  - 領域ごとに 0 以上
- 非区分モデルとパーティション間等速度モデル
  - 枝長：OTU 数 × 2 − 3
- 比例モデル

# データから推定するパラメータ数のまとめ

- 置換速度行列
  - DNA では $r_{XY}$ が 0 ～ 5
  - DNA では $\pi_X$ が 0 ～ 3
- 座位間の置換速度不均質性
  - 領域ごとに 0 以上

- 非区分モデルとパーティション間等速度モデル
  - 枝長：OTU 数 × 2 − 3
- 比例モデル
  - 枝長：OTU 数 × 2 − 3
  - 枝長比：領域数− 1

# データから推定するパラメータ数のまとめ

- 置換速度行列
  - DNA では $r_{XY}$ が 0 ～ 5
  - DNA では $\pi_X$ が 0 ～ 3
- 座位間の置換速度不均質性
  - 領域ごとに 0 以上

- 非区分モデルとパーティション間等速度モデル
  - 枝長：OTU 数 × 2 − 3
- 比例モデル
  - 枝長：OTU 数 × 2 − 3
  - 枝長比：領域数− 1
- 分離モデル

# データから推定するパラメータ数のまとめ

- 置換速度行列
  - DNA では $r_{XY}$ が 0 ～ 5
  - DNA では $\pi_X$ が 0 ～ 3
- 座位間の置換速度不均質性
  - 領域ごとに 0 以上

- 非区分モデルとパーティション間等速度モデル
  - 枝長：OTU 数 × 2 − 3
- 比例モデル
  - 枝長：OTU 数 × 2 − 3
  - 枝長比：領域数− 1
- 分離モデル
  - 枝長：(OTU 数 × 2 − 3) × 領域数

で、結局

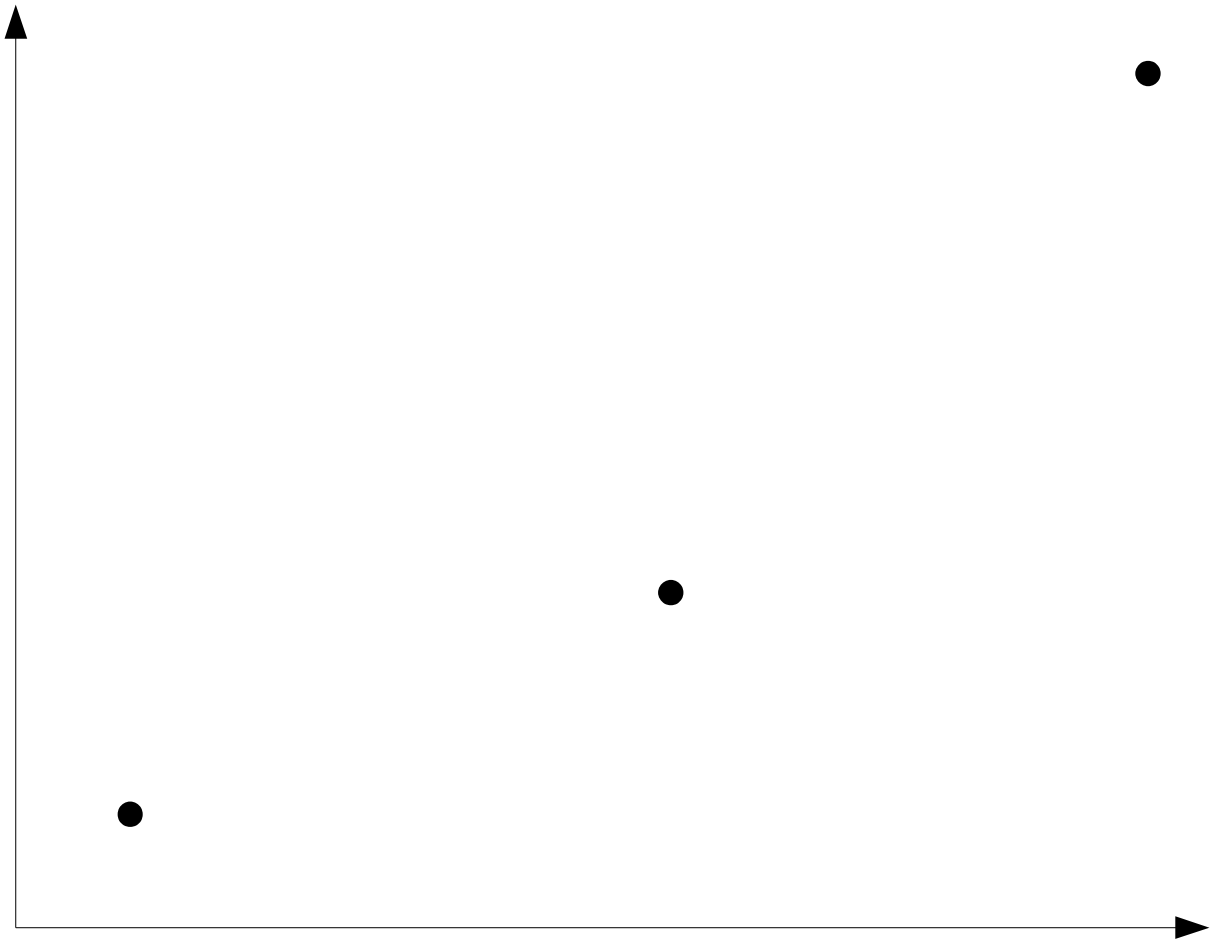
どれが**最適**

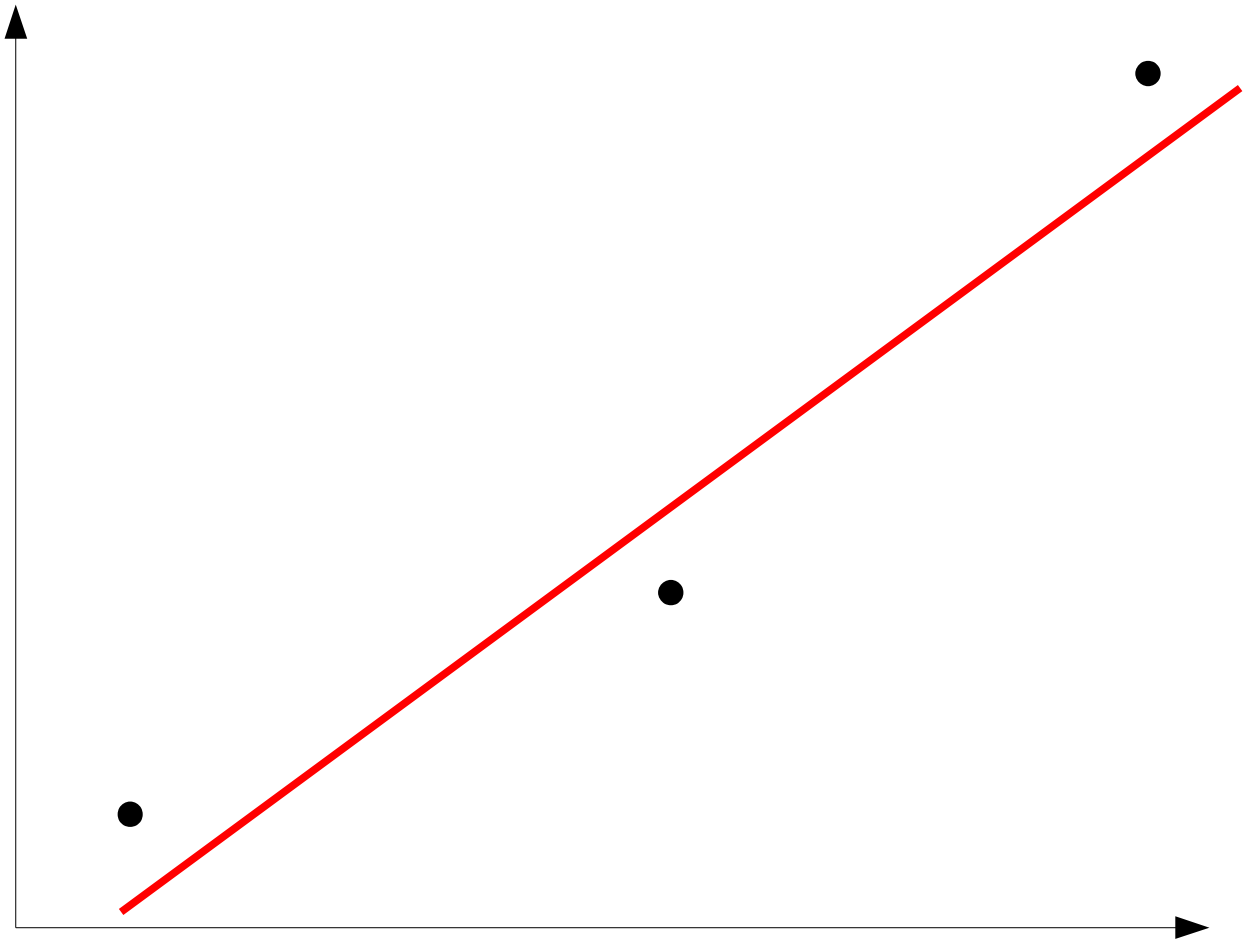
なのか？

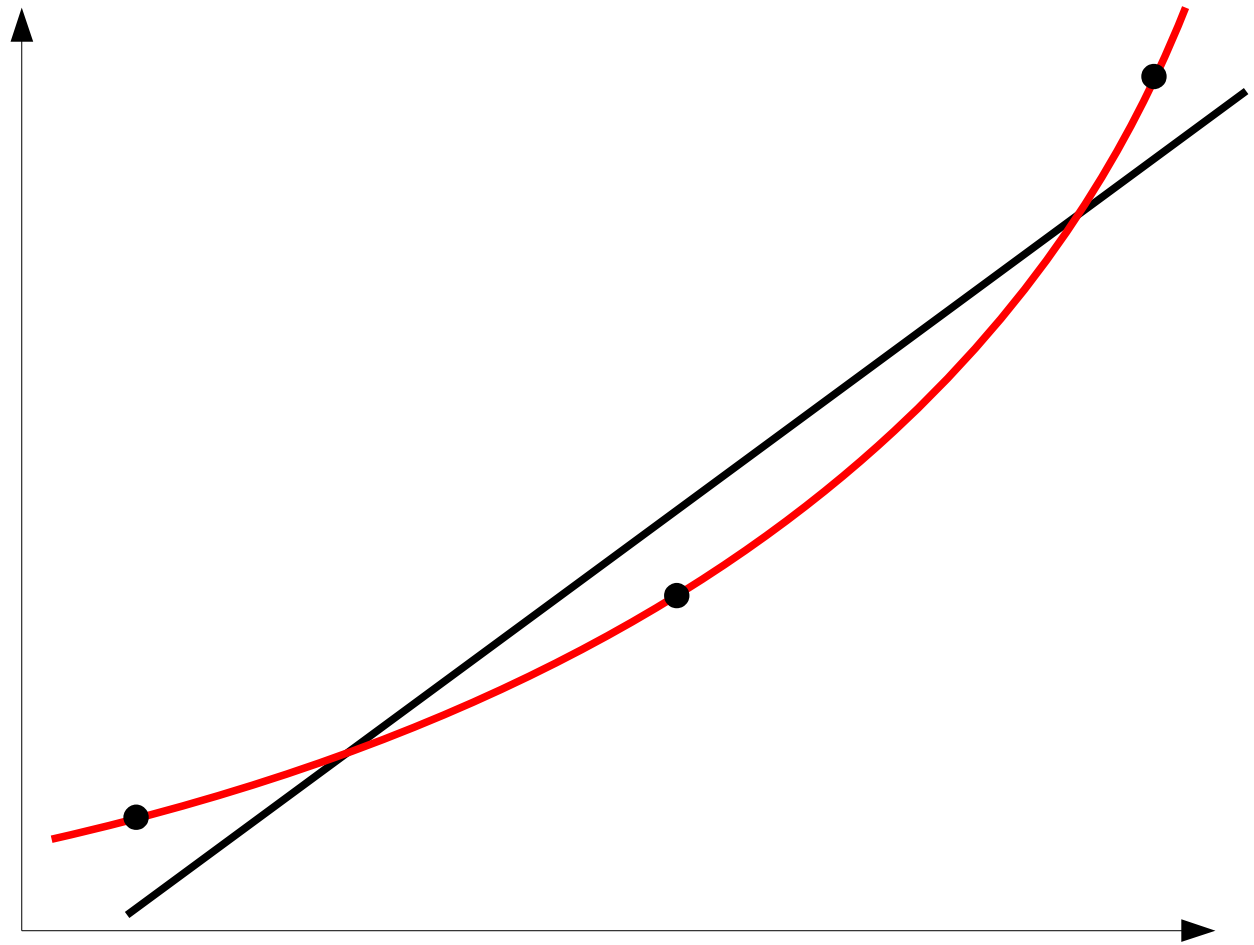
# モデル選択

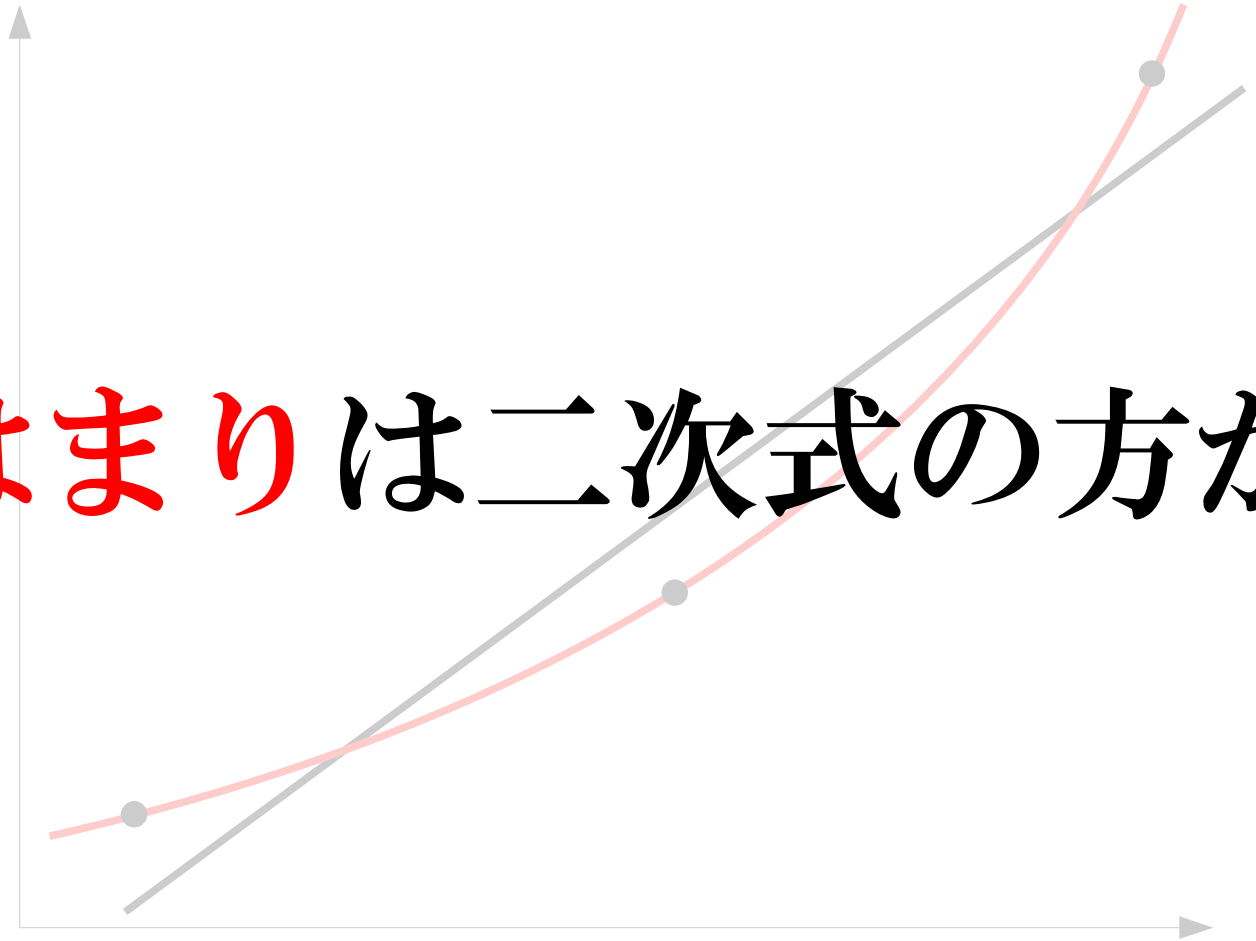
しよう

# モデル選択って何？



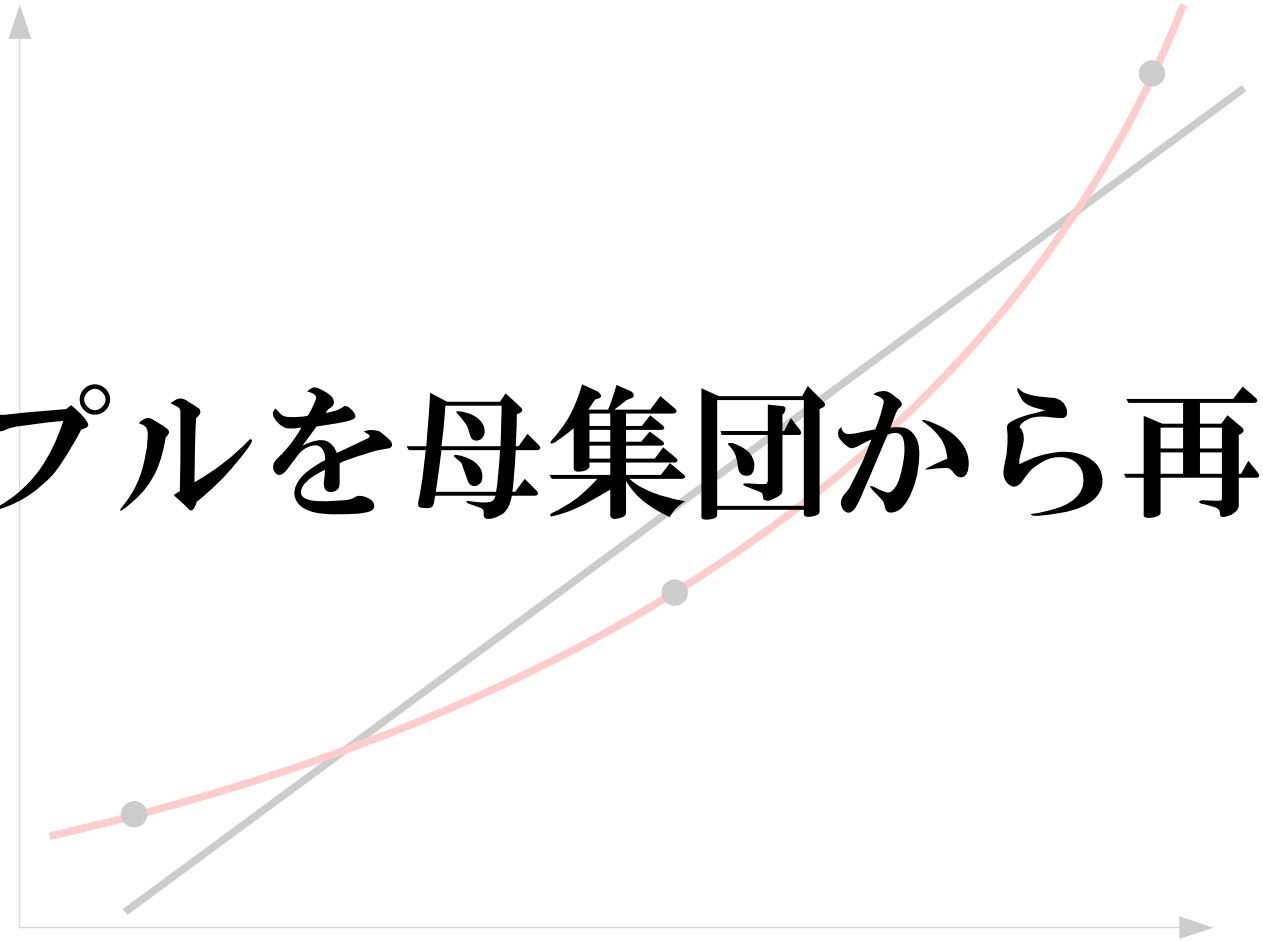


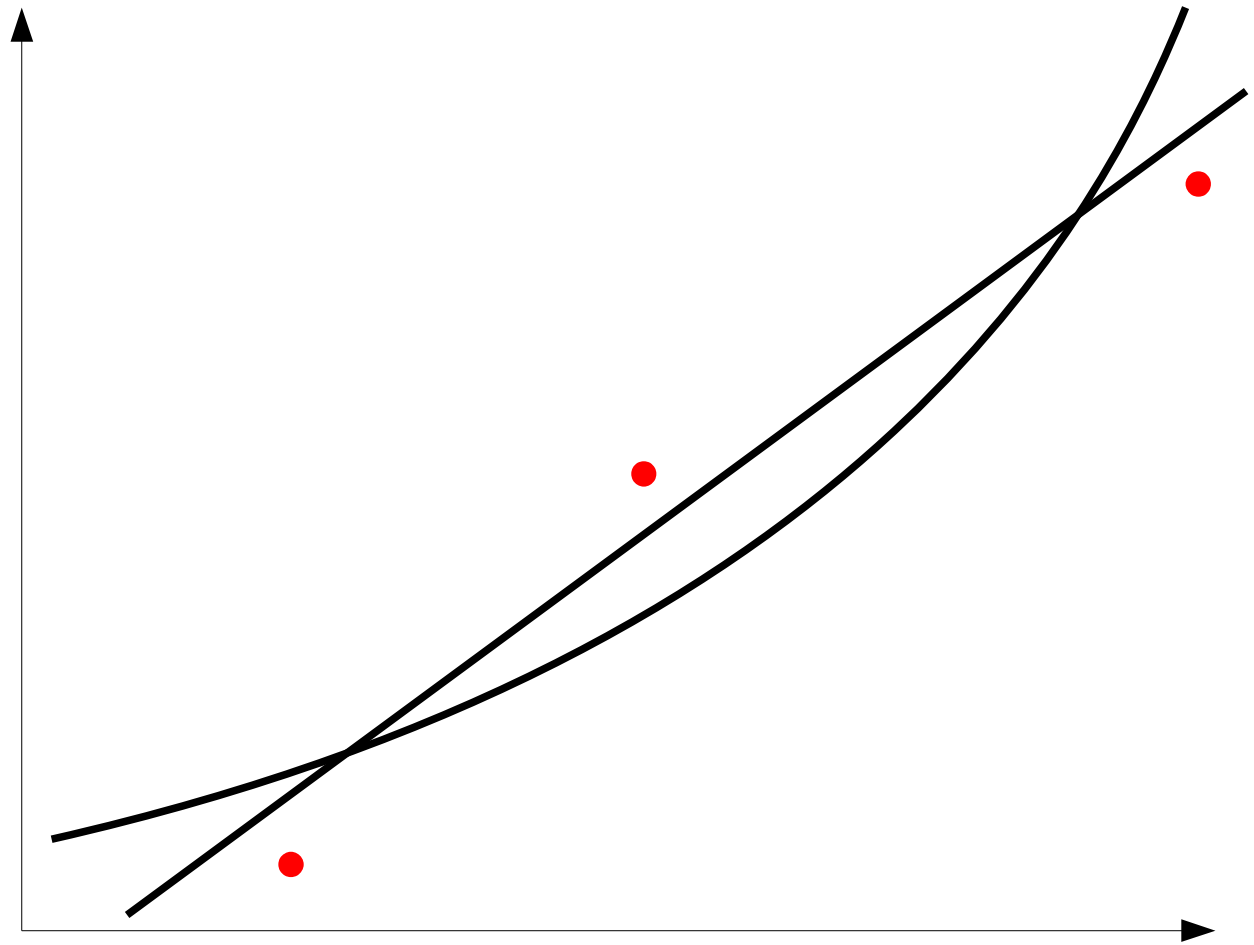


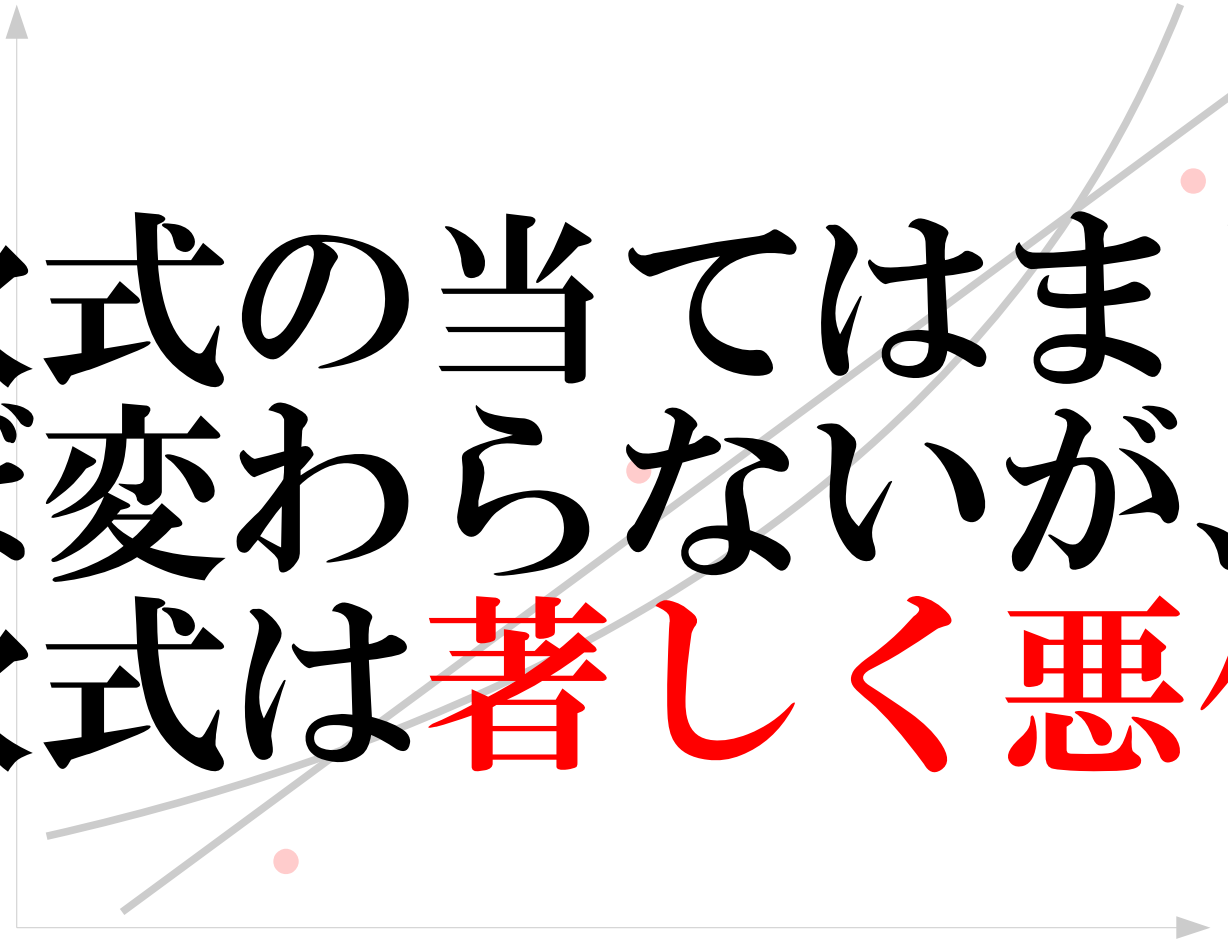


# 当てはまりは二次式の方が良い

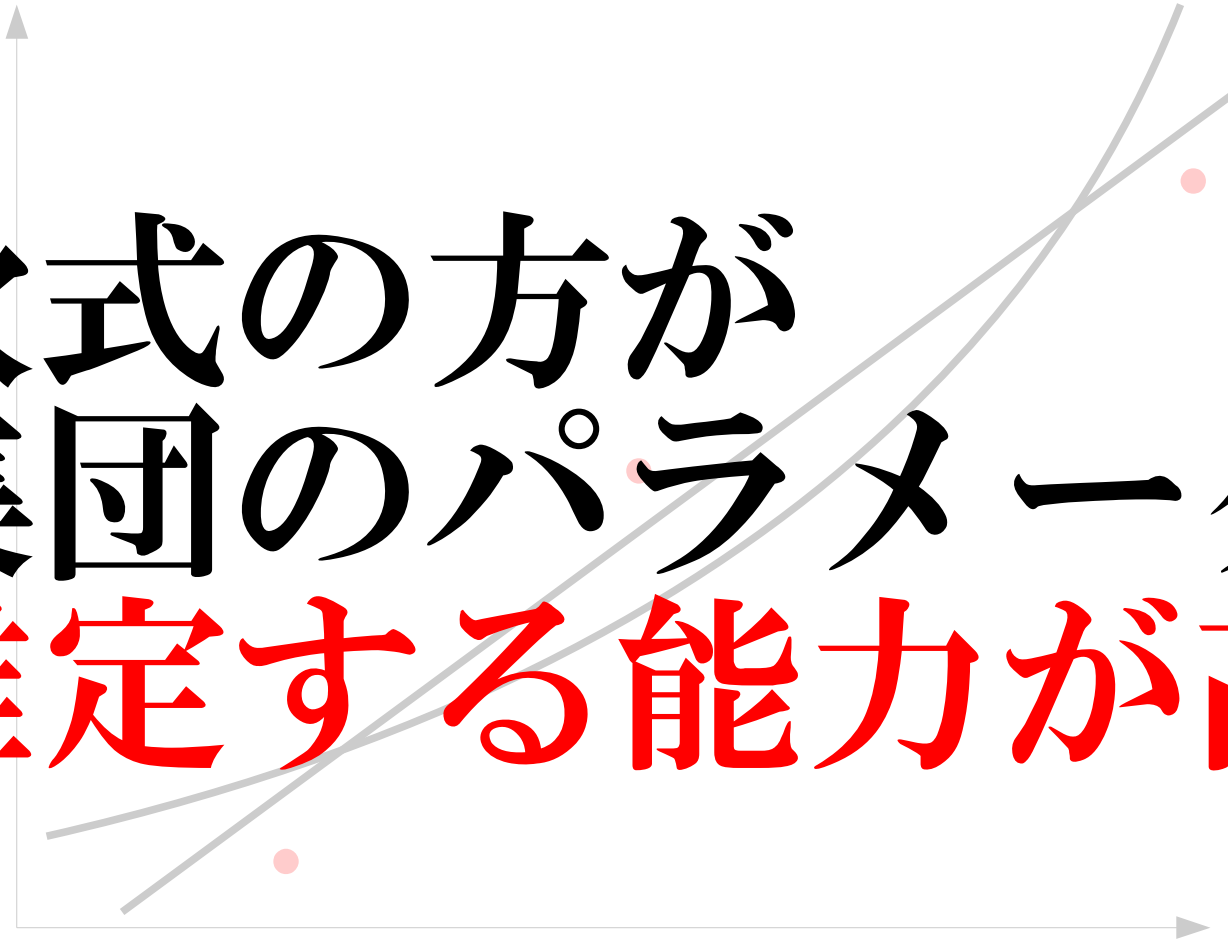
# サンプルを母集団から再抽出







一次式の当てはまりは
ほぼ変わらないが、
二次式は**著しく悪化**



一次式の方が
母集団のパラメータを
推定する能力が高い

# 当てはまりの良さ

# 当てはまりの良さ

|

# パラメータ数

当てはまりの良さ

|

パラメータ数

||

**推定能力**

$$\mathrm{AIC} = -2\ln L + 2K$$

($L$ は最大化尤度・ $K$ はパラメータ数 )