Likelihood theory of DNA insertions/deletions and truthful Multiple Sequence Alignment (MSA)
-- background and outline --

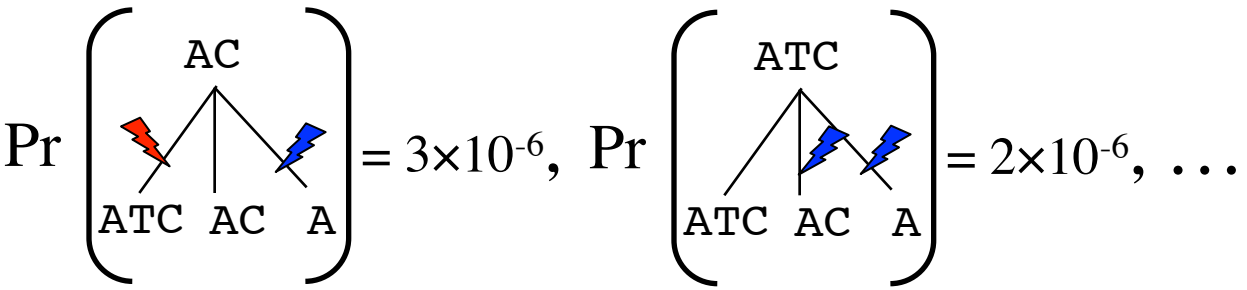
DNA 挿入/欠失の尤度理論と正しい Multiple Sequence Alignment (MSA)
-- 背景と概要 --

矢田研セミナー（2014/06/18）
担当: 江澤 潔

プロジェクトの目的

1. **MSA** (multiple sequence alignment) から過去のDNA **挿入/欠失**（insertions/deletions）を推定する**尤度理論**（likelihood theory）を構築し、

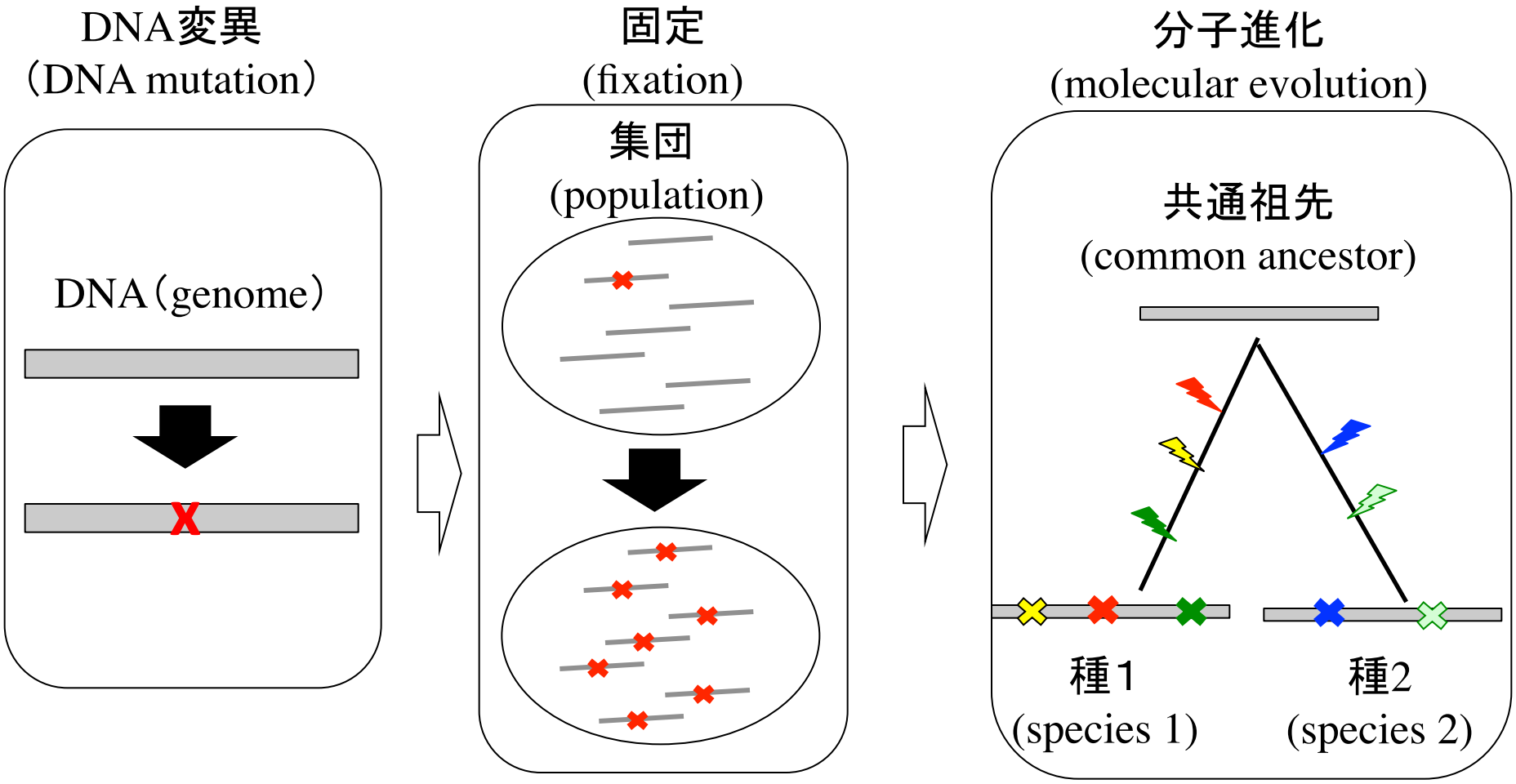
```
Seq1   ATC
Seq2   A-C
Seq3   A--
```



2. その理論を利用して、**MSA** を「**より正しく**」復元する方法を開発する。

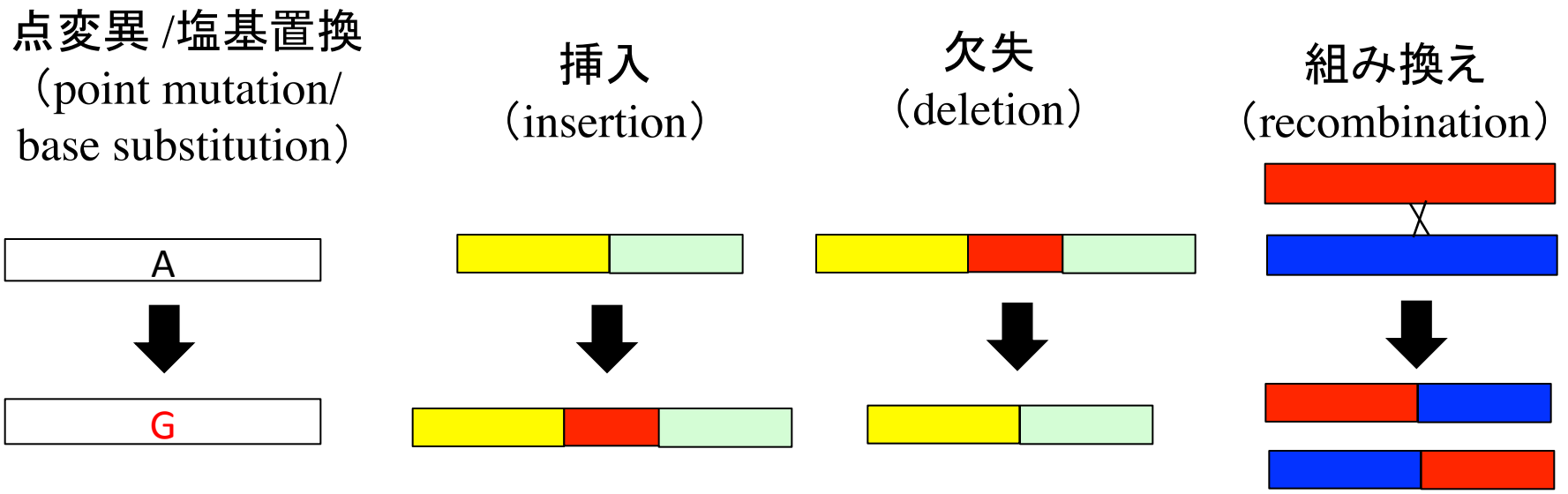
このセミナーでは、プロジェクトの背景と全体像について説明する。

DNA変異は分子進化の駆動力



分子進化は(集団に固定した)DNA変異が蓄積した結果と考えることができる。

DNA変異のタイプ



＊組み換えは通常、同一集団内で起こり、種間配列解析で検出するのは難しい。
[が、異所性組み換え（ectopic recombination）や水平伝搬（horizontal transfer）等は種間配列解析でも検出し得る。]

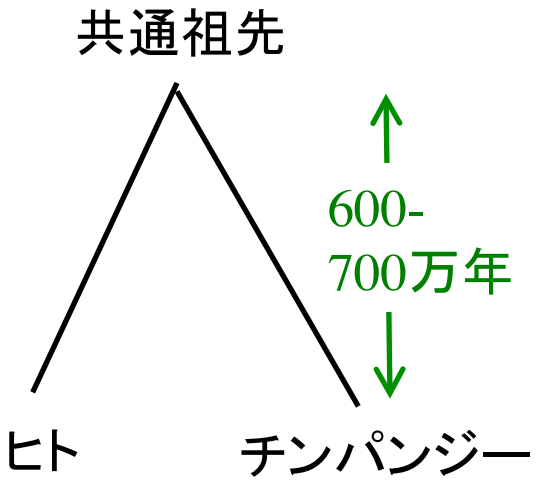
＊　その他、重複(duplication), 逆位(inversion) 等の **genome rearrangements** も重要な変異であるが、(global) alignment では（広い意味で）挿入/欠失とみなされる（こともある）。

挿入/欠失はDNA変化の大部分を占める

例）ヒト（human）とチンパンジー（chimpanzee）のゲノム比較
（e.g., Britten, 2002;
The international Chimpanzee Chromosome 22 Consortium, 2004;
The Chimpanzee Sequencing and Analysis Consortium, 2005）

塩基置換による違いは 1.2-1.5%
挿入/欠失も加えると　　4-5%

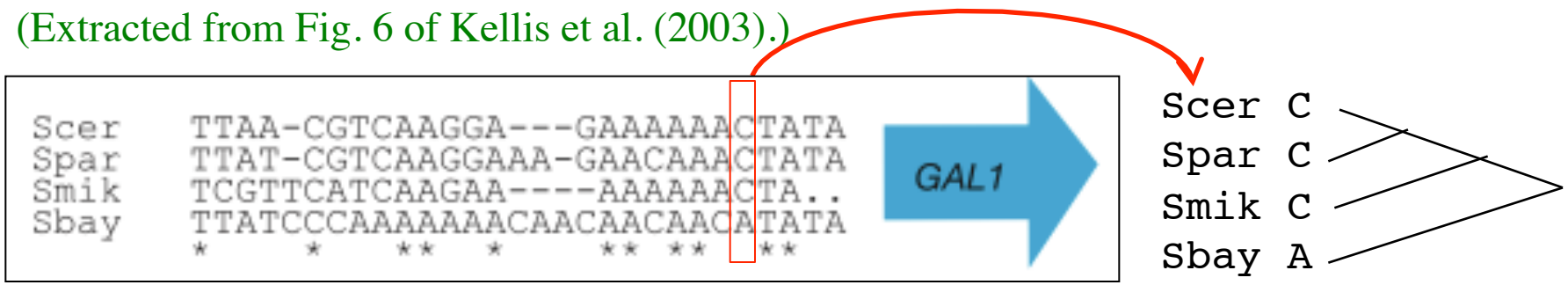
… 挿入/欠失はゲノム間の違い（塩基数）
　の約70%を占める！



他の種の近縁ゲノム比較でも挿入/欠失が塩基の違いの大部分（75-98%）を占める（Britten et al., 2003）

（問題点1）現在の分子進化解析は点変異中心

(Extracted from Fig. 6 of Kellis et al. (2003).)



（理由１）データの入手しやすさ。

、、、多少質の悪いMSAでも「良質な」領域を選んで解析できる。

（理由２）（尤度に基づく）理論の確立と便利な解析ツールの存在。

（e.g., Felsenstein, 2004; Yang, 2006）

しかし、、、ゲノム変化の大多数を知るには、**DNA 挿入/欠失**もまじめに調べなければならない。

=> 正しい multiple sequence alignment (MSA) と、

DNA 挿入/欠失を推定する理論　が必須になる。

MSAは（相同）配列解析の Holy Grail (Gusfield, 1997)

Holy Grail:
Jesus の最後の晩餐で使われた丼（bowl）、杯（cup）、皿（plate）。
（Arthur 王伝説等に登場し、騎士達が追い求めたらしい。）
=> （比喩）大いなる欲望と追求の対象。

(Image from http://home.messiah.edu/~tp1180/page%204.html)

MSA (multiple sequence alignment)

(Extracted from Fig. 6 of Kellis et al. (2003).)

MSAは(相同)配列解析の Holy Grail (Gusfield, 1997)

Holy Grail:
Jesus の最後の晩餐で使われた丼(bowl)、杯(cup)、皿(plate)。
(Arthur 王伝説等に登場し、騎士達が追い求めたらしい。)
=> (比喩)大いなる欲望と追求の対象。

(Image from http://home.messiah.edu/~tp1180/page%204.html)

MSA (multiple sequence alignment)

(Extracted from Fig. 6 of Kellis et al. (2003).)

進化解析

系統関係推定
進化速度推定
自然選択圧推定
変異/進化 pattern 推定
進化 events /祖先状態推定

構造解析

2次構造推定
3次構造推定
4次構造推定

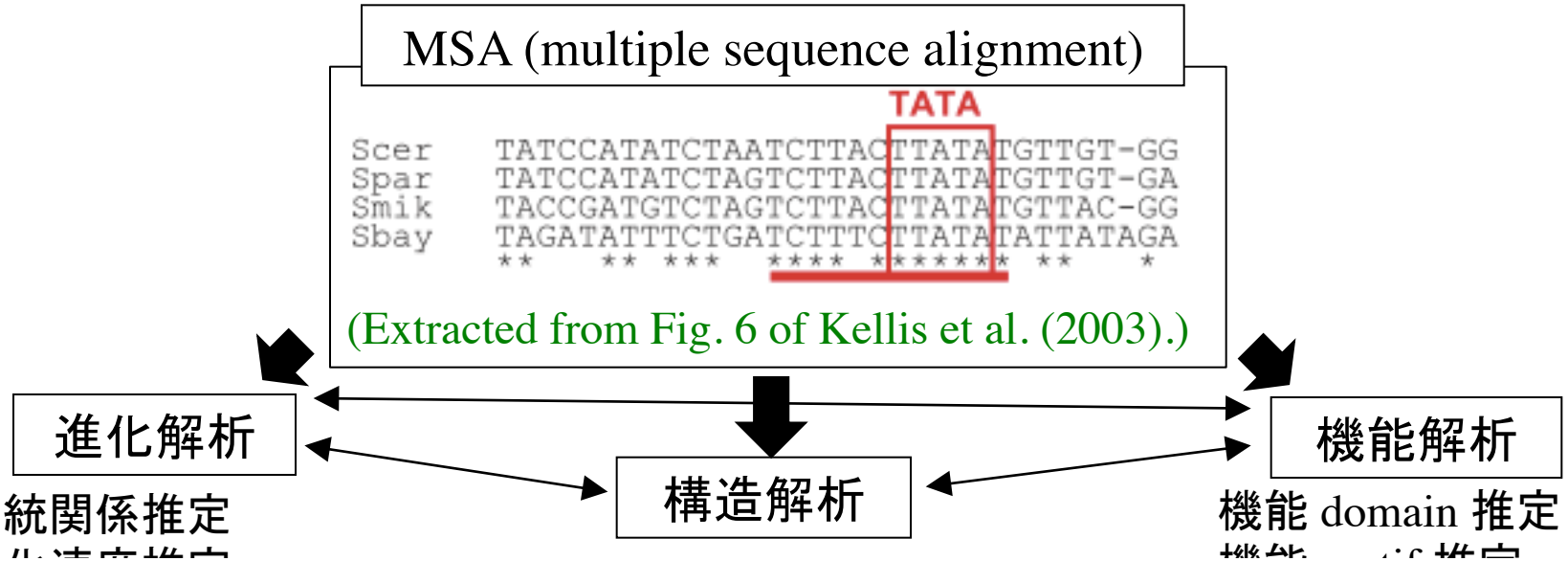
機能解析

機能 domain 推定
機能 motif 推定
相互作用部位/領域推定
活性部位推定
病因変異推定(?)

MSAは(相同)配列解析の Holy Grail (Gusfield, 1997)

Holy Grail:
Jesus の最後の晩餐で使われた丼(bowl)、杯(cup)、皿(plate)。
(Arthur 王伝説等に登場し、騎士達が追い求めたらしい。)
=> (比喩)大いなる欲望と追求の対象。

(Image from http://home.messiah.edu/~tp1180/page%204.html)

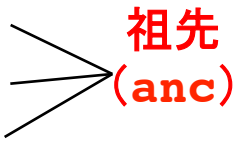


より正確な MSAを得ることは分子生物学全体の進展につながる

正しいMSA は分子進化 events の推定に使える（1/2）

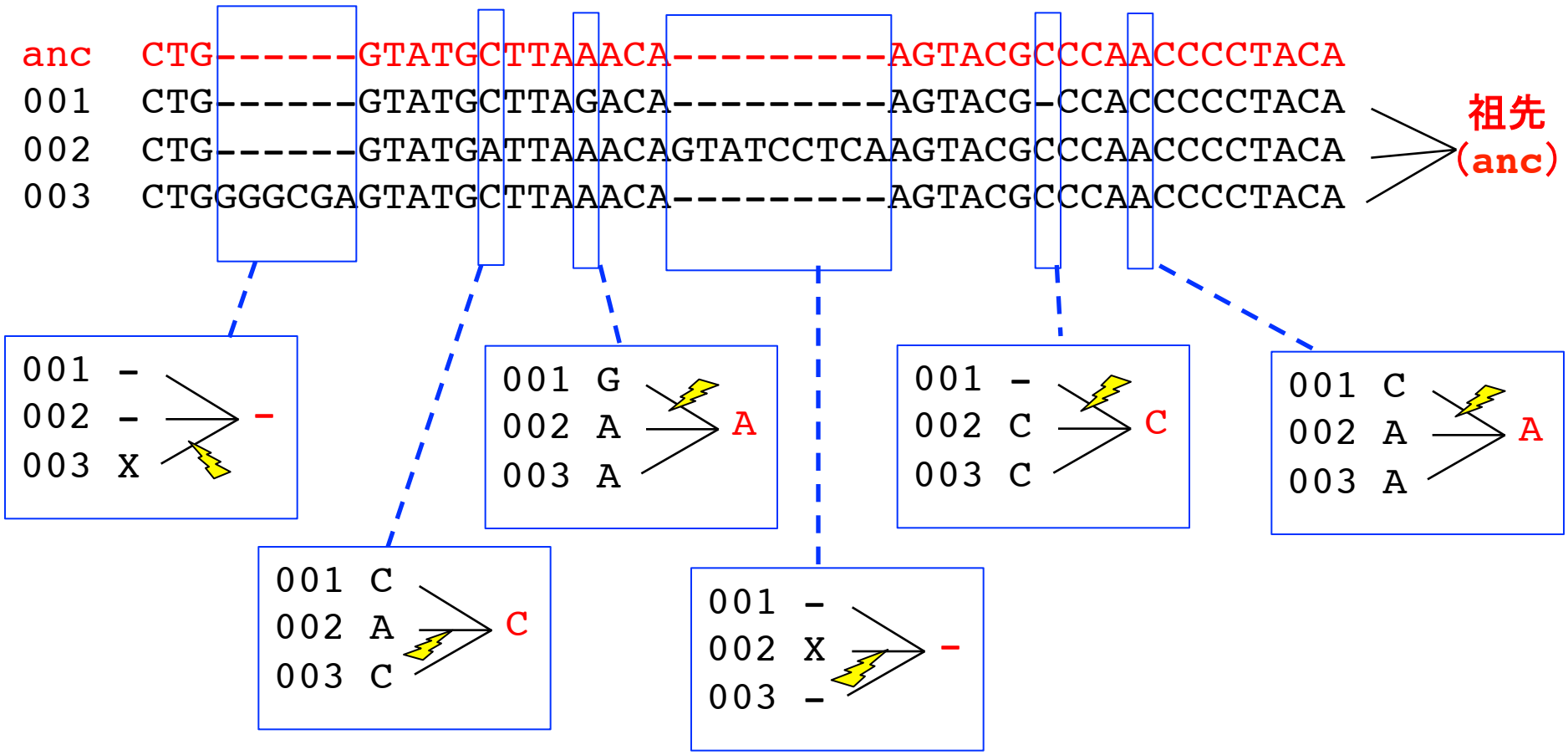
（枝長：すべて 0.05 置換/塩基； 挿入頻度＝欠失頻度＝0.005 event/site）

```
anc   CTG------GTATGCTTAAACA---------AGTACGCCCAACCCCTACA
001   CTG------GTATGCTTAGACA---------AGTACG-CCACCCCCTACA
002   CTG------GTATGATTAAACAGTATCCTCAAGTACGCCCAACCCCTACA
003   CTGGGGCGAGTATGCTTAAACA---------AGTACGCCCAACCCCTACA
```



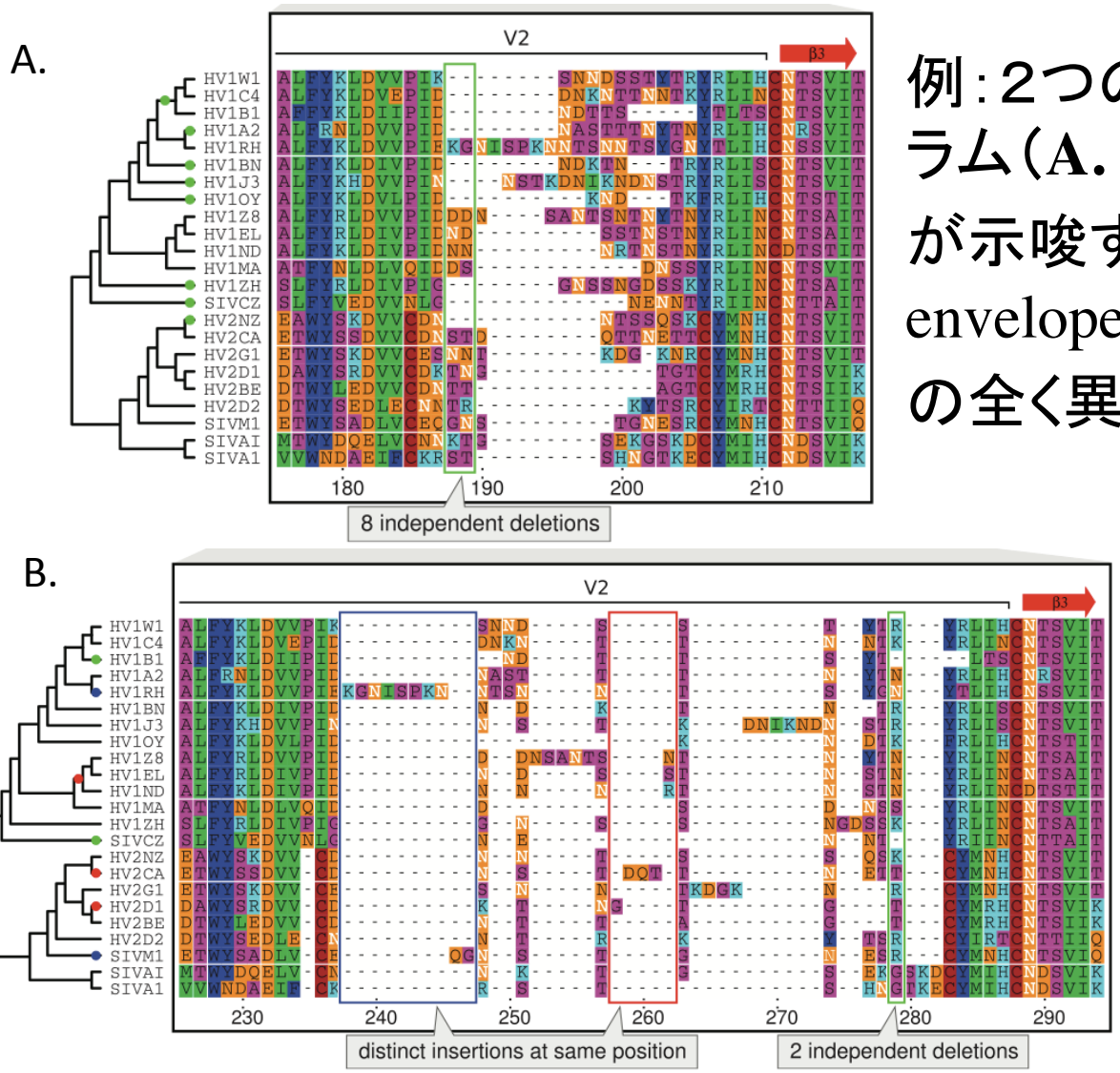
正しいMSA は分子進化 events の推定に使える(2/2)

(枝長:すべて 0.05 置換/塩基; 挿入頻度＝欠失頻度＝0.005 event/site)



＊ 枝が短いときは再節約法(maximum parsimony)が充分良く近似する。
(枝が長いときは events /祖先状態の確率分布を用いた方が安全。)

（問題点2）現存のMSA 復元プログラムは十分に正確ではない (1/3)



例：2つの異なるMSA復元プログラム（**A. ClustalW; B. PRANK$_{+F}$**）が示唆するHIV & SIV の envelope 糖タンパク、gp120、の全く異なる進化シナリオ。

（From Figure 1 of Löytynoja and Goldman (2008).）

- - - MSA errors は分子進化機構の推定 errors に直接つながる!!

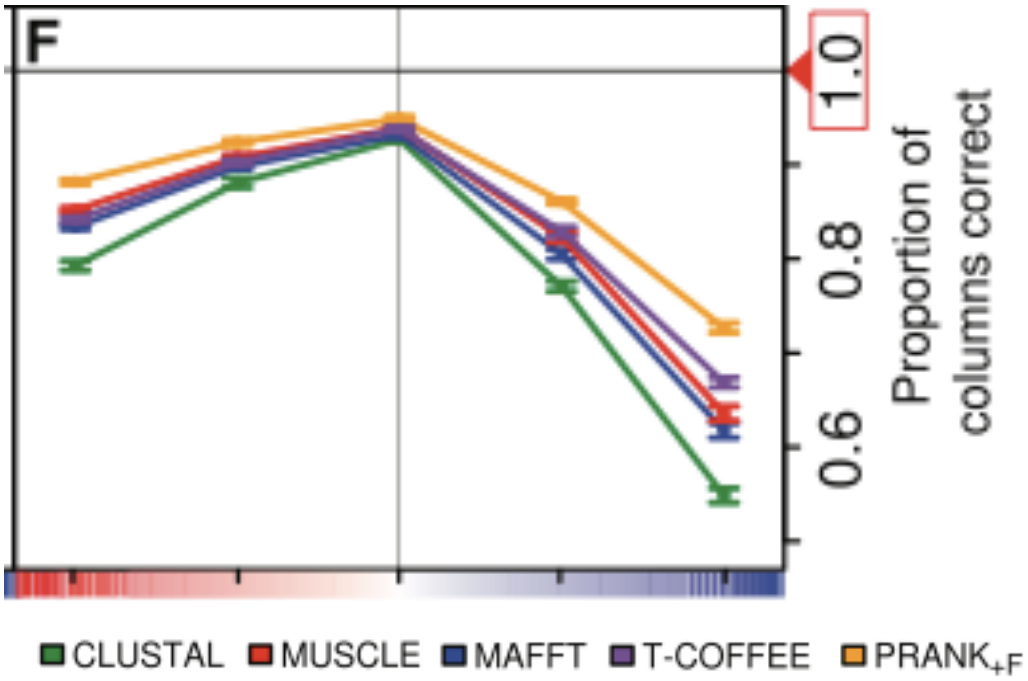
（問題点2）現存のMSA 復元プログラムは十分に正確ではない (2/3)

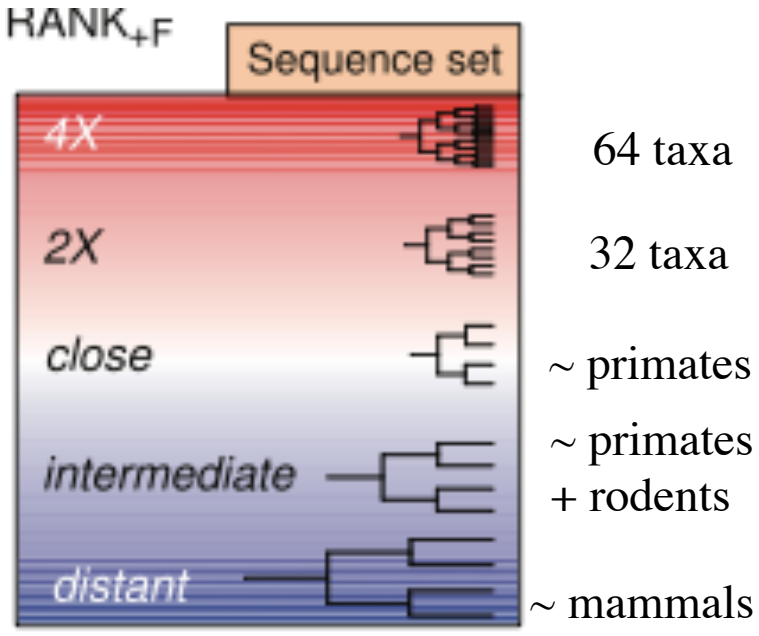
間違った MSA columns（列）の割合 :

5%-50%　（Loytynoja and Goldman, 2008）

50%-90%　（Landan and Graur, 2009）

（条件によってかなり異なるが、一般的には無視できない。）





(Extracted from Figure 3 of Löytynoja and Goldman (2008))

（問題点2）現存のMSA 復元プログラムは十分に正確ではない (3/3)

MSA 復元 errors の主な要因:

（要因１）MSA空間探索は本質的に不完全;

（要因2）score 計算がDNA挿入/欠失の歴史には無頓着;

（要因3）Affine gap-penalty は実際の挿入/欠失の長さ分布に fit しない;

（要因4）そもそも、分子進化は確率過程、従って、（真の）最適解が本当のMSAとは限らない。

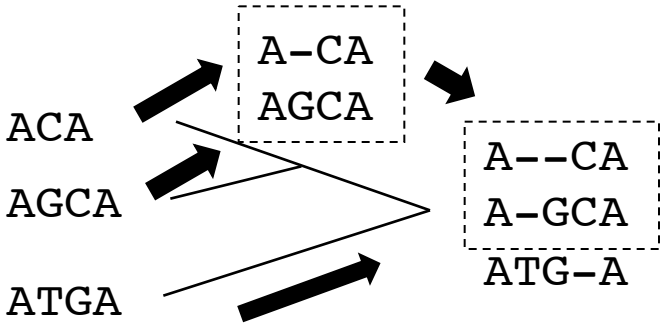
. . . 以下、各々の要因について説明する。

（要因１）MSA空間探索は本質的に不完全（1/2）

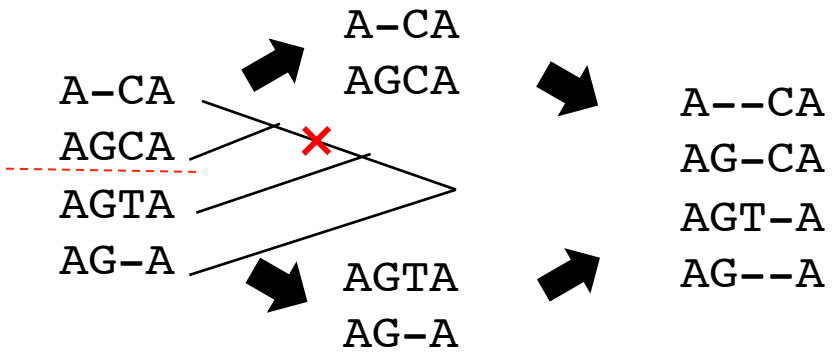
網羅的なMSA空間探索は実質的には不可能である。
（N 個の長さ L の配列を align するには、少なくとも $O(L^N)$ の時間を要する）

=> 現在のほぼすべてのMSA復元は
（group 対 group）pairwise alignment (PWA) の繰り返しによる:

Progressive alignment



Iterative refinement



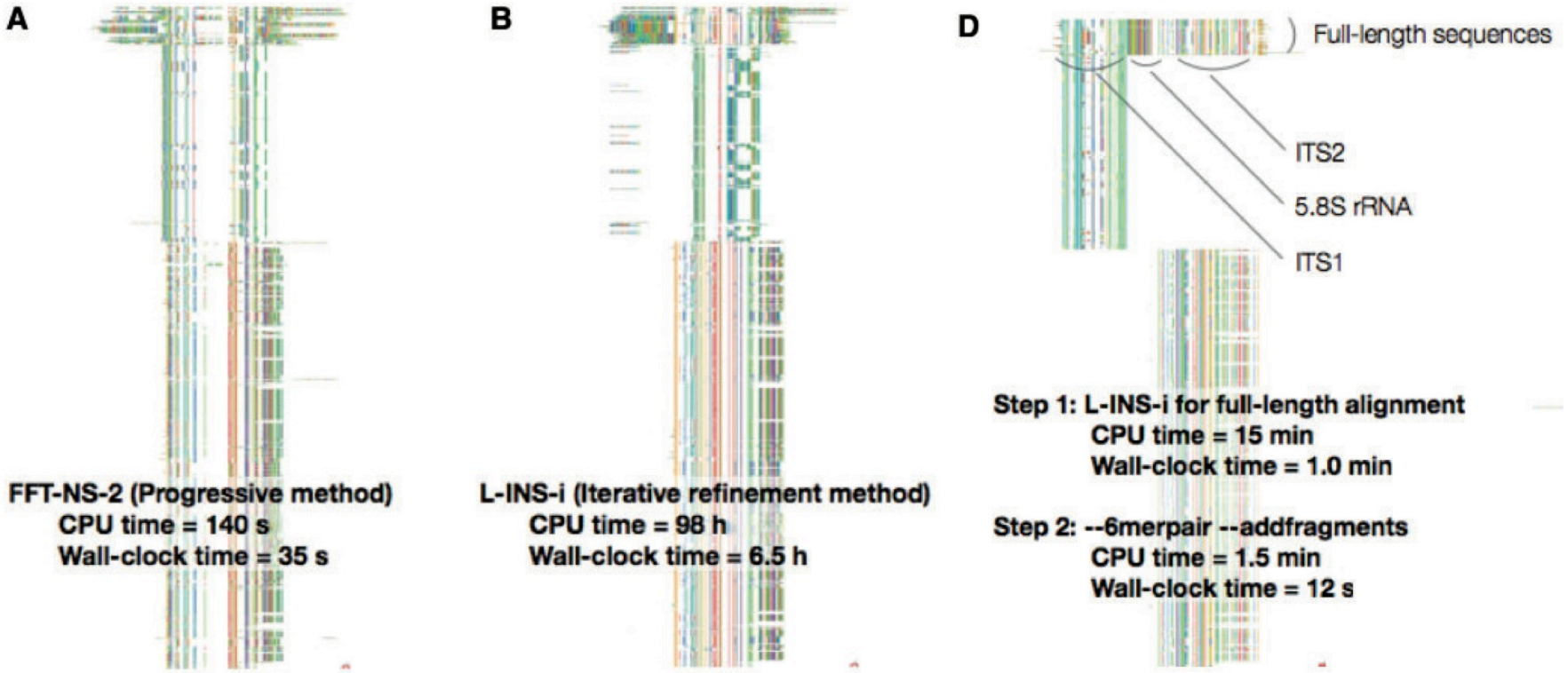
時間はかなり節約できる（典型的には $O(N \times L^2)$ ）が、
局所的最適解（local optimum）に trap される恐れがある。

（要因１）MSA空間探索は本質的に不完全（2/2）

不完全なMSA空間探索による「悲劇」の例：

MAFFT の標準 options で得られる明らかに間違った復元MSAs

筆者推奨の二段階法で得られる“正しい” 復元MSAs

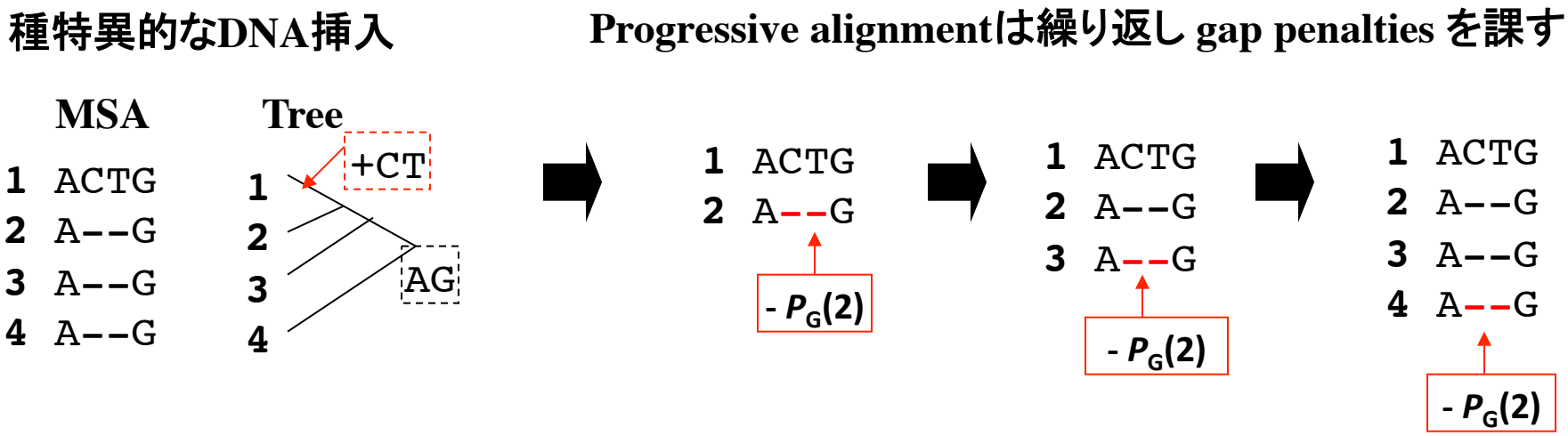


(Extracted from Figure 2 of Katoh and Standley (2013).)

(要因2)現在主流の score 計算はDNA挿入/欠失の歴史には無頓着(1/2)

- 現在主流の MSA プログラムは(特にタンパク質や構造RNA等の)保存領域の同定に重点を置いている。

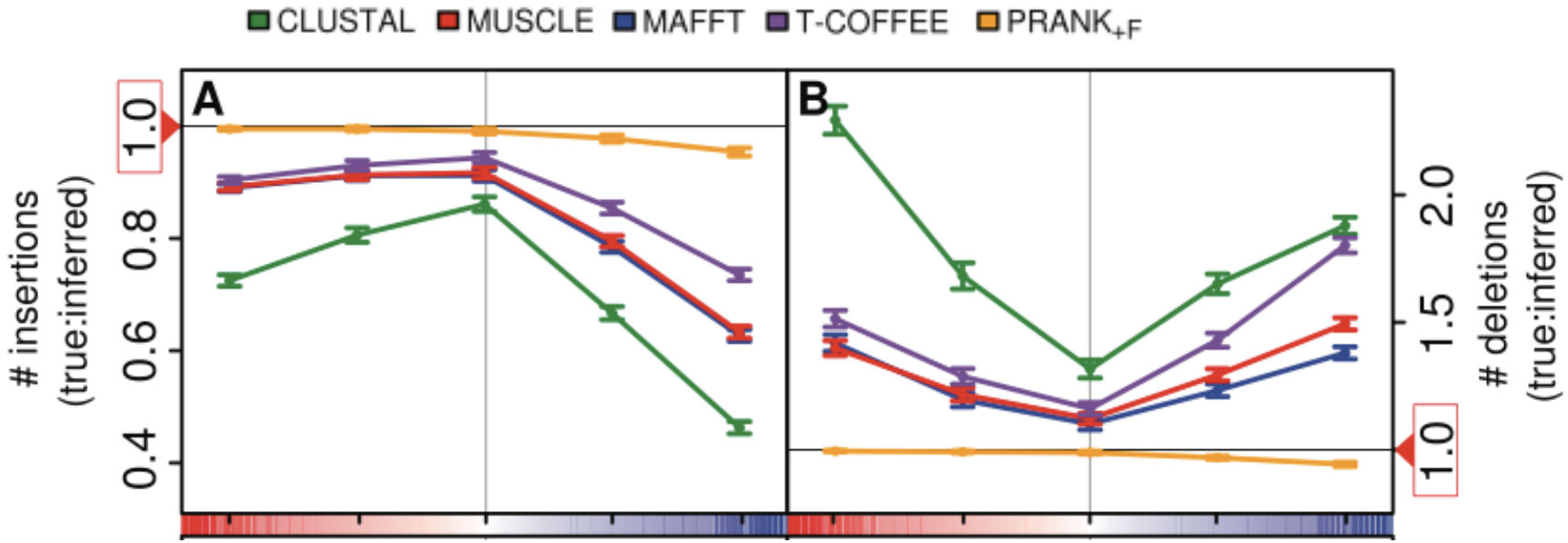
  …多くの heuristics (それが適用できない問題には無力)を導入。
- MSA errors の一因は、(特に progressive alignment での) **gap penalties の計算法** にもある(例:下図)。



(要因2)現在主流の score 計算はDNA挿入/欠失の歴史には無頓着(2/2)

実際、現存のMSA 復元プログラムのほとんど(PRANK以外)は、

DNA挿入の頻度を過小評価し、DNA欠失の頻度を過大評価する。



(Extracted from Figure 3 of Löytynoja and Goldman (2008))

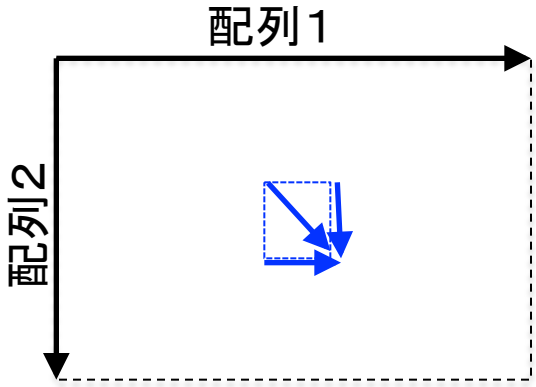
(過去のデータ解析でしばしば観測された「DNA欠失の優勢」
も実はこの estimation bias によるのかも知れない、、、)

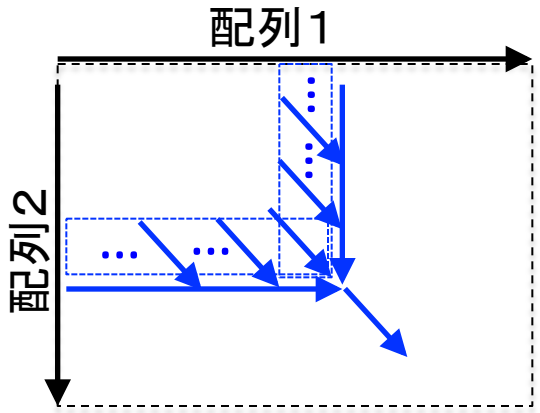
（要因3）Affine gap-penalty（幾何分布）は実際の挿入/欠失の長さ分布に fit しない（1/3）

I. 現在の alignment 復元で常用されているのはAffine gap-penalty

初期の Dynamic Programming (DP)
(Needlesman and Wunsch, 1970)

改良型 DP (Gotoh, 1982)





Linear gap penalty: $P_G(l) = \alpha l$

Affine gap penalty: $P_G(l) = \alpha l + \beta$

一塩基挿入/欠失 model

挿入/欠失長は幾何分布: $P_{I/D}[l] = (1 - \varepsilon)\,\varepsilon^{l-1}$

$[\ \varepsilon = \exp(-\alpha)\ ]$

（要因3）Affine gap-penalty（幾何分布）は実際の挿入/欠失の長さ分布に fit しない（2/3）

II. 実際の挿入／欠失長は power-law に従う

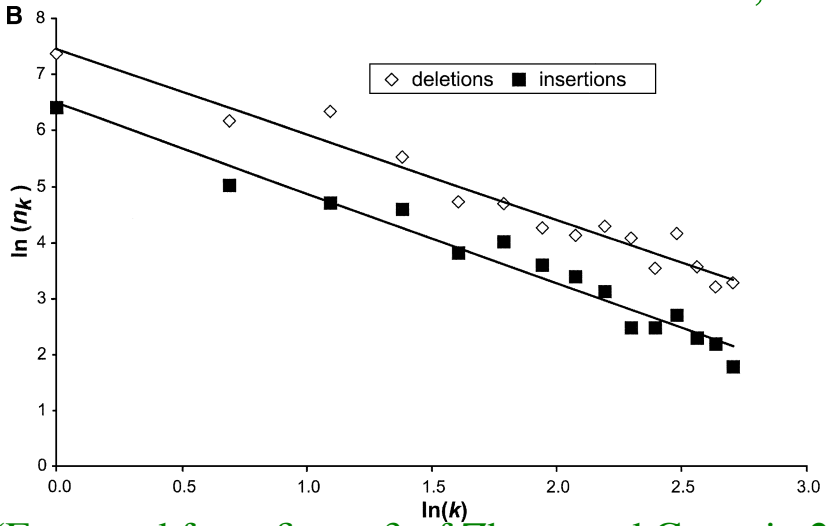
過去の大規模解析では、挿入/欠失された DNA /アミノ酸配列の長さは

power-law: $P_{I/D}[l] = C\, l^{-\gamma}$ $(\gamma = 1\sim2)$

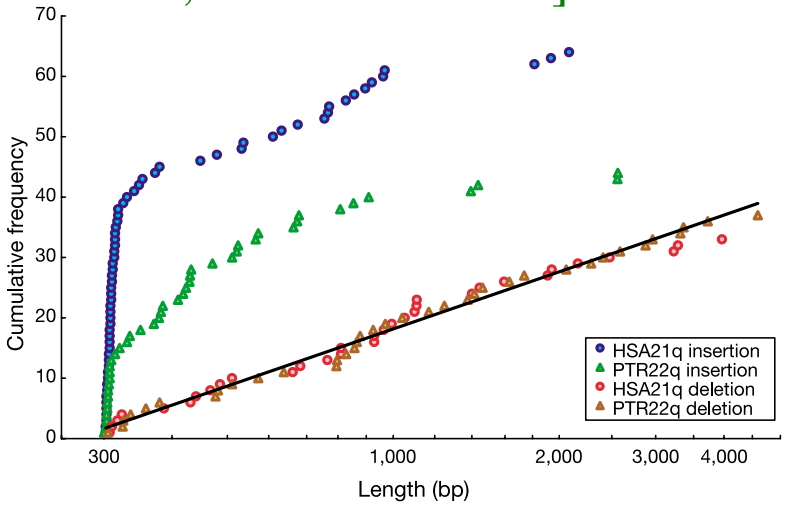
に従うことが観測された。

[タンパク質 : Gonnet et al. 1992; Benner et al. 1993; Chang and Benner 2004.

DNA : Gu and Li 1995; Zhang and Gerstein 2003; The International Chimpanzee Chromosome 22 Consortium 2004; Yamane et al. 2006; Fan et al. 2007.]



(Extracted from figure 3 of Zhang and Gerstein 2003)

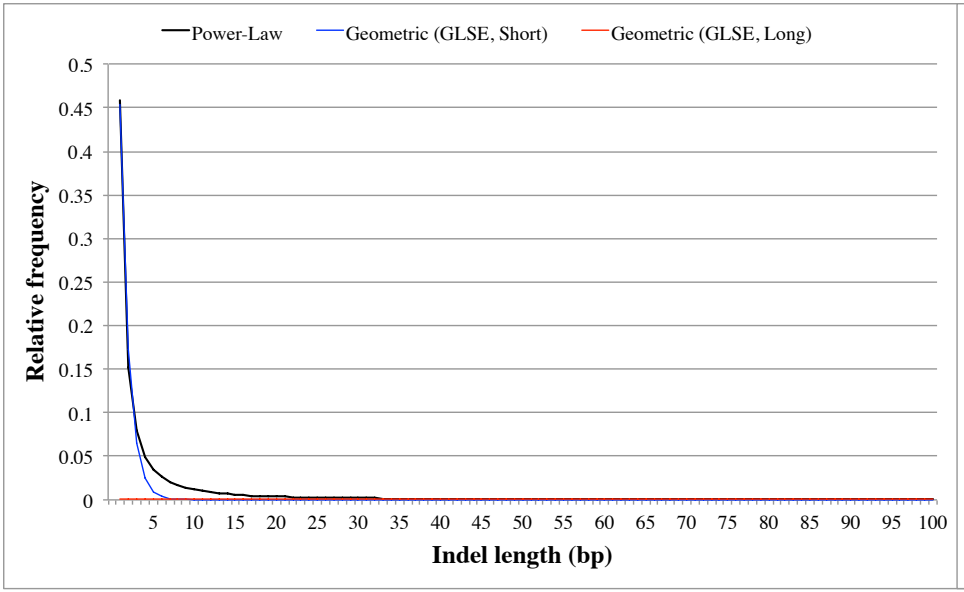


(Figure 3 of Chimp Chr. 22 Consortium 2004)

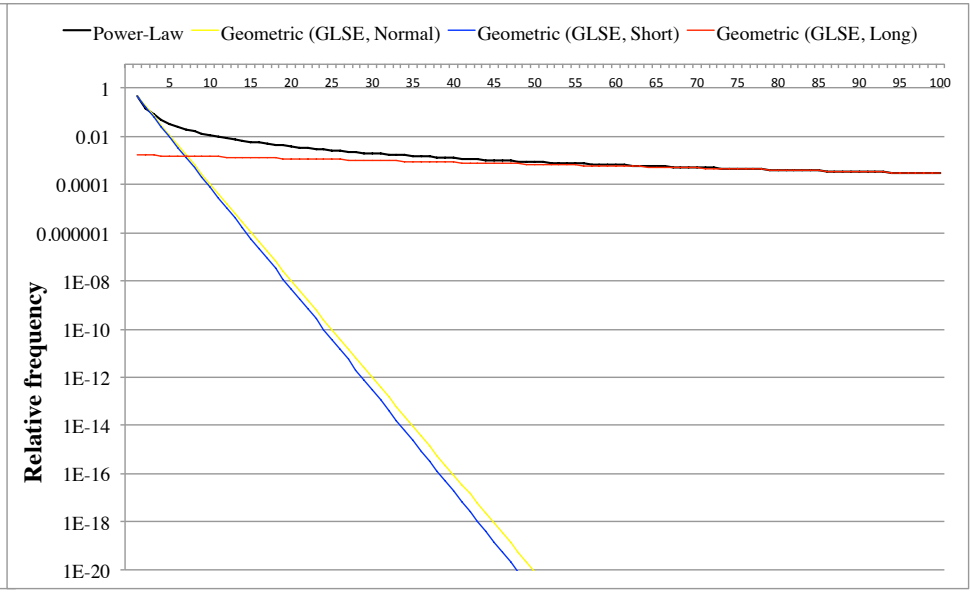
（要因3）Affine gap-penalty（幾何分布）は実際の挿入/欠失の長さ分布に fit しない（3/3）

III. 幾何分布（geometric distribution）は power-law に fit しない

最小二乗 (LS) fitの結果（線型表示）



最小二乗 (LS) fitの結果（対数表示）

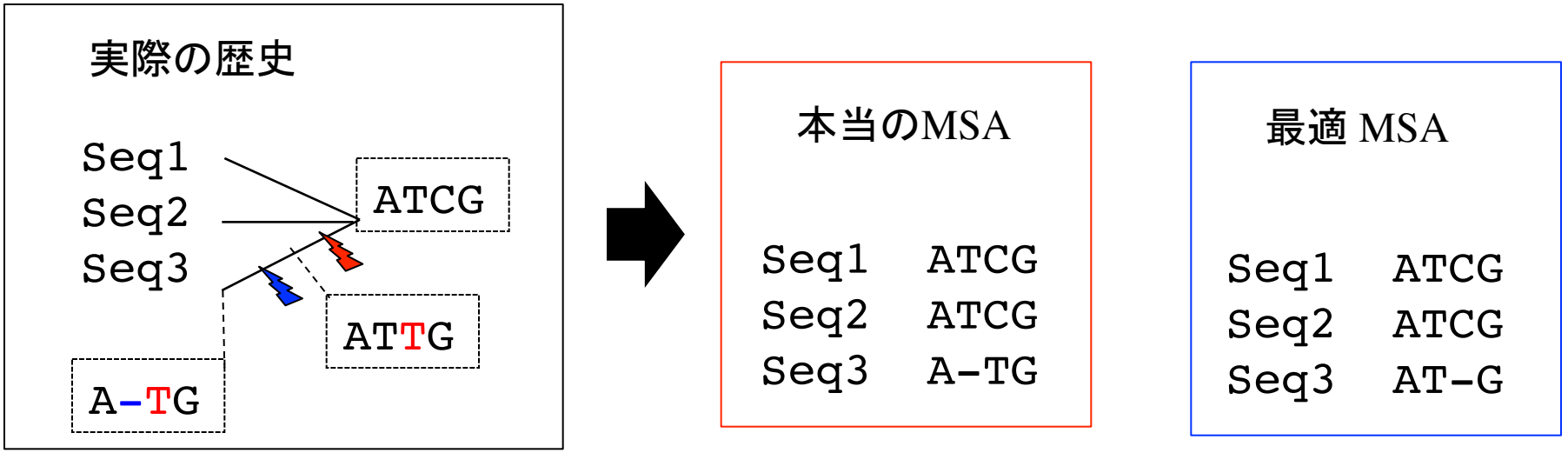


実際、power-law を用いると PWA や(PW)配列比較解析の精度や整合性が向上する (Cartwright 2006, 2009)。

(要因4)そもそも、分子進化は確率過程

仮に「完璧な」MSA score を用いて「完全な」MSA空間探索をして、
「(真の)最適MSA」が得られたとしても、
それが本当のMSAであるとは限らない。

例)



実際、ある大規模 simulation 解析(Lunter et al. 2008)によると、
PWA errors の大半はこのような分子進化の偶然性(**stochasticity**)が
原因であると見積もられた。

問題点のまとめ(1)

(問題点1)現在の解析は点変異中心

… 進化上ののDNA変化の約30%しか扱えない。

=> 残り 70% を扱うには、

DNA 挿入/欠失に着目する必要。その為に、

* 挿入/欠失を解析する理論

と

* 正しいMSA

が必要。

問題点のまとめ(1)と解決策

(問題点1)現在の解析は点変異中心

… 進化上ののDNA変化の約30%しか扱えない。

=> 残り 70% を扱うには、

DNA 挿入/欠失に着目する必要。その為に、

* 挿入/欠失を解析する理論<= 挿入/欠失の尤度理論の構築

と

* 正しいMSA

が必要。しかし、、、

問題点のまとめ(2)

（問題点２）現在のMSA復元は十分に正確ではない

（要因１）MSA空間探索は本質的に不完全;

（要因2）score 計算がDNA挿入/欠失の歴史には無頓着;

（要因3）Affine gap-penalty (幾何分布)は実際の挿入/欠失の長さ分布（power-law） に fit しない;

（要因4）そもそも、分子進化は確率過程、従って、（真の）最適解が本当のMSAとは限らない。

問題点(2)の解決策(案)(1)

(問題点2)現在のMSA復元は十分に正確ではない

(要因1)MSA空間探索は本質的に不完全;

(要因2)score 計算がDNA挿入/欠失の歴史には無頓着;

(要因3)Affine gap-penalty (幾何分布)は実際の挿入/欠失の長さ分布(power-law) に fit しない;

(要因4)そもそも、分子進化は確率過程、従って、(真の)最適解が本当のMSAとは限らない。

共/準最適解の含包

挿入/欠失の尤度理論×塩基置換の尤度理論

挿入/欠失の尤度理論×塩基置換の尤度理論を用いたMSA score（1/2）

（基本的考え）

MSAの**尤度**（正確には、MSAの条件付き確率）:

$$\Pr[\text{MSA} \mid \text{進化モデル}]$$

は、**理想的なMSA score**として使える筈である。

（理由１）原則として、尤度が高ければ高い程、そのMSAは実現しやすい;

（理由２）尤度は、分子進化過程（置換、挿入/欠失）の歴史をきちんと考慮して計算される（次のスライド）

=> 問題点２の要因２は自然に克服できる;

（理由３）実際の挿入/欠失長の分布（*e.g.*, power-law）を考慮に入れれば更に精度が向上する筈である。

挿入/欠失の尤度理論×塩基置換の尤度理論を用いたMSA score（2/2）

（第一近似では、）MSAの尤度は挿入/欠失部分と塩基置換部分に分解できる。

例）

$$\Pr\left(\begin{matrix}1 & \mathrm{ATCG}\\ 2 & \mathrm{ATCG}\\ 3 & \mathrm{A-TG}\end{matrix}\middle|\ \text{進化モデル}\right) = \Pr\left(\begin{matrix}1 & \mathrm{NNNN}\\ 2 & \mathrm{NNNN}\\ 3 & \mathrm{N-NN}\end{matrix}\middle|\ \text{挿入/欠失モデル}\right) \times \Pr\left(\begin{matrix}1 & \mathrm{ATCG}\\ 2 & \mathrm{ATCG}\\ 3 & \mathrm{ANTG}\end{matrix}\middle|\ \text{塩基置換モデル}\right)$$

$$\Pr\left(\begin{matrix}1 & \mathrm{NNNN}\\ 2 & \mathrm{NNNN}\\ 3 & \mathrm{N-NN}\end{matrix}\middle|\ \text{挿入/欠失モデル}\right) = \Pr\left(\begin{matrix}1\\ 2\\ 3\end{matrix}\middle|\ \begin{matrix}\mathrm{NNNN}\\ \mathrm{N-NN}\end{matrix}\right) + \Pr\left(\begin{matrix}1\\ 2\\ 3\end{matrix}\middle|\ \begin{matrix}\mathrm{NNNN}\\ \mathrm{NNNNN}\\ \mathrm{N-NN}\end{matrix}\right) + \ldots$$

$$\Pr\left(\begin{matrix}1 & \mathrm{ATCG}\\ 2 & \mathrm{ATCG}\\ 3 & \mathrm{ANTG}\end{matrix}\middle|\ \text{塩基置換モデル}\right) = \Pr\left(\begin{matrix}1\\ 2\\ 3\end{matrix}\middle|\ \begin{matrix}\mathrm{ATCG}\\ \mathrm{ANTG}\end{matrix}\right) + \Pr\left(\begin{matrix}1\\ 2\\ 3\end{matrix}\middle|\ \begin{matrix}\mathrm{ATCG}\\ \mathrm{ATTG}\\ \mathrm{ANTG}\end{matrix}\right) + \ldots$$

（問題点A）計算時間が余計にかかる（PWAでも最低 $O(L^3)$ ）。

共/準最適解の含包

共/準最適解 (co-/sub-optimum solution):

最適解（optimum solution）と同じ、もしくは少しだけ低い尤度を持つ MSA。

（案）これらは少なからぬ実現可能性があるので、尤度で重み付けして

「選択肢」に含めることにより、本当のMSAを見逃す可能性を減らす。

以前の例）

最適 MSA

$$\Pr\left(\begin{array}{ll}1 & \mathtt{ATCG}\\ 2 & \mathtt{ATCG}\\ 3 & \mathtt{AT\text{-}G}\end{array}\middle|\ \text{進化モデル}\right) = 6\times10^{-5},$$

本当のMSA（準最適）

$$\Pr\left(\begin{array}{ll}1 & \mathtt{ATCG}\\ 2 & \mathtt{ATCG}\\ 3 & \mathtt{A\text{-}TG}\end{array}\middle|\ \text{進化モデル}\right) = 1\times10^{-5}, \ldots$$

「推定 MSA」 =

$$\begin{array}{ll}1 & \mathtt{ATCG}\\ 2 & \mathtt{ATCG}\\ 3 & \mathtt{AT\text{-}G}\end{array} \quad 72\% \quad + \quad \begin{array}{ll}1 & \mathtt{ATCG}\\ 2 & \mathtt{ATCG}\\ 3 & \mathtt{A\text{-}TG}\end{array} \quad 12\% \quad + \ldots$$

問題点(2)の解決策(案)(1')

(問題点2)現在のMSA復元は十分に正確ではない

(要因1)MSA空間探索は本質的に不完全;

(要因2)score 計算がDNA挿入/欠失の歴史には無頓着;

(要因3)Affine gap-penalty (幾何分布)は実際の挿入/欠失の長さ分布(power-law) に fit しない;

(要因4)そもそも、分子進化は確率過程、従って、(真の)最適解が本当のMSAとは限らない。

共/準最適解の含包

挿入/欠失の尤度理論×塩基置換の尤度理論

- - - (問題点A)計算時間が余計にかかる

問題点(2)の解決策(案)(2)

局所的相同性を利用した
効率的な MSA 空間探索

(問題点2)現在のMSA復元は十分に正確ではない

(要因1)MSA空間探索は本質的に不完全;

(要因2)score 計算がDNA挿入/欠失の歴史には無頓着;

(要因3)Affine gap-penalty (幾何分布)は実際の挿入/欠失の長さ分布(power-law) に fit しない;

(要因4)そもそも、分子進化は確率過程、従って、(真の)最適解が本当のMSAとは限らない。

共/準最適解の含包

挿入/欠失の尤度理論×塩基置換の尤度理論

- - - (問題点A)計算時間が余計にかかる

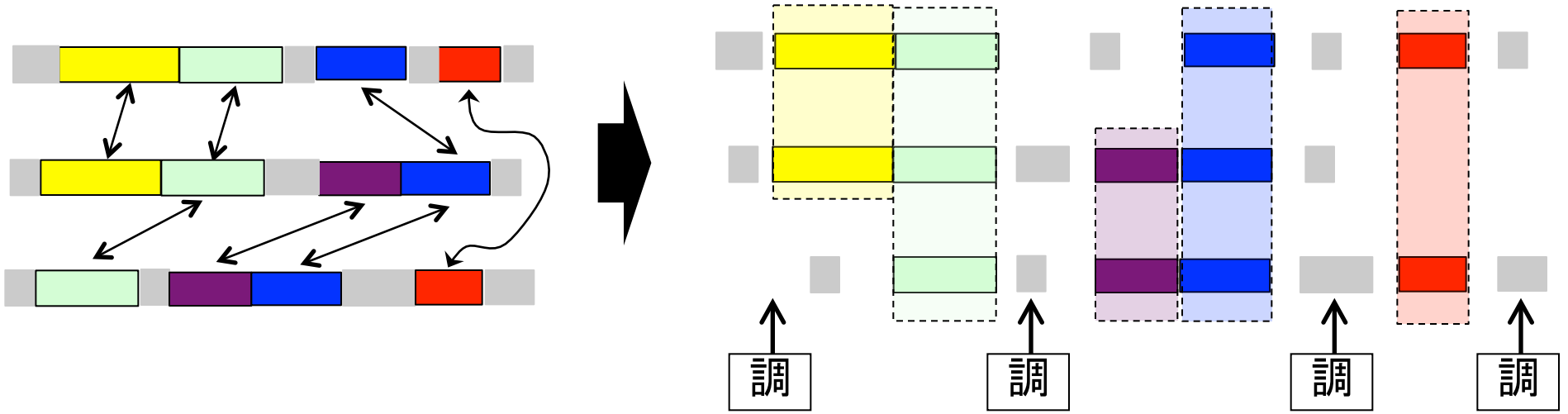
局所的相同性を利用した効率的な MSA 空間探索（1/3）

（問題点２；要因１）従来の（大局的）progressive alignment や iterative refinement では局所的最適解にtrapされる危険がある。

（問題点A）（現実的な挿入/欠失長分布を用いた）尤度計算は時間がかかる。

（案）まず、配列間の**局所的 PWA** を行い、

明らかに相同な領域のペア（ほぼgapなし）はそのまま align したままにしておいて、**残った領域**（普通 gap の近辺）のみ改めて調べる。

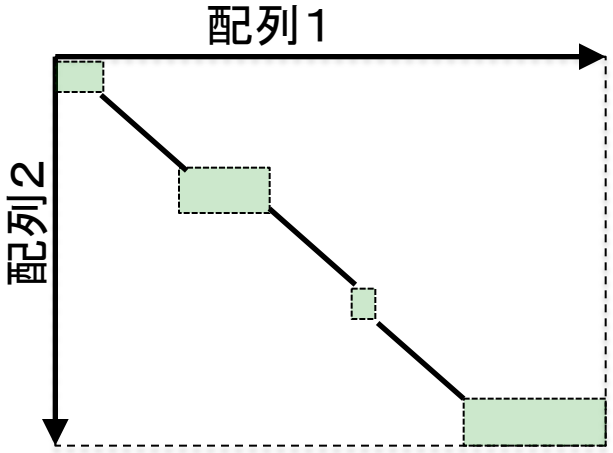


局所的相同性を利用した効率的な MSA 空間探索（2/3）

. . . この戦略により、探索空間は著しく縮小する！！

従来の（バカ正直な）
global alignmentの探索空間

効率的なMSA 空間探索
戦略の探索空間



* この戦略は、MLAGAN (Brudno et al. 2003) やMISHIMA (Kryukov and Saitou 2010) 等で導入された **divide-and-conquer 戦略**の発展版と考えることができる。

* また、この戦略は、TBA（(B)LASTZ の結果から直接（断片的）MSAを構築）(Blanchette et al. 2003)等で使われる algorithm の洗練版と見なすこともできる。

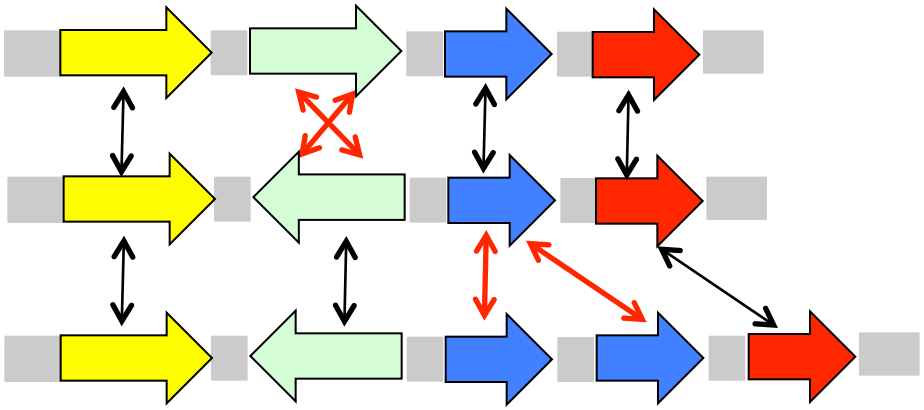
局所的相同性を利用した効率的な MSA 空間探索（3/3）

（余談）この戦略は現存の（global) MSA プログラムでは扱うのが難しい**ゲノム再編成（genome rearrangements）**

(*e.g.*, indels, duplications, inversions, translocations, transpositions) の頻発する、**不安定ゲノム領域の alignment** にも応用できるかもしれない。

逆位（inversion）

重複
（duplication）

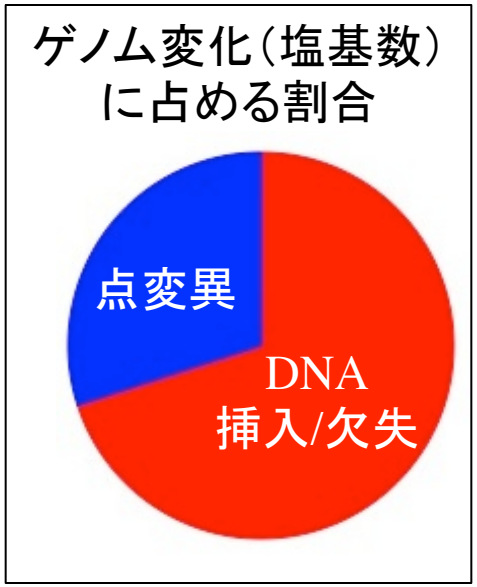


問題は、、、どうに結果を表現（表示）するか、、、

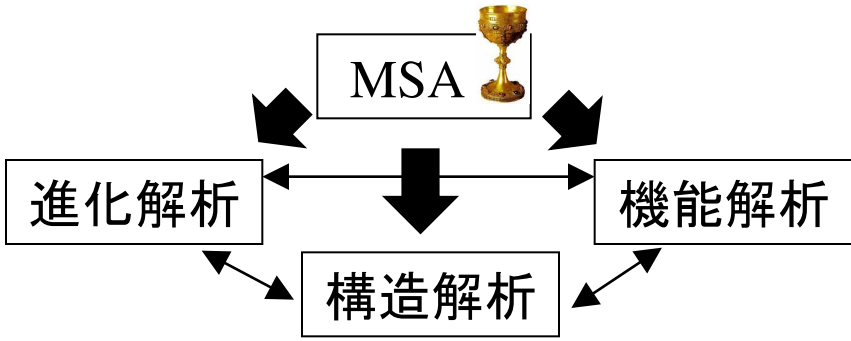
まとめ（背景）

1. DNA挿入/欠失はゲノム変化の約70%を占める。

. . . しかし、これまでは点変異の解析が中心だった。



2. MSAは配列解析の Holy Grail。



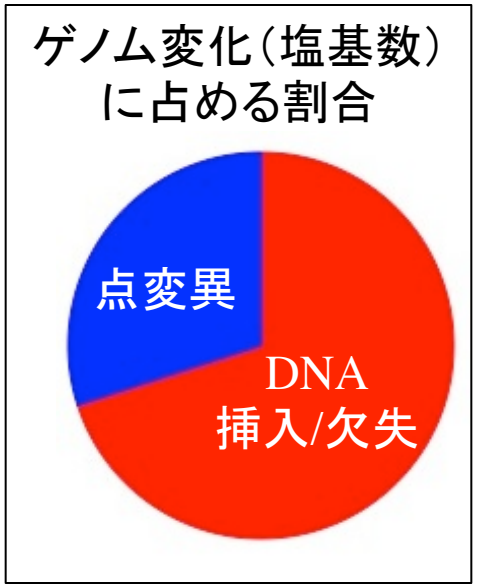
. . . しかし、正しいMSAの復元は容易ではなかった。

まとめ（概要）

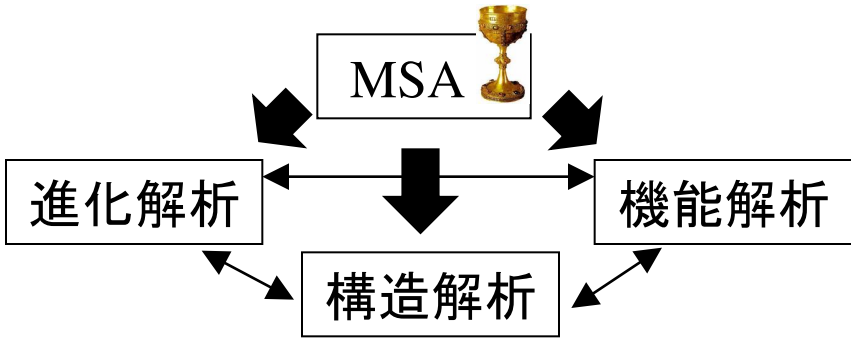
1. DNA挿入/欠失はゲノム変化の約70%を占める。

. . . しかし、これまでは点変異の解析が中心だった。

=> 我々は、**DNA 挿入/欠失の尤度理論**で
「無視されてきた70%」の研究への突破口を開く！



2. MSAは配列解析の Holy Grail。



. . . しかし、正しいMSAの復元は容易ではなかった。

=> 我々は、上記尤度理論と**効率的MSA空間探索戦略**で
MSAの「正しい復元」（truthful reconstruction）に迫る！

謝辞（Acknowledgements）

入門（initiation）:
斎藤　成也　教授　（遺伝研）
Dr. Kryukov, Kirill（遺伝研 -> 東海大学）

祖先プロジェクト&共同研究（ancestral project & collaboration）:
Prof. Graur, Dan (University of Houston)
Dr. Landan, Giddy (UH -> Heinrich-Heine University)

創造的刺激（inspiration）:
Dr. Cartwright, Reed A. (UH -> Arizona State University)

支援＆共同研究（見込み）（support & (prospective) collaboration）:
矢田　哲志　教授　　（九州工業大学）

恩師 &その他の恩人 (mentors & saviors):
吉川圭二博士（故：元大阪大学教授）、二宮正雄博士（元基研教授）、
垣谷俊昭博士（元名古屋大学教授）、五條堀孝博士（元遺伝研教授）、
池尾一穂博士（遺伝研准教授）、印南秀樹博士（総研大葉山准教授）

参考文献 (1/4)

- S.A. Benner, M.A. Cohen, G.H. Gonnet (1993), "Empirical and structural model for insertions and deletions in the divergent evolution of proteins." *J. Mol. Biol.* 229:1065-1082.
- M. Blanchette, W.J. Kent, C. Riemer, L. Elnitski, A.F.A. Smit, K.M. Roskin, R. Baertsch, K. Rosenbloom, H. Clawson, E.D. Green, D. Haussler, W. Miller (2003), "Aligning multiple genomic sequences with the threaded blockset aligner." *Genome Res.* 14:708-715.
- R.J. Britten (2002), "Divergence between samples of chimpanzee and human DNA sequences is 5%, counting indels." *PNAS* 99:13633-13635.
- R.J. Britten et al. (2003), "Majority of divergence between closely related DNA samples is due to indels." *PNAS* 100:4661-4665.
- M. Brudno, C.B. Do, G.M. Cooper, M.F. Kim, E. Davydov, NISC Comparative Sequencing Program, E.D. Green, A. Sidow, S. Batzoglou (2003), "LAGAN and multi-LAGAN: Efficient tools for large-scale multiple alignment of genomic DNA." Genome Res. 13:721-731.
- R.A. Cartwright (2006), "Logarithmic gap costs decrease alignment accuracy." *BMC Bioinformatics* 7:527.
- R.A. Cartwright (2009), "Problems and solutions for estimating indel rates and length distributions." *Mol. Biol. Evol.* 26:473-480.

参考文献 (2/4)

- M.S.S. Chang, S.A. Benner (2004), “Empirical analysis of protein insertions and deletions determining parameters for the correct placement of gaps in protein sequence alignments.” *J. Mol. Biol.* 341:617-631.
- Y. Fan, W. Wang, G. Ma, L. Liang, Q. Shi, S. Tao (2007), “Patterns of insertions and deletion in mammalian genomes.” *Curr. Genomics* 8:370-378.
- G.H. Gonnet, M.A. Cohen, S.A. Benner (1992), “Exhaustive matching of the entire protein sequence database.” *Science* 256:1443-1445.
- O. Gotoh (1982), “An improved algorithm for matching biological sequences.” *J. Mol. Biol.* 162:705-708.
- X. Gu, W.-H. Li (1995), “The size distribution of insertions and deletions in human and rodent pseudogenes suggests the logarithmic gap penalty for sequence alignment.” *J. Mol. Evol.* 40:464-473.
- D. Gusfield (1997), “*Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology*.” (Cambridge Univ. Press, Cambridge, NY).
- K. Katoh and D.M. Standley (2013), “MAFTT multiple sequence alignment software version 7: improvements in performance and usability.” *Mol. Biol. Evol.* 30:772-780.
- M. Kellis et al. (2003), “Sequencing and comparison of yeast species to identify genes and regulatory elements.” *Nature* 2003:241-254.

参考文献 (3/4)

- W.J. Kent et al. (2003), “Evolution’s cauldron: Duplication, deletion, and rearrangement in the mouse and human genomes.” *PNAS* 100:11484-11489.
- K. Kryukov, N. Saitou (2010), “MISHIMA-a new method for high speed multiple alignment of nucleotide sequences of bacterial genome scale data.” *BMC Bioinformatics* 11:142.
- G. Landan and D. Graur (2009), “Characterization of pairwise and multiple sequence alignment errors.” *Gene* 441:141-147.
- A. Löytynoja and N. Goldnam (2008), “Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis.” *Science* 320:1632-1635.
- G. Lunter, A. Rocco, N. Mimouni, A. Heger, A. Caldeira, J. Hein (2008), “Uncertainty in homology inferences: Assessing and improving genomic sequence alignment.” *Genome Res*. 18:298-309.
- S.B. Needleman and C.D. Wunsch (1970), “A general method applicable to the search for similarities in the amino acid sequence of two proteins.” *J. Mol. Biol.* 48:443-453.

参考文献 (4/4)

- The Chimpanzee Sequencing and Analysis Consortium (2005), “Initial sequence of the chimpanzee genome and comparison with the human genome.” *Nature* 437:69-87.
- The International Chimpanzee Chromosome 22 Consortium (2004), “DNA sequence and comparative analysis of chimpanzee chromosome 22.” *Nature* 429:382-388.
- Y. Yamane, K. Yano, T. Kawahara (2006), “Pattern and rate of indel evolution inferred from whole chloroplast intergenic regions in sugarcane, maize, and rice.” *DNA Res*. 13:197-204.
- Z. Zhang, M. Gerstein (2003), “Patterns of nucleotide substitution, insertion, and deletion in the human genome inferred from pseudogenes.” *Nucl. Acids Res*. 31:5338-5348.