

# ギャップ情報を利用した分子系統樹推定

筑波大・院生命環境科学 田辺晶史  
✉ a.s.tanabe@fifthdimension.jp  
✉ http://www.fifthdimension.jp/

## Summary

Q

多重配列整列と整列の不確実な領域の除去によって、ギャップ情報を用いた系統樹の推定法は改善or改悪するか？

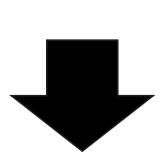
A

整列によって正解は減少し間違いは増加した。整列の不確実な領域の除去は大幅な正解の減少と間違いの増加を招いた。

## Introduction

ギャップ情報を用いる系統推定法は以下の2つがある

- ・整列と系統推定の同時計算
  - ・ギャップ情報を2値や多値形質データに符号化して加える
- ただし、それでは以下の問題がある
- ・前者は計算量が膨大で現在はほとんど使われていない
  - ・後者は符号化方法と置換モデルの最適な組み合わせが不明



最適なギャップ情報の符号化方法と置換モデルをシミュレーションデータ解析によって探索  
多重配列整列と整列の不確実な領域の除去がどのような影響を及ぼすかを調査

## Bayesian tree inference

MrBayes5D 3.1.2.2010.05.16を用いて系統推定

- ・置換モデルは配列生成時と同じ
- ・ASDSF $\leq 0.002$ でサンプル数1,000に達したら終了
- ・最大300万ステップ
- ・ASDSF $\leq 0.002$ にならなくても最後の1,000サンプルでASDSF $\leq 0.01$ であれば解析に用いた
- ・1,000本の系統樹から95%多数決合意樹を作成
- ・当てはめたモデルは以下の3つ

Equal frequency (ef) model	Variable frequency (vf) model	Variable composition (vc) model
To	To	From
0 1	0 1	$r\pi_0$
From 0	From 1	$r\pi_1$

0から1と1から0への変異確率が等しい

$\pi_0$ : 0の頻度

$\pi_1$ : 1の頻度

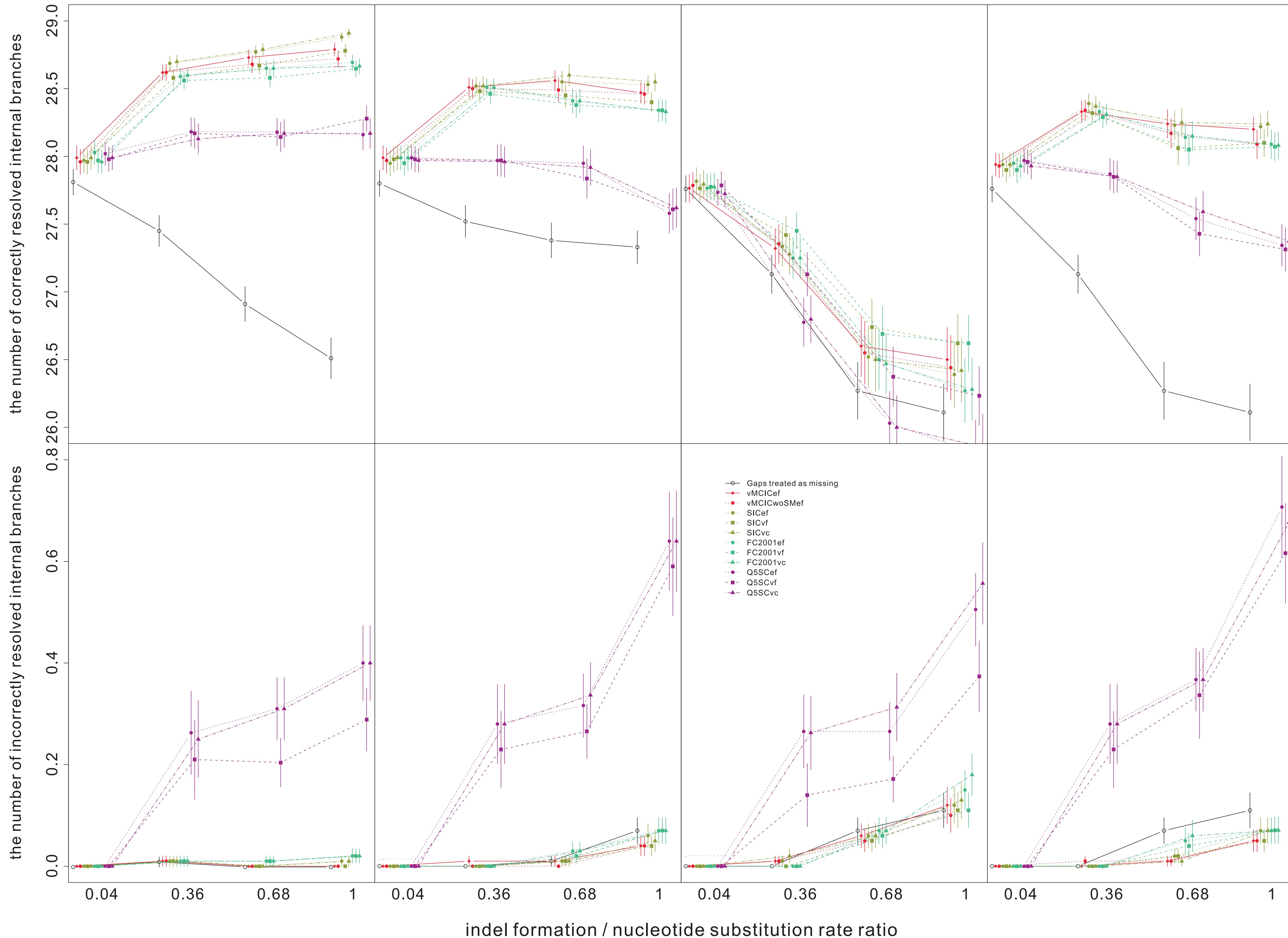
## Results

true alignment data

aligned data

trimmed data

trimmed seq + aligned gap



置換モデル間の違いはほとんど見られなかった

符号化法はSICとvMCICが優れていた

整列は正解の減少と間違いの増加を招いた

除去も正解の減少と間違いの増加を招いた

除去前の配列から作成したギャップ情報配列を除去後の配列に組み合わせると正解は増加し間違いも減少したが、除去前には劣る

## Discussion

01の頻度で変異確率の違いを表すことはできていない  
vfモデルは無意味

vcモデルの効果が確認できなかったのは、indelの起きやすさの座位間の違いを配列生成に組み込んでいないからかも

gap extensionが配列生成に組み込まれていない  
FC2001やQ5SCに不利になっているかも

非対称型系統樹では除去量が極めて多くなる  
対称型系統樹やより長い配列では異なる結果になるかも