A cancelled oral presentation (4AT27-01) at ConBio2017

自然な挿入／欠失を含む配列進化モデルを扱える理論形式とその配列アラインメントへの応用

◦ 江澤　潔　(無所属) &
矢田　哲士 (九工大、情報工•生命情報)

Theoretical formulation to handle sequence evolutionary models with natural insertions/deletions and its application to sequence alignment

◦ Kiyoshi Ezawa (Independent) &
Tetsushi Yada (Kyushu Inst. of Technol., Dept. Biosci. & Bioinform.)

無所属の江澤です。

この様な発表の機会を下さった大会関係者の方々に心より感謝致します。

この発表では、私が九工大にいた時から続けている矢田哲士教授との共同研究についてお話致します。

Table of Contents（目次）

For details on this study, please refer to（詳細は以下参照）:

Ezawa K. 2016. BMC Bioinformatics 17: 304 (theoretical basis);

Ezawa K. 2016. BMC Bioinformatics 17: 397 (computations & simulations);

Ezawa K. 2016. BMC Bioinformatics 17: 133 (application to multiple sequence alignments).

1

この研究に関しましては既に3本の論文が BMC Bioinformatics に掲載されておりますので、

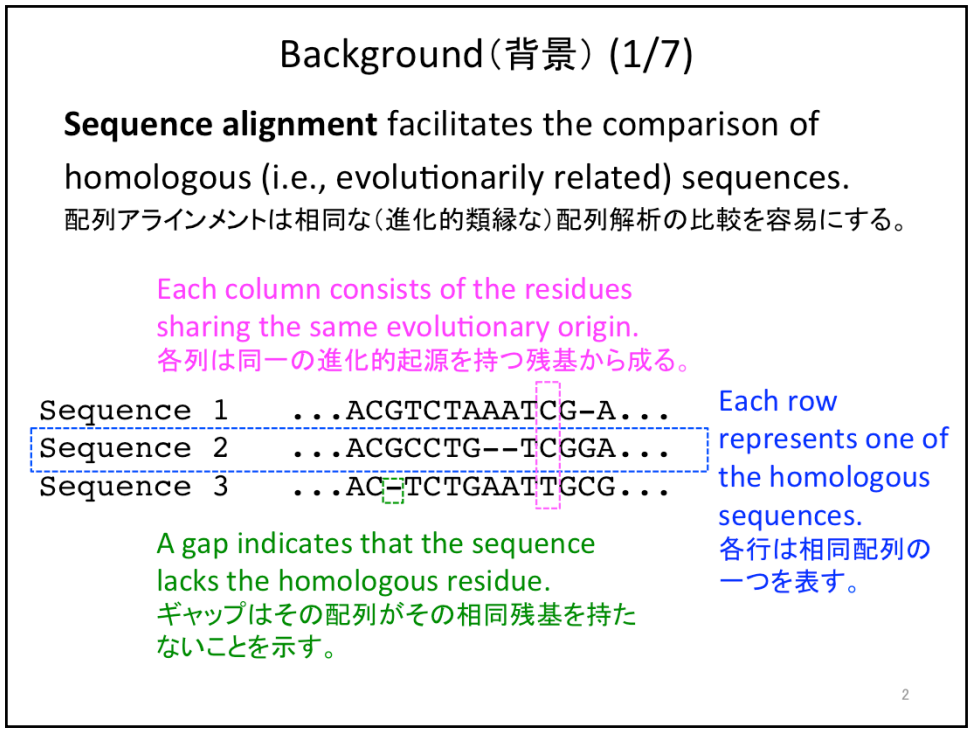
詳細をお知りになりたい方は是非これらの論文をご覧下さい。

ここでは、「何故このような研究が大事なのか？」を分かって頂く為に、

まず、研究の背景についての説明にかなり時間を割きたいと思います。

その後で、結果の中でも柱となるものについて手短に説明し、

それから、その後の研究の進捗状況を非常に簡潔に報告させて頂きます。



それでは研究の背景の説明に入ります。

まず、この研究の主題である配列アライメントについてですが、

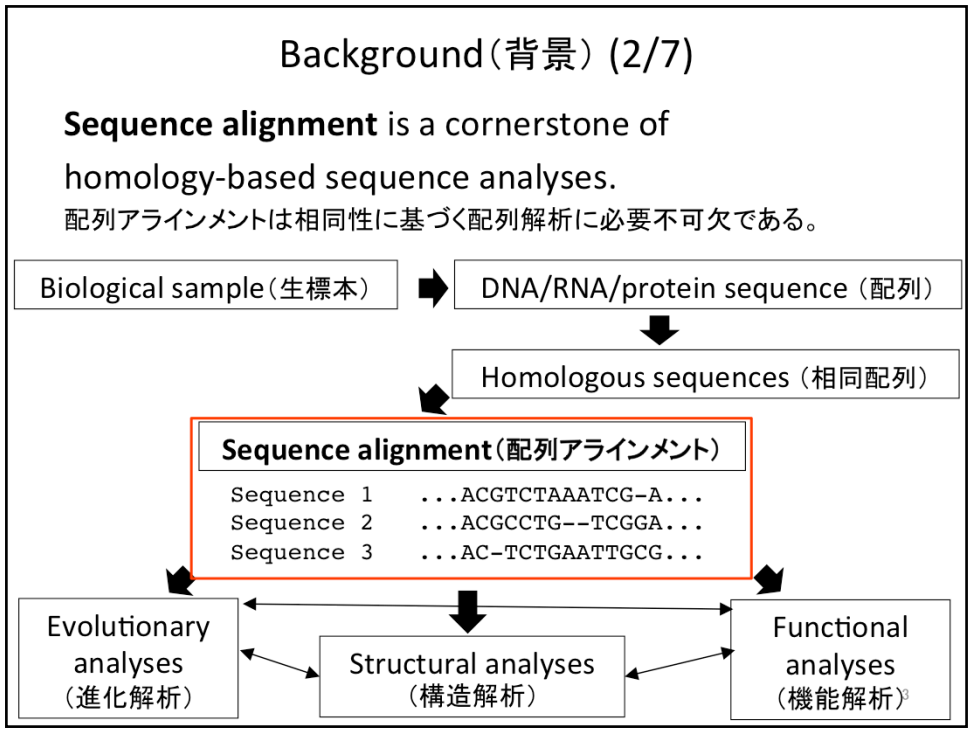
これは進化的類縁関係のある、いわゆる「相同」配列の比較を容易にする二次的データです。

よく用いられるこの様な「行列表示」では、

各行が相同配列のうちの一つを表し、

各列は同一の進化的起源を持つ残基から成っています。

そして、ハイフンで表されるギャップは、その配列がその相同残基を持たないことを示します。



実際の相同性に基づく配列解析は、

生標本から抽出したDNA/RNA/タンパクシツの配列を決定（sequencing & assembly）してから、

その相同配列を検索して、

その相同配列のセットから復元された配列アラインメントを用いて、

さらに下流の、進化解析、構造解析、機能解析等を行う、

と言う様な流れで進められます。

その中でも配列アラインメントは要に位置しますので、

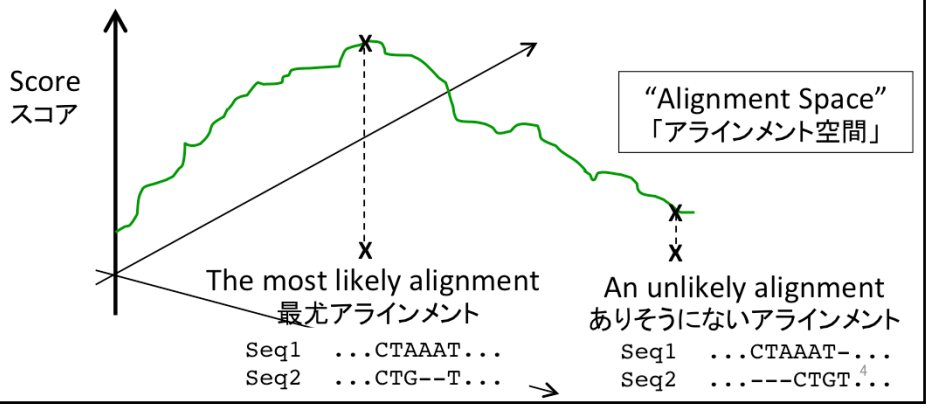
その善し悪しが配列解析全体の善し悪しを左右すると言っても過言ではありません。

Background（背景）(3/7)

In general,

an objective score function is intended to identify the most likely sequence alignment under an (explicit or implicit) model .

通常、スコア関数はある（明示的もしくは暗黙の）モデルの下で最尤の配列アラインメントを同定する様に意図されている。



配列アラインメントの復元は普通、ある客観的なスコア関数を最適化することによって為されますが、

通常、スコア関数は、ある明示的もしくは暗黙に与えられたアラインメントのモデルの下で

最も起こり易いアラインメントが最高のスコアを持つ様に意図して定義されます。

（と言い換えることが出来ます）

Background（背景）(4/7)

Roughly speaking,

$$\text{gap penalty} = -\log\ (\text{probability of the gap})$$

ギャップ•ペナルティ　　ギャップの確率

Especially,

affine gap penalty アフィン•ギャップ•ペナルティ		geometric distribution 幾何分布
$f(l) = a + b * l$	⟷	$Prob[l] = A * g^{l}$

(l: gap length ギャップ長)

* Affine gap penalty has enabled efficient dynamic programming algorithms for sequence alignment (e.g., Gotoh O. 1982. JMB 162:705-8; Durbin R. et al. 1998. Biological Sequence Analysis. Cambridge Univ. Press). 5

さて、スコア関数の重要な要素の一つであるギャップ＝ペナルティは、大雑把に言って、

ギャップの出現確率の対数のマイナスで与えられます。

特に、大変広く用いられているアフィン＝ギャップ＝ペナルティはギャップ長の幾何分布に対応します。

なお、アフィン＝ギャップ＝ペナルティがよく使われる一因は、

幾つかの効率の良い配列アラインメントの動的計画アルゴリズムを可能にした事にあると思います。

Background（背景）(5/7)

Unfortunately,

the affine gap penalty does *not* reflect realistic sequence evolution.

残念ながら、アフィン•ギャップ•ペナルティは現実的な配列進化を反映していない。

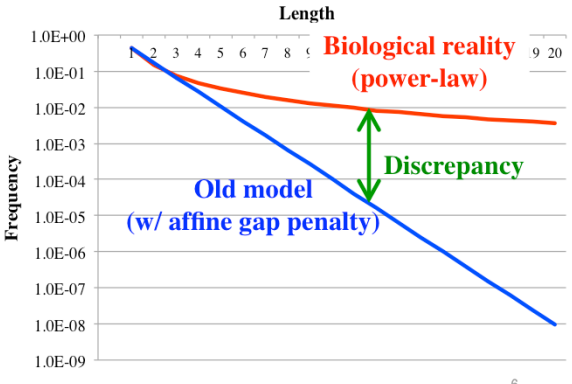
(1) Evidence from sequence data（配列データからの証拠）:

Power-law distribution of insertion/deletion lengths

挿入／欠失長のベキ乗則分布:

$$Prob[l] = C^* l^{-\gamma}$$

(e.g., Gonnet G. et al. 1992. Science 256:1443-5; Gu X. and Li WH. 1995. JME 40:464-73; Zhang Z. and Gerstein M. 2003. NAR 31:5338-48; Yamane et al. 2006. DNA Res. 13:197-204; Fan Y. et al. 2007. Curr Genomics 8:370-8)



(Adapted from: https://www.youtube.com/watch?v=L0NGsIJOuf4)

しかしながら、残念なことに、これまでの研究によって、

アフィン＝ギャップ＝ペナルティは現実的な配列進化を反映していないことが分かっています。

先ず、多数の配列データ解析により、挿入／欠失長はベキ乗分布に従う事が分かりました。

このことは、アフィン＝ギャップ＝ペナルティを使ったモデルでは、

長いギャップの出現頻度が著しく過小評価されてしまうことを意味します。

Background（背景）(6/7)

Unfortunately,

the affine gap penalty does *not* reflect realistic sequence evolution.

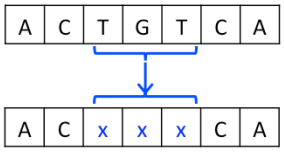
残念ながら、アフィン•ギャップ•ペナルティは現実的な配列進化を反映していない。

(2) Theoretical evidence（理論的証拠）:

the affine gap penalty requires **unnatural assumptions** on sequence evolutionary models.

アフィン•ギャップ•ペナルティは配列進化モデルに**不自然な仮定**を要する。

indivisible fragments
分割不能な小配列
(e.g., Thorne JL. et al. 1992. JME 34:3-16)



"memory" of deleted residues
欠失した残基の「記憶」
(e.g., Rivas E. and Eddy S. 2015. BMCBI 16:406)

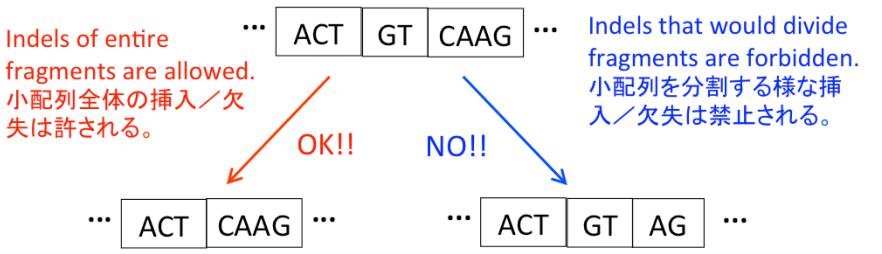
更に、
理論的な研究の結果、
配列進化モデルでアフィン＝ギャップ＝ペナルティを得るためには、
分割不能な小配列や欠失した残基の「記憶」等といった、
不自然な仮定を課さねばならない事が分かりました。

(Note) Unnatural assumptions on sequence evolutionary models
(註)　配列進化モデルへの**不自然な仮定**

(1) Indivisible fragments（分割不能な小配列）

(e.g., Thorne JL. et al. 1992. JME 34:3-16)

In a model with affine gap-penalty,
アフィン＝ギャップ＝ペナルティを持つモデルでは、



However, natural evolution allows both types of indels.
しかしながら、自然な進化では両方のタイプの挿入／欠失が許される。

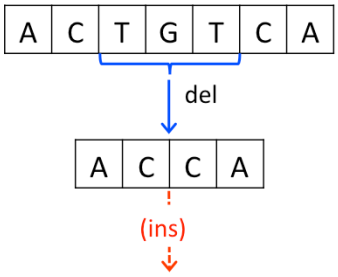
8

ここで不自然な仮定についてもう少し説明します。
アフィン＝ギャップ＝ペナルティを得るために導入された初めの仮定は
各配列を分割不能な小配列の連なりで表す事です。
この仮定では、
小配列全体の挿入／欠失は許されますが、
小配列を分割する様な挿入／欠失は禁止されます。
しかしながら、実際の自然な配列進化では
両方のタイプの挿入／欠失が許されますので、
この仮定は配列進化の忠実な記述を阻害します。

(Note) Unnatural assumptions on sequence evolutionary models
(註)　配列進化モデルへの**不自然な仮定**

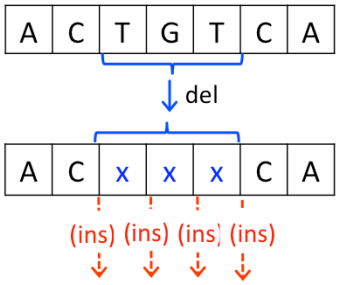
(2) "Memory" of deleted residues（欠失した残基の「記憶」）
(e.g., Rivas E. and Eddy S. 2015. BMCBI 16:406)

In natural evolution,
自然な進化では、



deleted residues leave nothing.
欠失した残基は何も残さない。

To have affine gap-penalty,
アフィン＝ギャップ＝ペナルティを得るには、



"memory" of deleted residues must be retained,
which distorts indel probabilities.
欠失した残基の「記憶」を保持しなければならず、
それが挿入／欠失の確率を歪める。

9

もう一つ必要な仮定が、欠失した残基の「記憶」です。
自然な進化では当然、欠失した残基は後に何も残しません。
しかしながら、
アフィン＝ギャップ＝ペナルティを得るためには、
欠失した残基の「記憶」を保持しなければならず、
それがその後の挿入／欠失の確率を自然なものから歪めてしまいます。

Background（背景）(7/7)

Unfortunately,

the affine gap penalty does *not* reflect realistic sequence evolution.

残念ながら、アフィン•ギャップ•ペナルティは現実的な配列進化を反映していない。

Clearly,

the current situation *contradicts* the foundation of natural science (including molecular evolution),

that is, to describe natural phenomena as faithfully as possible.

明らかに、この現状は「自然現象を出来るだけ忠実に記述する」という
（分子進化学も含めた）自然科学の根幹に反する。

=> We need to rectify the situation.
状況を是正せねばならない。

10

明らかに、この現状は「自然現象を出来るだけ忠実に記述する」という
分子進化学も含めた自然科学の根幹に反しますので、
なんとしてでも是正しなければなりません。

Goals of this study（研究目的）(1/1)

(1) to construct a **theoretical formulation** that can **handle stochastic sequence evolutionary models with natural insertion/deletion processes** (including power-law length distributions) in a mathematically exact yet intuitively comprehensible manner（自然な確率論的配列進化モデルを扱える理論形式の構築）;

(2) to develop methods (and algorithms) to **compute the** occurrence **probabilities of local sequence alignments**, based on the formulation（局所配列アラインメントの確率を計算する手法の開発）;

(3) to apply the methods to the **analyses of** multiple sequence **alignments**（多重配列アラインメントの解析への応用）.

11

そこで、この研究の目的ですが、

一つ目は、理論形式を構築して、
（べき乗則分布も含む）自然な挿入／欠失を取り入れた確率論的配列進化モデルを数学的に正確で、それでいて直感的に分かり易く扱えるようにすること、

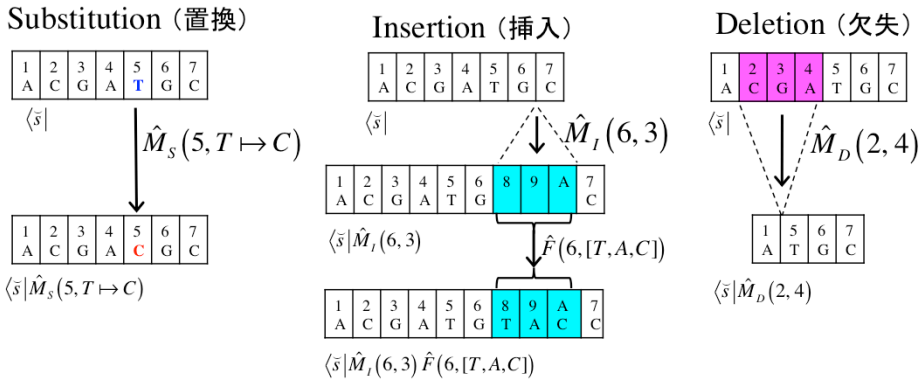
二つ目は、その理論形式に基づいて、
局所配列アラインメントの出現確率を計算する手法やアルゴリズムを開発すること、

そして三つ目は、
開発した手法を多重配列アラインメントの解析に応用することでした。

Results（結果）(1/6)

Theoretical formulation（理論形式） Ezawa K. 2016. BMCBI 17: 304

Operator representation of mutations (変異の演算子による表現):



...This greatly facilitates the mathematical handling of natural stochastic sequence evolutionary models!!

12

ここから主な結果を説明します。

まず、理論形式の構築についてですが、

我々の形式では、置換、挿入、欠失といった個々の変異をそれに対応する

「演算子」の配列状態ベクトルへの作用で表現します。

こうする事により、自然な確率論的配列進化モデルの数学的扱いが著しく容易になりました。

Results（結果）(2/6)

Theoretical formulation（理論形式） Ezawa K. 2016. BMCBI 17: 304

Factorizing alignment probability (アラインメントの確率の因数分解):

Under a set of conditions that we found,

$$Prob \begin{pmatrix} \texttt{ACGTCTAAATCG-ACTAGCATACGC} \\ \texttt{ACGCCTG--TCGGACTG--AAATGG} \\ \texttt{AC-TCTGAATTGCGCTGG---ATGC} \end{pmatrix}$$

Gapped segment ギャップあり領域

$$= Prob_0 \begin{pmatrix} \texttt{AC\#TCTA\#\#TCG\#ACTA\#-\#\#ACGC} \\ \texttt{AC\#CCTG\#\#TCG\#ACTG\#-\#\#ATGG} \\ \texttt{AC\#TCTG\#\#TTG\#GCTG\#-\#\#ATGC} \end{pmatrix} \times \mu_P \begin{pmatrix} \texttt{G} \\ \texttt{G} \\ \texttt{-} \end{pmatrix} \times \mu_P \begin{pmatrix} \texttt{AA} \\ \texttt{--} \\ \texttt{AA} \end{pmatrix} \times \mu_P \begin{pmatrix} \texttt{-} \\ \texttt{G} \\ \texttt{C} \end{pmatrix} \times \mu_P \begin{pmatrix} \texttt{GCAT} \\ \texttt{--AA} \\ \texttt{G---} \end{pmatrix}$$

... This drastically reduces the time complexity !!

13

その様な演算子表現を用いて、我々は、

ある特定の条件の下では、配列アラインメントの出現確率が、

ギャップのない部分で決まる因子と各々のギャップを含む領域からの寄与の積で表す事が出来る

事を証明しました。

このような因数分解は、アラインメントの確率の計算の時間的コストを劇的に減らします。

Results（結果）(3/6)

Computation of local alignment probabilities
（局所アラインメントの確率の計算） Ezawa K. 2016. BMCBI 17: 397

Algorithm for the 1st-approximate probability of multiple sequence alignment (多重配列アラインメントの確率の一次近似を計算するアルゴリズム):

Local alignment probability 局所アラインメントの確率	≈	Total contribution from parsimonious histories 最節約的履歴からの全寄与

ex.) $Prob\begin{pmatrix} AA \\ -- \\ AA \end{pmatrix}$

$Prob$ (ancestor AA → del → AA / -- / AA)

14

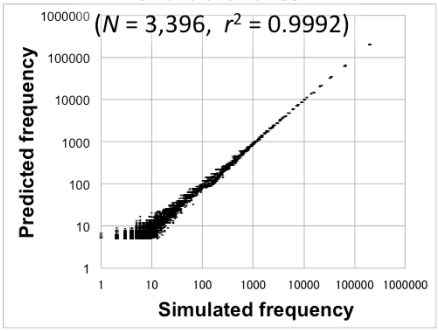
次に、局所アラインメントの確率の計算については
幾つかの手法およびアルゴリズムを開発しましたが、
特に代表的なのは、
多重配列アラインメントの確率の第一近似を計算するアルゴリズムです。
ここで、「第一近似」は、
各局所アラインメントの確率を、
それを与える様な（一般には複数ある）最節約的な挿入／欠失の履歴からの全寄与
で近似することを意味します。

Results（結果）(4/6)

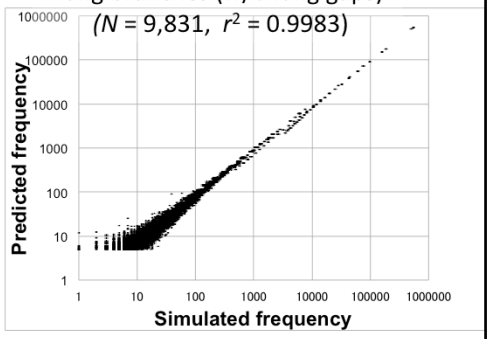
Computation of local alignment probabilities
（局所アラインメントの確率の計算） Ezawa K. 2016. BMCBI 17: 397

Comparison of theoretical predictions against simulated frequencies
(理論による予言とシミュレーションでの頻度の比較):

Short branches



Long branches (w/o long gaps)



... The 1st approximation works well when branches and/or gaps are short enough.

15

第一近似の良さを調べるため、

我々は、第一近似によって理論的に計算したギャップ＝パターンの頻度と

分子進化シミュレーションの結果生じた正しい配列アラインメント中での頻度を比較しました。

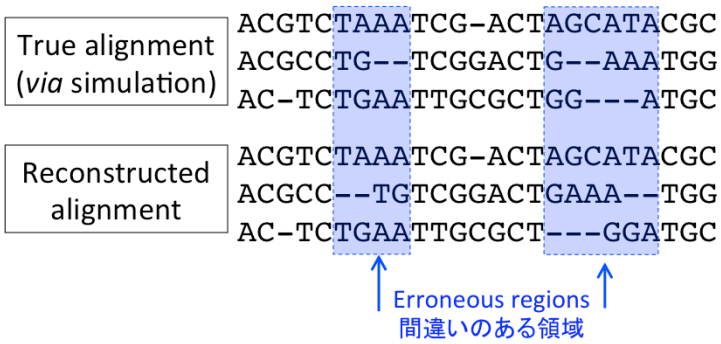
ご覧の通り、

第一近似は、系統樹の枝および／もしくはギャップが十分に短ければ、確率をかなり正確に評価することが分かりました。

Results（結果）(5/6)

Analyses of errors in multiple sequence alignments
（多重配列アラインメントの間違いの解析） Ezawa K. 2016. BMCBI 17: 133

Simulation analysis -- methods (シミュレーション解析 -- 方法):



=> Compare the probabilities of the two alignments in each erroneous region.

16

この様に、アラインメントの出現確率がかなり正確に計算できると、

多重配列アラインメントの間違いについてこれまでと違った角度から調べることが出来ます。

この解析では、

まず、配列進化シミュレーションによって正しいアラインメントを生成し、

それから、同じ相同配列を使ってあるアラインメントプログラムでアラインメントを復元します。

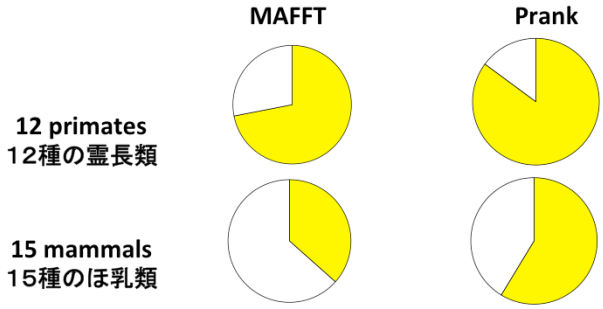
二つのアラインメントを比較すると、（青い影をつけた様な）間違って復元された領域が見つかります。

こうした各々の間違った領域で、二つのアラインメントの出現確率を比較してみました。

Results（結果）(6/6)

Analyses of errors in multiple sequence alignments
（多重配列アラインメントの間違いの解析） Ezawa K. 2016. BMCBI 17: 133

Simulation analysis -- results (シミュレーション解析 -- 結果):



... In a (near) majority of errors, the true alignment is *not* more likely than the reconstructed alignment.

17

結果をパイグラフで表します。

2つの異なる系統樹にそって進化シミュレーションして得たそれぞれ約1万のアラインメントを

2つの代表的なアラインメントプログラムを用いて復元しましたが、

いずれの場合も、

間違いのうちの過半数あるいはほぼ過半数では、

正しいアラインメントが復元されたアラインメントよりも出現確率が高くはない、

と言う結果となりました。

これは、（ほぼ）過半数の間違いでは、正しいアラインメントは最尤（あるいは最適）のアラインメントではない事を意味し、

間違いが進化過程の本質的な確率性に起因することを示します。

Summary（まとめ）(1/1)

1) Our **theoretical formulation** facilitates the handling of stochastic sequence evolutionary models with **natural insertions/deletions**（自然な配列進化モデルの扱いを容易にする理論形式）.
2) We found a set of **conditions** under which **alignment probabilities are factorable**（アラインメント確率が分解可能になる条件）.
3) We developed an **algorithm** to calculate **the 1st-approximate probabilities** of multiple sequence alignments（アラインメント確率の第一近似）.
4) The 1st-approximation works well if the branches and/or gaps are short enough（第一近似はうまく行く）.
5) Our simulation analysis indicated that a (near) majority of alignment errors are due to **the stochastic nature of evolutionary processes**, which suggests **the need for statistical alignment methods**（エラーの大半は進化過程の確率性に因る）.

18

主な結果をまとめますと、

まず、我々が構築した理論形式を用いれば、自然な挿入／欠失を含めた確率論的配列進化モデルを容易に扱う事ができます。

2つ目に、我々は、アラインメントの確率が因数分解可能になる条件を見つけました。

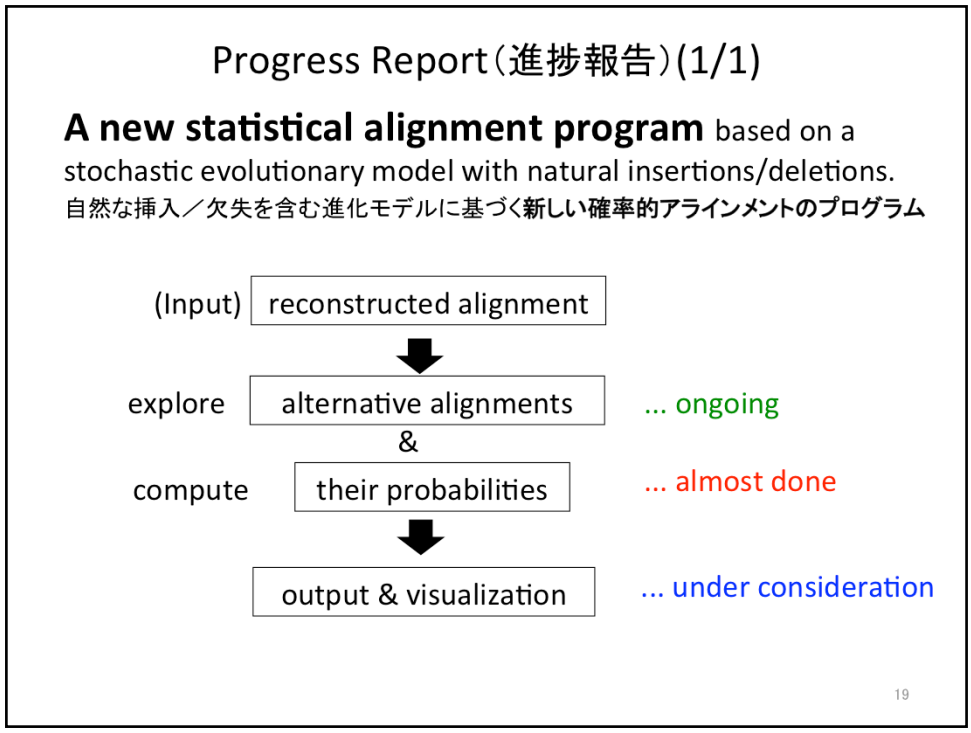
3つ目に、我々は、多重配列アラインメントの確率の第一近似を計算するアルゴリズムを開発しました。

4つ目に、枝および／あるいはギャップが十分短ければ、第一近似はかなり正確であることが分かりました。

そして、5つ目に、シミュレーション解析の結果、

過半数あるいはほぼ過半数のアラインメントの間違いは進化過程の確率性に因る事が示されました。

そしてそのことは確率的アラインメント法の必要性を示唆します。



これらの結果を踏まえて、現在我々は、

自然な挿入／欠失を含む進化モデルに基づく**新しい確率的アラインメントのプログラム**

を開発しております。

非常に大雑把に言いますと、

このプログラムは、他のプログラムで復元されたアラインメントを入力として受け取り、

その「近傍」にある替わりのアラインメントを探索し、

それらの出現確率を計算します。

そして最後に結果を出力および視覚化する予定です。

確率計算のモジュールはほぼできあがり、

替わりのアラインメントを探索するモジュールは現在開発中です。

そして、出力に関しては、どのような形式にすべきか検討しています。

近いうちにプログラムを完成させたいと思いますので、どうか楽しみに待って頂けたら幸いです。

Acknowledgements（謝辞）(1/1)

Secretariat & Organizers of ConBio2017

Prof. Dan Graur (University of Houston (UH))
Dr. Giddy Landan (UH -> Heinrich-Heine University)
Dr. Reed A Cartwright (UH -> Arizona State University)
Prof. Naruya Saitou (National Institute of Genetics (NIG))
Dr. Kirill Kryukov (NIG -> Tokai University)
Dr. Ian H Holmes (University of California, Berkeley)
Dr. Toshiaki Kakitani (former professor at Nagoya University)
Dr. Naoyuki Iwabe (Kyoto University)
Dr. Hideki Innan (Graduate School for Advanced Studies)
Prof. Takashi Gojobori (NIG)
Dr. Kazuho Ikeo (NIG)
Prof. Masao Ninomiya (former professor at Yukawa Institute for Theoretical Physics)
Dr. Keiji Kikkawa (deceased, former professor at Osaka University)

& All other people who supported and/or helped us.

20

最後に、この場をお借りして、お世話になったすべての方々に感謝の意を表します。
どうもありがとうございました。m(_ _)m
これで発表を終わりにします。
ご清聴どうもありがとうございました。m(_ _)m