

分子系統樹推定に適した 配列データセットの作成

系統樹推定に利用可能なデータ

系統樹推定に利用可能なデータ

- binary data

系統樹推定に利用可能なデータ

- binary data
 - 形質がある場合を 1、ない場合を0 とする

系統樹推定に利用可能なデータ

- binary data
 - 形質がある場合を 1、ない場合を0 とする
 - 異なる座位の 0 が同じ意味を持つ

系統樹推定に利用可能なデータ

- binary data
 - 形質がある場合を 1、ない場合を 0 とする
 - 異なる座位の 0 が同じ意味を持つ
- multistate/ordered data

系統樹推定に利用可能なデータ

- binary data
 - 形質がある場合を 1、ない場合を 0 とする
 - 異なる座位の 0 が同じ意味を持つ
- multistate/ordered data
 - 形質状態 0 ・ 1 ・ 2... のそれぞれを 0 ・ 1 ・ 2... とする

系統樹推定に利用可能なデータ

- binary data
 - 形質がある場合を 1、ない場合を 0 とする
 - 異なる座位の 0 が同じ意味を持つ
- multistate/ordered data
 - 形質状態 0 ・ 1 ・ 2... のそれぞれを 0 ・ 1 ・ 2... とする
 - **ordered** では数値が量を表す

系統樹推定に利用可能なデータ

- binary data
 - 形質がある場合を 1、ない場合を 0 とする
 - 異なる座位の 0 が同じ意味を持つ
- multistate/ordered data
 - 形質状態 0 ・ 1 ・ 2... のそれぞれを 0 ・ 1 ・ 2... とする
 - ordered では数値が量を表す
 - 異なる座位の 0 は同じ意味を持たない

系統樹推定に利用可能なデータ

- binary data
 - 形質がある場合を 1、ない場合を 0 とする
 - 異なる座位の 0 が同じ意味を持つ
- multistate/ordered data
 - 形質状態 0 ・ 1 ・ 2... のそれぞれを 0 ・ 1 ・ 2... とする
 - ordered では数値が量を表す
 - 異なる座位の 0 は同じ意味を持たない
- nucleotide/amino-acid data

系統樹推定に利用可能なデータ

- binary data
 - 形質がある場合を 1、ない場合を 0 とする
 - 異なる座位の 0 が同じ意味を持つ
- multistate/ordered data
 - 形質状態 0 ・ 1 ・ 2... のそれぞれを 0 ・ 1 ・ 2... とする
 - ordered では数値が量を表す
 - 異なる座位の 0 は同じ意味を持たない
- nucleotide/amino-acid data
 - 4 形質状態または 20 形質状態のデータ

系統樹推定に利用可能なデータ

- binary data
 - 形質がある場合を 1、ない場合を 0 とする
 - 異なる座位の 0 が同じ意味を持つ
- multistate/ordered data
 - 形質状態 0・1・2... のそれぞれを 0・1・2... とする
 - ordered では数値が量を表す
 - 異なる座位の 0 は同じ意味を持たない
- nucleotide/amino-acid data
 - 4 形質状態または 20 形質状態のデータ
 - 異なる座位の A が同じ意味を持つ

系統樹推定に利用可能なデータ

- binary data
 - 形質がある場合を 1、ない場合を 0 とする
 - 異なる座位の 0 が同じ意味を持つ
- multistate/ordered data
 - 形質状態 0・1・2... のそれぞれを 0・1・2... とする
 - ordered では数値が量を表す
 - 異なる座位の 0 は同じ意味を持たない
- nucleotide/amino-acid data
 - 4 形質状態または 20 形質状態のデータ
 - 異なる座位の A が同じ意味を持つ
- rRNA/tRNA data with secondary structure information

系統樹推定に利用可能なデータ

- binary data
 - 形質がある場合を 1、ない場合を 0 とする
 - 異なる座位の 0 が同じ意味を持つ
- multistate/ordered data
 - 形質状態 0・1・2... のそれぞれを 0・1・2... とする
 - ordered では数値が量を表す
 - 異なる座位の 0 は同じ意味を持たない
- nucleotide/amino-acid data
 - 4 形質状態または 20 形質状態のデータ
 - 異なる座位の A が同じ意味を持つ
- rRNA/tRNA data with secondary structure information
 - rRNA/tRNA の stem 部に専用モデルを適用するためのもの

データサンプリングのバイアスについて

データサンプリングのバイアスについて

- SNP や制限酵素座位の有無、形態は

変異がある

形質だけがサンプリングされるので、偏っている

データサンプリングのバイアスについて

- SNP や制限酵素座位の有無、形態は

変異がある

形質だけがサンプリングされるので、偏っている

- これを ascertainment bias とか coding bias という

データサンプリングのバイアスについて

- SNP や制限酵素座位の有無、形態は

変異がある

形質だけがサンプリングされるので、偏っている

- これを ascertainment bias とか coding bias という
- 最尤法やベイズ法では、バイアスを補正するモデルを適用する必要がある (RAxML ・ MrBayes などが対応)

データサンプリングのバイアスについて

- SNP や制限酵素座位の有無、形態は

変異がある

形質だけがサンプリングされるので、偏っている

- これを ascertainment bias とか coding bias という
- 最尤法やベイズ法では、バイアスを補正するモデルを適用する必要がある (RAxML ・ MrBayes などが対応)
- 最節約法では気にしなくてよい

分子系統樹

の

推定に適している

とは？

相阿

である







相同







非相同

相同



The background of the image is a composite of two photographs. On the left, there is a close-up of a brown frog with large, dark eyes, resting on a rocky surface. On the right, there is a close-up of a beetle with a metallic blue-green sheen and yellowish-brown markings on its back. The text is overlaid on this background.

相同

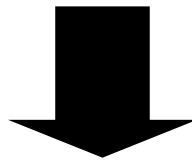
||

同一の祖先形質

に由来する

TaxonA AAGTGTGACTGGGGAT
TaxonB TGTGACTGCAT
TaxonC AATGTGGCTGCGAT
TaxonD TCTGACTG

TaxonA	AAGTGTGACTGGGGAT
TaxonB	TGTGACTGCAAT
TaxonC	AATGTGGCTGCGAT
TaxonD	TCTGACTG



TaxonA	AAGTGTGACTGGGGAT
TaxonB	---TGTGACTGC-AT
TaxonC	AA-TGTGGCTGCGAT
TaxonD	---TCTGACTG----

多重配列整列



TaxonA	AAGTGTGACTGGGGAT
TaxonB	---TGTGACTGC-AT
TaxonC	AA-TGTGGCTGCGAT
TaxonD	---TCTGACTG----

TaxonA AAGTGTGACTGGGGAT
TaxonB TGTGACTGCGAT
TaxonC AAGTGTGACTGGGGAT
TaxonD TCTGACTG

多重配列整列

||
↓

相同形質の同定

TaxonA AAGTGTGACTGGGGAT
TaxonB ---TGTGACTGC-AT
TaxonC AA-TGTGGCTGCGAT
TaxonD ---TCTGACTG----

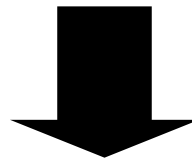
整列が

怪しいとき～

捨てれば
いいじゃない

TaxonA AAGTGTGACTGGGGAT
TaxonB - - - TGTGACTGCA - AT
TaxonC AA - TGTGGCTGCGAT
TaxonD - - - TCTGACTG - - - -

TaxonA	AAGTGTGACTGGGGAT
TaxonB	---TGTGACTGC-AT
TaxonC	AA-TGTGGCTGCGAT
TaxonD	---TCTGACTG-----



トリミング

TaxonA	TGTGACTG
TaxonB	TGTGACTG
TaxonC	TGTGGCTG
TaxonD	TCTGACTG

分子系統樹

の

推定に適していいない

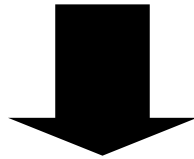
とは？

假定

に反する

TaxonA	GGGTTTGGTA-TAATTTCT
TaxonB	GGATTTGGCA-TAATCAGG
TaxonC	GGATT-GGTACTAATTAGT
TaxonD	GGATT-GGAACTAATTAGA

TaxonA	GGG	TTT	GG	TA	-	TAATTT	CT
TaxonB	GG	ATT	GG	CA	-	TAAT	CAGG
TaxonC	GG	ATT	-	GG	TAC	TAATT	AGT
TaxonD	GG	ATT	-	GG	AACT	TAATT	AGA



TaxonA	G	F	G	M	I	S
TaxonB	G	F	G	M	I	S
TaxonC	G	L	V	L	I	S
TaxonD	G	L	E	L	I	S

TaxonA	GGG	TTT	GGTA	-	TAATT	TTCT
TaxonB	GG	ATT	GGCA	-	TAAT	CAGG
TaxonC	GG	ATT	-GGT	ACT	TAATT	AGT
TaxonD	GG	ATT	-GGA	ACT	TAATT	AGA



TaxonA	G	F	G	M	I	S
TaxonB	G	F	G	M	I	S
TaxonC	G	L	V	L	I	S
TaxonD	G	L	E	L	I	S

TaxonA	GGG	TTT	GGTA	-	TAATTT	CT
TaxonB	GG	ATT	GGCA	-	TAAT	CAGG
TaxonC	GG	ATT	-	GGT	ACTA	ATTAGT
TaxonD	GG	ATT	-	GGA	ACTA	ATTAGA

多数のアミノ酸が一度に置換

TaxonA	G	F	G	M	I	S
TaxonB	G	F	G	M	I	S
TaxonC	G	L	V	L	I	S
TaxonD	G	L	E	L	I	S

TaxonA	GGGTTTGGTA-TAATTTCT
TaxonB	GGATTTGGCA-TAATCAGG
TaxonC	GGATT-GGTACTAATTAGT
TaxonD	GGATT-GGAAC TAATTAGA

多数のアミノ酸が一度に置換

TaxonA	G	F	G	M	I	S
TaxonB	G	F	G	M	I	S
TaxonC	G	L	V	L	I	S
TaxonD	G	L	E	L	I	S

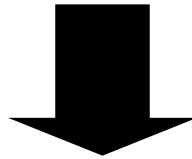
TaxonA GGGTTTGGTA-TAATTTCT
TaxonB GATTTGGCA-TATCAGG
TaxonC TT-GGT-AATTAGT
TaxonD GG-T-GCTAATTAGA

多数のアミノ酸が一度に置換

TaxonA F G I S
TaxonB F G I S
TaxonC G L V L I S
TaxonD G L E L I S

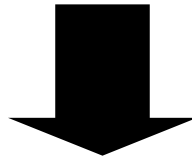
TaxonA	GGGTTTGGTTTAATT
TaxonB	GGATTTGGCTTAATC
TaxonC	GGATTTGGTTAAATT
TaxonD	GGATTTGGATAATTA

TaxonA	GGG	TTT	GG	TTT	AATT
TaxonB	GG	ATT	GG	CTT	AATC
TaxonC	GG	ATT	GG	TTA	AATT
TaxonD	GG	ATT	GG	ATA	ATTA



TaxonA	G	F	G	L	I
TaxonB	G	F	G	L	I
TaxonC	G	F	G	*	
TaxonD	G	F	G	*	

TaxonA	GGG	TTT	GGT	TTT	AATT
TaxonB	GG	ATT	GGC	TTA	ATC
TaxonC	GG	ATT	GGT	TAA	AATT
TaxonD	GG	ATT	GG	ATA	ATTA



TaxonA	G	F	G	L	I
TaxonB	G	F	G	L	I
TaxonC	G	F	G	*	
TaxonD	G	F	G	*	

TaxonA	GGG	TTT	GGT	TTT	AATT
TaxonB	GG	ATT	TGG	CTT	AATC
TaxonC	GG	ATT	TGG	TAA	AATT
TaxonD	GG	ATT	TGG	ATA	ATTA

多数のアミノ酸が消滅



TaxonA	G	F	G	L	I
TaxonB	G	F	G	L	I
TaxonC	G	F	G	*	
TaxonD	G	F	G	*	

TaxonA	GGG	TTT	GGT	TTT	AATT
TaxonB	GG	ATT	GGC	TTA	ATC
TaxonC	GG	ATT	GGT	TAA	ATT
TaxonD	GG	ATT	GG	ATA	ATT

多数のアミノ酸が消滅

TaxonA	G	F	G	L	I
TaxonB	G	F	G	L	I
TaxonC	G	F	G	*	
TaxonD	G	F	G	*	

TaxonA GGGTTTGGTTTAATT
TaxonB GGATTTGGGCAATC
TaxonC GGATTTGGGAAATT
TaxonD GGATTTGGGATAATTA

多数のアミノ酸が消滅

TaxonA G F G L I
TaxonB G F G I
TaxonC G F G *
TaxonD G F G *

分子系統樹に適したデータセットとは？

分子系統樹に適したデータセットとは？

- 「仮定」から明らかに逸脱した部分を含まない

分子系統樹に適したデータセットとは？

- 「仮定」から明らかに逸脱した部分を含まない
 - 分子進化モデルは 1 形質から 1 形質への変異を仮定している．以下の変異はこの仮定に反する

分子系統樹に適したデータセットとは？

- 「仮定」から明らかに逸脱した部分を含まない
 - 分子進化モデルは 1 形質から 1 形質への変異を仮定している．以下の変異はこの仮定に反する
 - フレームシフト突然変異や逆位

分子系統樹に適したデータセットとは？

- 「仮定」から明らかに逸脱した部分を含まない
 - 分子進化モデルは 1 形質から 1 形質への変異を仮定している．以下の変異はこの仮定に反する
 - フレームシフト突然変異や逆位
 - 開始・終始コドンの変異やイントロンのスプライセオソーム認識配列の変異

分子系統樹に適したデータセットとは？

- 「仮定」から明らかに逸脱した部分を含まない
 - 分子進化モデルは 1 形質から 1 形質への変異を仮定している．以下の変異はこの仮定に反する
 - フレームシフト突然変異や逆位
 - 開始・終始コドンの変異やイントロンのスプライセオソーム認識配列の変異
 - 挿入・欠失

分子系統樹に適したデータセットとは？

- 「仮定」から明らかに逸脱した部分を含まない
 - 分子進化モデルは 1 形質から 1 形質への変異を仮定している．以下の変異はこの仮定に反する
 - フレームシフト突然変異や逆位
 - 開始・終始コドンの変異やイントロンのスプライセオソーム認識配列の変異
 - 挿入・欠失
 - → 整列が信頼できるならギャップを - にすることである程度対応可能

分子系統樹に適したデータセットとは？

- 「仮定」から明らかに逸脱した部分を含まない
 - 分子進化モデルは 1 形質から 1 形質への変異を仮定している．以下の変異はこの仮定に反する
 - フレームシフト突然変異や逆位
 - 開始・終始コドンの変異やイントロンのスプライセオソーム認識配列の変異
 - 挿入・欠失
 - → 整列が信頼できるならギャップを - にすることである程度対応可能
 - 系統樹上で分子進化モデルは共通と仮定している．以下のデータはこの仮定に反する可能性が高い

分子系統樹に適したデータセットとは？

- 「仮定」から明らかに逸脱した部分を含まない
 - 分子進化モデルは 1 形質から 1 形質への変異を仮定している．以下の変異はこの仮定に反する
 - フレームシフト突然変異や逆位
 - 開始・終始コドンの変異やイントロンのスプライセオソーム認識配列の変異
 - 挿入・欠失
 - → 整列が信頼できるならギャップを - にすることである程度対応可能
 - 系統樹上で分子進化モデルは共通と仮定している．以下のデータはこの仮定に反する可能性が高い
 - タンパクコード塩基配列なのに遺伝暗号が共通でない
 - タンパクコード塩基配列でコドン使用頻度が共通でない
 - 塩基・アミノ酸配列で塩基頻度・アミノ酸頻度が共通でない

分子系統樹に適したデータセットとは？

分子系統樹に適したデータセットとは？

- 明らかに選択圧の異なる部分を含まない

分子系統樹に適したデータセットとは？

- 明らかに選択圧の異なる部分を含まない
 - 異なる分子進化モデルを適用可能であれば問題無し

分子系統樹に適したデータセットとは？

- 明らかに選択圧の異なる部分を含まない
 - 異なる分子進化モデルを適用可能であれば問題無し
 - あまりに局所的だと個別の分子進化モデルを適用してまで使う価値＝情報量は無い

分子系統樹に適したデータセットとは？

- 明らかに選択圧の異なる部分を含まない
 - 異なる分子進化モデルを適用可能であれば問題無し
 - あまりに局所的だと個別の分子進化モデルを適用してまで使う価値＝情報量はない
 - フレームシフト突然変異や逆位
 - 開始・終止コドンの変異やイントロンのスプライセオソーム認識配列の変異
 - 開始コドン
 - 終止コドン
 - 複数の遺伝子に共用されている部位

その他の注意点

その他の注意点

- 波形データが怪しければ縮重コードを用いる

その他の注意点

- 波形データが怪しければ縮重コードを用いる
 - もちろん除去するのがベスト

その他の注意点

- 波形データが怪しければ縮重コードを用いる
 - もちろん除去するのがベスト
- ギャップは「-」に、ギャップかどうかすら不明は「?」に、ギャップではないがどの塩基・アミノ酸か不明は「N/X」にして区別する

その他の注意点

- 波形データが怪しければ縮重コードを用いる
 - もちろん除去するのがベスト
- ギャップは「-」に、ギャップかどうかすら不明は「?」に、ギャップではないがどの塩基・アミノ酸か不明は「N/X」にして区別する
 - もちろん除去するのがベスト

その他の注意点

- 波形データが怪しければ縮重コードを用いる
 - もちろん除去するのがベスト
- ギャップは「-」に、ギャップかどうかすら不明は「？」に、ギャップではないがどの塩基・アミノ酸か不明は「N/X」にして区別する
 - もちろん除去するのがベスト
- **タンパクコード塩基配列では読み枠が変化しないように編集する**

その他の注意点

- 波形データが怪しければ縮重コードを用いる
 - もちろん除去するのがベスト
- ギャップは「-」に、ギャップかどうかすら不明は「?」に、ギャップではないがどの塩基・アミノ酸か不明は「N/X」にして区別する
 - もちろん除去するのがベスト
- タンパクコード塩基配列では読み枠が変化しないように編集する
- 削除した位置とその配列がすぐわかるようにする