分子系統解析における
様々な問題について

そもそもどこの配列を使うべき?

そもそもどこの配列を使うべき?

- 置換が早すぎず遅すぎない(=多すぎず少なすぎない)

そもそもどこの配列を使うべき?

- 置換が早すぎず遅すぎない(=多すぎず少なすぎない)
- 連続長は長い方が良い

そもそもどこの配列を使うべき?

- 置換が早すぎず遅すぎない(=多すぎず少なすぎない)
- 連続長は長い方が良い
- 遺伝子重複が起きていない(=パラログでない)

で、そういう領域をどうやって探す?

で、そういう領域をどうやって探す?

- 外群種と内群種のゲノム・トランスクリプトームがある場合

で、そういう領域をどうやって探す?

- 外群種と内群種のゲノム・トランスクリプトームがある場合
  - BLASTで類似箇所を探す

で、そういう領域をどうやって探す?

- 外群種と内群種のゲノム・トランスクリプトームがある場合
  - BLASTで類似箇所を探す
  - 類似度が高く、アライメント長が長く、そういうのが1件だけのものを採用

で、そういう領域をどうやって探す?

- 外群種と内群種のゲノム・トランスクリプトームがある場合
  - BLASTで類似箇所を探す
  - 類似度が高く、アライメント長が長く、そういうのが1件だけのものを採用
  - HiMAP (Highly Multiplexed Amplicon-based Phylogenomics) というプログラムで自動的に領域選択＋プライマー設計できる

で、そういう領域をどうやって探す?

- 外群種と内群種のゲノム・トランスクリプトームがある場合
  - BLASTで類似箇所を探す
  - 類似度が高く、アライメント長が長く、そういうのが1件だけのものを採用
  - HiMAP (Highly Multiplexed Amplicon-based Phylogenomics) というプログラムで自動的に領域選択＋プライマー設計できる
- ゲノム・トランスクリプトームがない場合

で、そういう領域をどうやって探す?

- 外群種と内群種のゲノム・トランスクリプトームがある場合
  - BLASTで類似箇所を探す
  - 類似度が高く、アライメント長が長く、そういうのが1件だけのものを採用
  - HiMAP (Highly Multiplexed Amplicon-based Phylogenomics) というプログラムで自動的に領域選択＋プライマー設計できる
- ゲノム・トランスクリプトームがない場合
  - 全ゲノム解読する

で、そういう領域をどうやって探す?

- 外群種と内群種のゲノム・トランスクリプトームがある場合
  - BLASTで類似箇所を探す
  - 類似度が高く、アライメント長が長く、そういうのが1件だけのものを採用
  - HiMAP (Highly Multiplexed Amplicon-based Phylogenomics) というプログラムで自動的に領域選択＋プライマー設計できる
- ゲノム・トランスクリプトームがない場合
  - 全ゲノム解読する
  - トランスクリプトーム解析を行う

多遺伝子座連結解析の問題

多遺伝子座連結解析の問題

- パラログ混入や浸透交雑、水平伝播、incomplete lineage sortingで、遺伝子座間で支持する系統樹が異なる(不調和)

多遺伝子座連結解析の問題

- パラログ混入や浸透交雑、水平伝播、incomplete lineage sortingで、遺伝子座間で支持する系統樹が異なる(不調和)
  - 連結解析のブートストラップ値悪化やアーティファクトの原因

多遺伝子座連結解析の問題

- パラログ混入や浸透交雑、水平伝播、incomplete lineage sortingで、遺伝子座間で支持する系統樹が異なる(不調和)
  - 連結解析のブートストラップ値悪化やアーティファクトの原因
- Internode Certainty, ICAll, TreeC, TCA値で不調和を評価

多遺伝子座連結解析の問題

- パラログ混入や浸透交雑、水平伝播、incomplete lineage sortingで、遺伝子座間で支持する系統樹が異なる(不調和)
  - 連結解析のブートストラップ値悪化やアーティファクトの原因
- Internode Certainty, ICAll, TreeC, TCA値で不調和を評価
  - IC, ICAは系統仮説ごとに出るがTC, TCAは系統樹全体で1つ

多遺伝子座連結解析の問題

- パラログ混入や浸透交雑、水平伝播、incomplete lineage sortingで、遺伝子座間で支持する系統樹が異なる(不調和)
  - 連結解析のブートストラップ値悪化やアーティファクトの原因
- Internode Certainty, ICAll, TreeC, TCA値で不調和を評価
  - IC, ICAは系統仮説ごとに出るがTC, TCAは系統樹全体で1つ
  - ICの範囲は1～0で、ICAは1～マイナス?、小さいほど不調和

多遺伝子座連結解析の問題

- パラログ混入や浸透交雑、水平伝播、incomplete lineage sortingで、遺伝子座間で支持する系統樹が異なる(不調和)
  - 連結解析のブートストラップ値悪化やアーティファクトの原因
- Internode Certainty, ICAll, TreeC, TCA値で不調和を評価
  - IC, ICAは系統仮説ごとに出るがTC, TCAは系統樹全体で1つ
  - ICの範囲は1～0で、ICAは1～マイナス?、小さいほど不調和
  - TC, TCAはIC, ICAの総和．OTU数-3で割ってデータ間比較

多遺伝子座連結解析の問題

- パラログ混入や浸透交雑、水平伝播、incomplete lineage sortingで、遺伝子座間で支持する系統樹が異なる(不調和)
  - 連結解析のブートストラップ値悪化やアーティファクトの原因
- Internode Certainty, ICAll, TreeC, TCA値で不調和を評価
  - IC, ICAは系統仮説ごとに出るがTC, TCAは系統樹全体で1つ
  - ICの範囲は1～0で、ICAは1～マイナス?、小さいほど不調和
  - TC, TCAはIC, ICAの総和．OTU数-3で割ってデータ間比較
- 使用する遺伝子座を選別する

多遺伝子座連結解析の問題

- パラログ混入や浸透交雑、水平伝播、incomplete lineage sortingで、遺伝子座間で支持する系統樹が異なる(不調和)
  - 連結解析のブートストラップ値悪化やアーティファクトの原因
- Internode Certainty, ICAll, TreeC, TCA値で不調和を評価
  - IC, ICAは系統仮説ごとに出るがTC, TCAは系統樹全体で1つ
  - ICの範囲は1～0で、ICAは1～マイナス?、小さいほど不調和
  - TC, TCAはIC, ICAの総和．OTU数-3で割ってデータ間比較
- 使用する遺伝子座を選別する
  - Clusterflock, Concaterpiller, Conclustador

多遺伝子座連結解析の問題

- パラログ混入や浸透交雑、水平伝播、incomplete lineage sortingで、遺伝子座間で支持する系統樹が異なる(不調和)
  - 連結解析のブートストラップ値悪化やアーティファクトの原因
- Internode Certainty, ICAll, TreeC, TCA値で不調和を評価
  - IC, ICAは系統仮説ごとに出るがTC, TCAは系統樹全体で1つ
  - ICの範囲は1～0で、ICAは1～マイナス?、小さいほど不調和
  - TC, TCAはIC, ICAの総和．OTU数-3で割ってデータ間比較
- 使用する遺伝子座を選別する
  - Clusterflock, Concaterpiller, Conclustador
- species tree methodを使う

多遺伝子座連結解析の問題

- パラログ混入や浸透交雑、水平伝播、incomplete lineage sortingで、遺伝子座間で支持する系統樹が異なる(不調和)
  - 連結解析のブートストラップ値悪化やアーティファクトの原因
- Internode Certainty, ICAll, TreeC, TCA値で不調和を評価
  - IC, ICAは系統仮説ごとに出るがTC, TCAは系統樹全体で1つ
  - ICの範囲は1～0で、ICAは1～マイナス?、小さいほど不調和
  - TC, TCAはIC, ICAの総和．OTU数-3で割ってデータ間比較
- 使用する遺伝子座を選別する
  - Clusterflock, Concaterpiller, Conclustador
- species tree methodを使う
  - STEM, BUCKy, ASTRAL, *BEAST, BEST(MrBayes)

タクソンサンプリング法

タクソンサンプリング法

- 全種サンプリングは必ずしも良くない

タクソンサンプリング法

- 全種サンプリングは必ずしも良くない
- 系統樹上の分岐点・端点の密度ができるだけ偏らない方が良い

タクソンサンプリング法

- 全種サンプリングは必ずしも良くない
- 系統樹上の分岐点・端点の密度ができるだけ偏らない方が良い
  - 同一配列や近縁配列が一部では多く一部では少ないのは×

パーティションの切り方

パーティションの切り方

- Kakusan4は以下を比較して選択

パーティションの切り方

- Kakusan4は以下を比較して選択
  - 遺伝子座間・コドン位置間全部切る
  - 遺伝子座間全部切る・コドン位置間全部切らない
  - 遺伝子座間・コドン位置間全部切らない

パーティションの切り方

- Kakusan4は以下を比較して選択
  - 遺伝子座間・コドン位置間全部切る
  - 遺伝子座間全部切る・コドン位置間全部切らない
  - 遺伝子座間・コドン位置間全部切らない
- もっと柔軟にな切り方があるのでは?

パーティションの切り方

- Kakusan4は以下を比較して選択
  - 遺伝子座間・コドン位置間全部切る
  - 遺伝子座間全部切る・コドン位置間全部切らない
  - 遺伝子座間・コドン位置間全部切らない
- もっと柔軟にな切り方があるのでは?
  - PartitionFinderで探索可能

χ^2検定で組成の均一性が棄却されたら

χ^2検定で組成の均一性が棄却されたら

- 塩基配列ではACGTをAGYやRYに変換する

χ^2検定で組成の均一性が棄却されたら

- 塩基配列ではACGTをAGYやRYに変換する
- アミノ酸配列はDayhoff coding法+GTR20モデルなどを使う

χ^2検定で組成の均一性が棄却されたら

- 塩基配列ではACGTをAGYやRYに変換する
- アミノ酸配列はDayhoff coding法+GTR20モデルなどを使う
  - 形質状態のいくつかを統合することで無理矢理均一に

χ^2検定で組成の均一性が棄却されたら

- 塩基配列ではACGTをAGYやRYに変換する
- アミノ酸配列はDayhoff coding法+GTR20モデルなどを使う
  - 形質状態のいくつかを統合することで無理矢理均一に
- nhPhyloBayesで系統樹上での組成変化を許す

χ^2検定で組成の均一性が棄却されたら

- 塩基配列ではACGTをAGYやRYに変換する
- アミノ酸配列はDayhoff coding法+GTR20モデルなどを使う
  - 形質状態のいくつかを統合することで無理矢理均一に
- nhPhyloBayesで系統樹上での組成変化を許す
  - より適しているがLinux上でしか動かない

例：塩基配列の第3コドン位置だけRYコード化

下記のコマンドを入力してEnter

```
pgrecodeseq \          …コマンド名
 --type=DNA \          …入力配列はDNA
 3-.\3 \               …3つめから最後まで3つおきに処理
 GT-AC \               …GをAに、TをCに置換
 inputfile \           …入力ファイル名
 outputfile            …出力ファイル名
```

「\」は「次の行に改行なしで続く」という意味であることに注意
ただしスペースは入れること

例：χ^2検定で不均質解消を確認

下記のコマンドを入力してEnter

```
pgtestcomposition \      …コマンド名
 --type=DNA \            …入力配列はDNA
 3-.\3 \                 …3つめから最後まで3つおきに処理
 inputfile \             …入力ファイル名
 outputfile              …出力ファイル名
```

「\」は「次の行に改行なしで続く」という意味であることに注意
ただしスペースは入れること

例：アミノ酸配列をDayhoffコード化

下記のコマンドを入力してEnter

```
pgrecodeseq \        …コマンド名
 --type=AA \         …入力配列はアミノ酸
 STGPNEQKHVILYW-AAAADDDRRMMMFF \
 inputfile \         …入力ファイル名
 outputfile          …出力ファイル名
```

「\」は「次の行に改行なしで続く」という意味であることに注意
ただしスペースは入れること

変換したデータ解析の注意

- RAxMLで解析するときはさらに01データにしてbinaryデータとして解析する
  - -m BINGAMMA

変換したデータ解析の注意

- RAxMLで解析するときはさらに01データにしてbinaryデータとして解析する
  - -m BINGAMMA
- RAxMLで解析するときはさらに0～9A～Vのデータにして multistateデータとして解析する
  - -m MULTIGAMMA -K GTR

データのギャップ情報を使いたいとき

データのギャップ情報を使いたいとき

- トリミング前の配列から、simple indel coding法でギャップの有無を01に符号化

データのギャップ情報を使いたいとき

- トリミング前の配列から、simple indel coding法でギャップの有無を01に符号化
- トリミング後の配列に加えてMrBayes, RAxML, PAUP*で系統樹推定

例：simple indel coding法でギャップ情報を01データ化

下記のコマンドを入力してEnter

```
pgencodegap \          …コマンド名
 --method=SIC \        …符号化法はSIC
 inputfile \           …入力ファイル名
 outputfile            …出力ファイル名
```

注：入力ファイル形式はNEXUSのみに対応

「\」は「次の行に改行なしで続く」という意味であることに注意
ただしスペースは入れること

例：ギャップの01データを塩基配列と連結

下記のコマンドを入力してEnter

```
pgconcatgap \                    …コマンド名
 --output=MrBayes \              …MrBayes向けの出力を行う
 DNAseqfile \                    …塩基配列ファイル名
 binarydatafile                  …01データファイル名
```

「\」は「次の行に改行なしで続く」という意味であることに注意
ただしスペースは入れること

変異がある座位だけのデータに関する注意事項

変異がある座位だけのデータに関する注意事項

- 形態形質・SNPなどのデータでは、変異がある座位しか含まれていない

変異がある座位だけのデータに関する注意事項

- 形態形質・SNPなどのデータでは、変異がある座位しか含まれていない
- これは、「データ収集にバイアスascertainment biasがある」

変異がある座位だけのデータに関する注意事項

- 形態形質・SNPなどのデータでは、変異がある座位しか含まれていない
- これは、「データ収集にバイアスascertainment biasがある」
- RAxMLでは以下のオプションで補正した尤度を使用する
  - -m ASC_BINGAMMA
  - -m ASC_MULTIGAMMA
  - -m ASC_GTRGAMMA
  - -m ASC_PROTGAMMA[matrixname](F)

“Selective Inference”問題

“Selective Inference”問題

- データから推定した最尤系統樹とデータから推定した対立仮説の系統樹の尤度を比較するときデータを何度も使用している

“Selective Inference”問題

- データから推定した最尤系統樹とデータから推定した対立仮説の系統樹の尤度を比較するときデータを何度も使用している
  - →多重性

“Selective Inference”問題

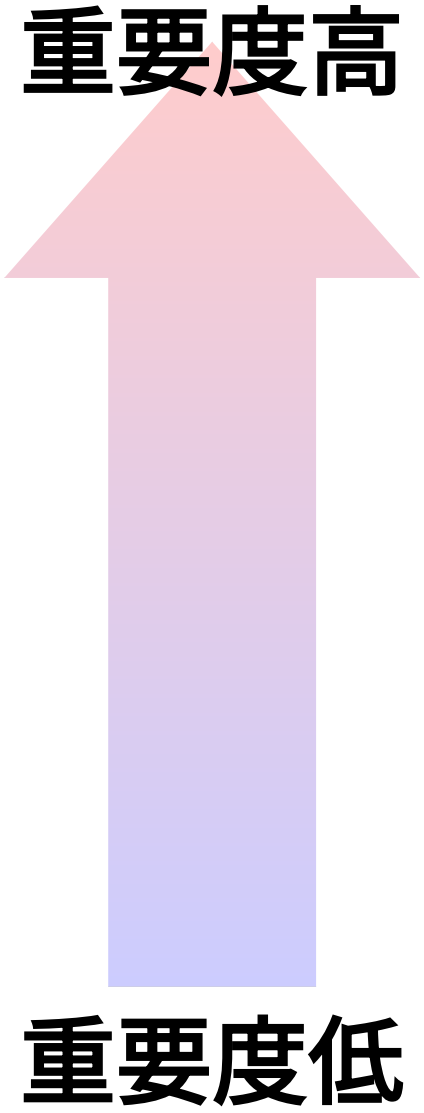
- データから推定した最尤系統樹とデータから推定した対立仮説の系統樹の尤度を比較するときデータを何度も使用している
  - →多重性
- データを複数回使用する際の多重性を適切に扱う枠組みがSelective Inference (Post-selection Inference)

“Selective Inference”問題

- データから推定した最尤系統樹とデータから推定した対立仮説の系統樹の尤度を比較するときデータを何度も使用している
  - →多重性
- データを複数回使用する際の多重性を適切に扱う枠組みがSelective Inference (Post-selection Inference)
- Selective InferenceのときのAU検定のp値を算出する方法が提案されている(Shimodaira and Terada 2019)

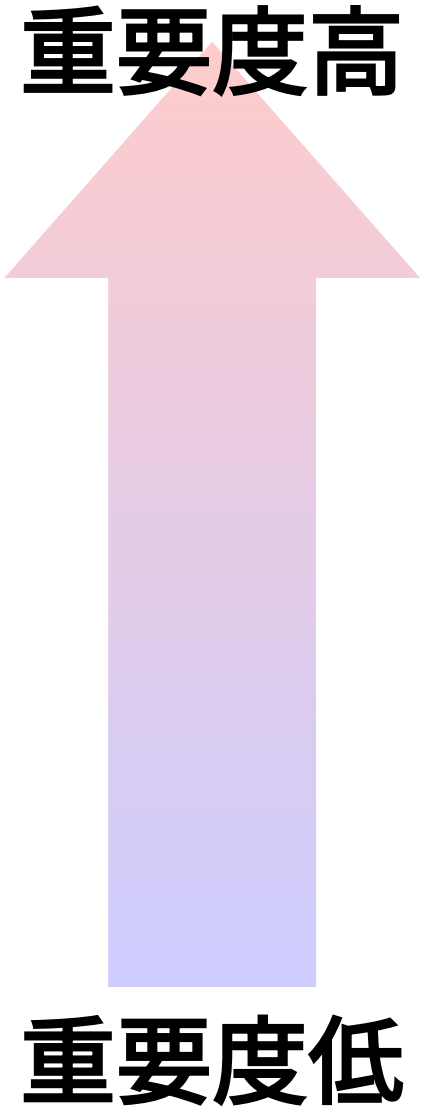
系統樹推定の勘所

系統樹推定の勘所



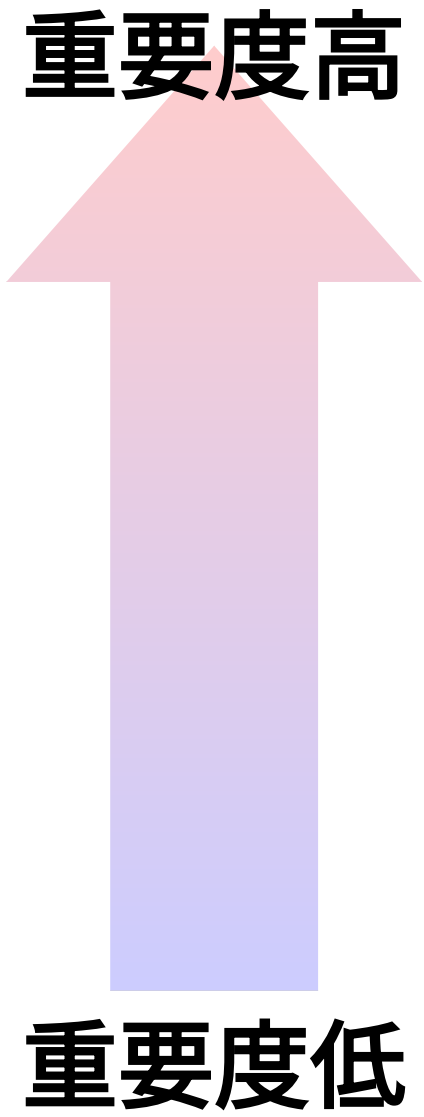
系統樹推定の勘所

- データの質



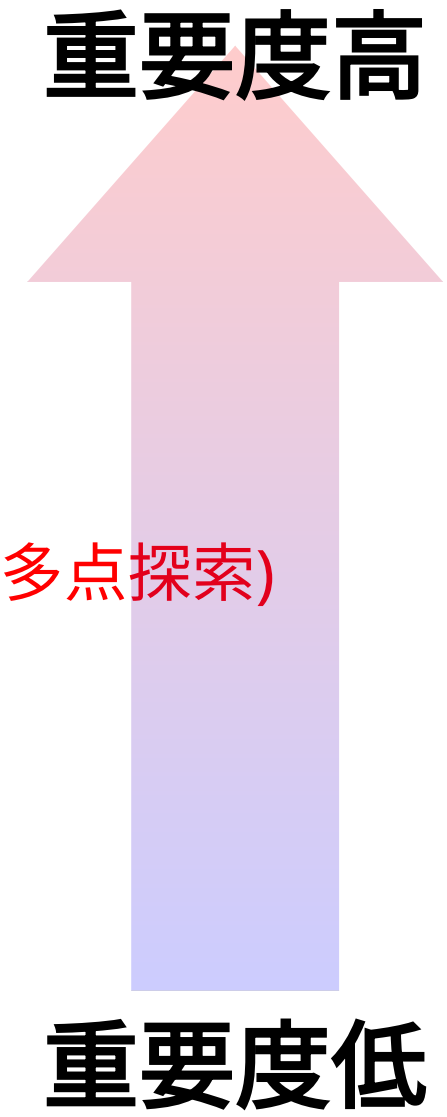
系統樹推定の勘所

- データの質
  - 多重整列とトリミング
  - 遺伝子座サンプリング
  - タクソンサンプリング
  - 不適な部分の除去



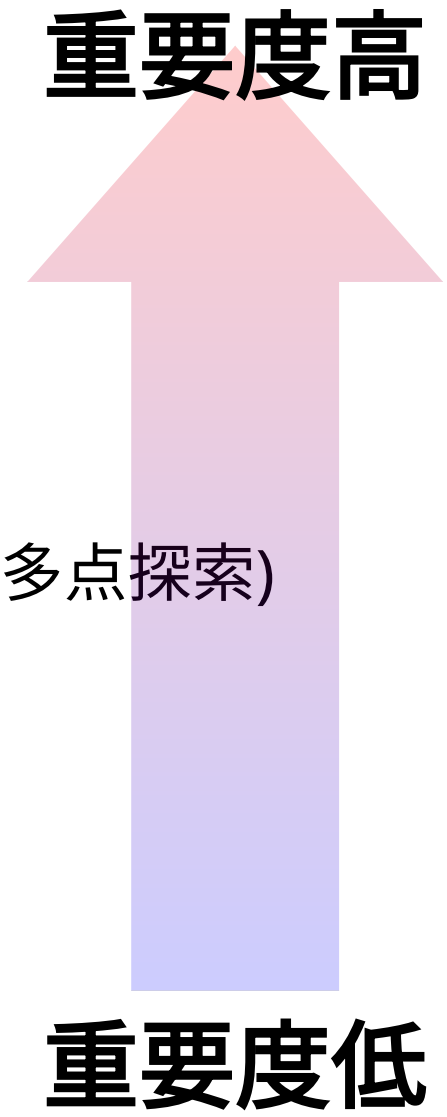
系統樹推定の勘所

- データの質
  - 多重整列とトリミング
  - 遺伝子座サンプリング
  - タクソンサンプリング
  - 不適な部分の除去
- 樹形探索範囲の広さ(NNI・SPR・TBR・多点探索)



系統樹推定の勘所

- データの質
  - 多重整列とトリミング
  - 遺伝子座サンプリング
  - タクソンサンプリング
  - 不適な部分の除去
- 樹形探索範囲の広さ(NNI・SPR・TBR・多点探索)
- パーティションの切り方



系統樹推定の勘所

- データの質
  - 多重整列とトリミング
  - 遺伝子座サンプリング
  - タクソンサンプリング
  - 不適な部分の除去
- 樹形探索範囲の広さ(NNI・SPR・TBR・多点探索)
- パーティションの切り方
- パーティション間モデル(等速度・比例・分離)

重要度高

重要度低

系統樹推定の勘所

- データの質
  - 多重整列とトリミング
  - 遺伝子座サンプリング
  - タクソンサンプリング
  - 不適な部分の除去
- 樹形探索範囲の広さ(NNI・SPR・TBR・多点探索)
- パーティションの切り方
- パーティション間モデル(等速度・比例・分離)
- パーティション内モデル(JC69～GTR+G)

重要度高

重要度低