2007－BIO－10 (9)
2007／9／14

エントロピー進化率を利用したマルチプルアライメント構築法

Multiple alignment algorithm with the entropy evolution rate

原 利英，佐藤 圭子，大矢 雅則
千葉県野田市山崎 2641 東京理科大学理工学部情報科学科

Toshihide Hara Keiko Sato Masanori Ohya
Tokyo University of Science Department of Information Sciences
2641 Yamazaki, Noda City, Chiba 278-8510 Japan

概要

マルチプルアライメント作成の一手法である累進法にエントロピー進化率を応用した手法を開発した．BAliBASE というマルチプルアライメントアルゴリズム間の比較を行うためのデータベースを用いた検証の結果，現在マルチプルアライメントを行うツールとして一般的に用いられている ClustalW を改善することに成功した．この結果から，累進法を利用したマルチプルアライメントアルゴリズム全般において，エントロピー進化率を利用することで精度が改善されることが期待できるといえる．

We developed the progressive method for the multipule alignment by means of the entropy evolution rate. Using the BAliBASE3.0 benchmark, the result based on our method is more accurate than that by the ClustalW which is usually used. Therefore we claim that the entropy evolution rate can improve all multiple alignment algorithms by the progressive method.

1 Introduction

近年，ヒトゲノム計画の進展に代表されるように，多種多様な生物のゲノム配列が決定され，データ量は指数的に増加し続けている．多くの生物種のゲノム配列が決定されたことにより，種間でゲノム配列を比較することができるようになった．これにより，種に共通の機能領域を特定したり，種特有の配列を見いだしたりできるようになった．こうした比較ゲノム解析において，複数本の塩基配列やアミノ酸配列を同時に比較するマルチプルアライメントは現在重要な役割を果たしている．

現在の比較ゲノム解析における手法は，マルチプルアライメントによって配列間の進化関係を明らかにしたり，データベースの類似性検索を中心としてゲノム配列から遺伝子の領域を取り出したり機能を推測したりしている．このような解析は，大規模に時間的・資金的資源を投入し行われるようにはなったが，その土台となる理論はまだまだ改善の余地があるといえる．これまでにも様々な理論が提案され，生命現象を理解するための試みがなされている．

数学的，特に確率論を用いて種と種の間に距離を定め，生物種の系統関係を探る研究が行われてきた．この研究の大きな土台となるものの 1 つに 1968 年に木村資生が提唱した分子進化の中立説 [1] があげられる．また，本研究室においても遺伝的差異を情報論的立場から計るエントロピー進化率が提案され，検証 [2][3] が行われてきた．

本論文では，このエントロピー進化率をマルチプルアライメント作成時に利用する手法について提案する．BAliBASE3.0[14] による検証の結果，現在マルチプルアライメント構築を行うツールとして一般的に用いられている ClustalW[12] を改善することに成功した．

2 Materials and methods

2.1 アライメント

アライメント（alignment）とは，複数のアミノ酸配列や塩基配列において進化的に対応する場所をそろえる作業のことを指す．また，その結果である配列グループも一般的にアライメントと呼ばれる．特に２つの配列間でのアライメントのことをペアワイズアライメント (pairwise alignment) と呼び，３本以上の場合はマルチプルアライメント（multiple alignment）と呼ぶ．

進化していく配列は，挿入（insertion）・欠損（deletion）・置換（mutation）を蓄積しているため，アライメントの際に対応する文字がなくなる場所ができる．そこで，対応する文字が無いことを空白文字である"–"や"*"を用いて表す．１つ以上の連続した空白文字のことをギャップと呼ぶ．

アライメントを行うにあたり，生物の進化，あるいは配列（アミノ酸配列，塩基配列など）間の差異（遺伝距離，あるいは単に距離）を数値として表す尺度を定義する．配列に対するアライメントとは，この尺度を用いて配列間の距離が最小となるように対応をとることと考えることができる．通常の方法では，挿入や欠損の結果であるギャップと置換の結果であるミスマッチをも含めた整列化された配列間の距離を定義する．その距離が最小となるようなギャップを挿入することをアライメントとする．

アライメントの手法として DP(Dynamic Programming 法) がある [8][9][10]．これは配列間の距離を最小にするようなアライメント結果を得ることができるが，各アミノ酸間（あるいは塩基間）やアミノ酸とギャップ間の距離の定義の仕方によって得られる結果が異なってくる．したがってその距離をどのように決定し，かつ適切なアライメント結果を導き出すかが重要となる．DP を用いたもののほかにも BLAST,FASTA といったデータベース検索を目的としたヒューリスティックな相同部位検索アルゴリズムも開発されてきた [4][5]．こういったアルゴリズムは，いくつかの仮定のもとに非常に高速に相同と思われる場所を検索することができる反面，最適解を検索できない場合もある．

2.2 累進法

累進法（progressive method）とはマルチプルアライメント構築のための一手法である．複数のアミノ酸配列に対して DP 法によりアライメントを構築することは，配列数の増加とともに計算資源的に大変困難なものとなる．そこで，累進法と呼ばれる手法では，複数のアミノ酸配列を順次アライメントして，長さのそろったアミノ酸配列のグループに束ねてゆき，最終的にアミノ酸配列全てを１つのグループに組み上げる．組み上げる順番は，案内木の情報に従う．案内木に基づく順番で配列と配列，配列とグループ，またはグループとグループのペアワイズ（グループ）アライメントを行う．基本的にペアワイズアライメントを順次行う手法であることから，計算資源の極端な消費を防ぐことができ，マルチプルアライメントを構成することが可能となる．しかし，できあがるアライメントは最適解であるとは限らないことに本質的な欠点があるといえる．つまり，この手法で重要なことは，アライメントを組み上げる際に挿入されたギャップは最後まで残る点である．そのため，結果の品質が利用する案内木に左右される．

2.3 エントロピー進化率 [2]

相同な配列（アミノ酸配列，塩基配列など）間の遺伝距離を求めるほとんどの手法は，残基の変異数をもとにした計算を行う．アミノ酸配列に対して一般的に使われるものとして，アライメントスコアや SP スコア (sum of pairs score) と呼ばれるものがあるが，これらはアミノ酸置換行列をもとに定義されたものであり，そしてこのアミノ酸置換行列は残基の変異数をカウントして作成されている．こうした手法にはいくつか問題点が指摘されている．たとえば，一本の配列の横のつながりを全く考慮せず，各サイトが独立して扱われる，挿入（insertion）・欠損（deletion）を考慮していない，などの点がある．

ここでは配列間の遺伝距離を表す他の手法として，Ohya により開発されたエントロピー進化率について説明する．

n 個の元からなる集合 $\boldsymbol{A}$ とその各元が起こる確率分布 $\boldsymbol{p}$ の組（$\boldsymbol{A},\boldsymbol{p}$）を完全事象系といい，二つの完全事象系（$\boldsymbol{A},\boldsymbol{p}$）と（$\boldsymbol{B},\boldsymbol{q}$）の事象の組が同時に起こる確率分布を r とするとき，（$\boldsymbol{A}\times\boldsymbol{B},\boldsymbol{r}$）を完全複合事象系という．この完全事象系，完全複合事象系を生物の塩基配列とアミノ酸配列において定める．

ここで，二つの配列 A,B で構成されるアライメントされた配列について考える．配列 A における構成文字 $a_1,a_2,a_3,\cdots$ と，それに付け加えたギャップ"$*$"の出現確率を $\boldsymbol{p}=(p(i))$ とし（$i=0$ がギャップ，$i=1,2,3,\cdots$ が構成文字 $a_1,a_2,a_3,\cdots$ に対応），配列 B における構成文字 $b_1,b_2,b_3,\cdots$ とギャップ"$*$"の出現確率を $\boldsymbol{q}=(q(j))$ とする．こうして，整列化された二つの配列 A,B に対する完全事象系 $(\boldsymbol{A},\boldsymbol{p}),(\boldsymbol{B},\boldsymbol{q})$ が作られるが，さらに，$\boldsymbol{A}$ の元 a_i と $\boldsymbol{B}$ の元 b_j とが対応づけられる同時確率分布 $\boldsymbol{r}=(r(i,j))$ を構成することができる．

ここで，配列のエントロピーと相互エントロピーは次のように計算できる．

$$S(A)=-\sum_i p(i)\log p(i)$$

$$I(A,B)=\sum_{i,j} r(i,j)\log\frac{r(i,j)}{p(i)q(j)}$$

なお，上記の式の和は，アミノ酸では $i,j=0,1,\cdots 20$，塩基では $i,j=0,1,2,3,4$ に対してとる．

相互エントロピーは A,B との間での情報のやりとりの精度を表すものであるため，この相互エントロピーを用いて生物間の類縁度を測ることができる．ここで，A と B との間の差異を表す量，エントロピー進化率 (Entropy evolution rate; EER) $\rho(A,B)$ を，A と B との情報量を介した相関を表す

$$r(A,B)=\frac{I(A,B)}{S(A)+S(B)-I(A,B)}$$

を用いて，

$$\rho(A,B)=1-r(A,B)\quad(0\le\rho(A,B)\le 1) \tag{1}$$

と定める．

2.4 エントロピー進化率を用いた累進法によるマルチプルアライメント構築法

ここでは，エントロピー進化率を用いた累進法によるマルチプルアライメント構築法について説明する．

累進法によりアライメントを構築するためには，案内木を作成する必要がある．案内木の作成法としては様々な手法が考えられるが，一般的には進化系統樹を作成するアルゴリズムがそのまま用いられている．エントロピー進化率を用いるためには，これが配列間の遺伝距離を計算する尺度であることから，案内木の作成法として距離法を用いることが適切であるといえる．そこで，エントロピー進化率と組み合わせる案内木作成法としてはUPGMA法 [6]，または近隣結合法 [7] を用いることとする．また，エントロピー進化率により配列間の距離を測るためには２つの配列長が等しい必要がある．そのため，配列間の距離をエントロピー進化率により計算する際に，事前にペアワイズアライメントを行う．

以上のことを整理すると，エントロピー進化率を用いた累進法の流れは以下の通りとなる．

1. 全ての配列ペアに対してペアワイズアライメント（グローバルアライメント，DP）を行い，エントロピー進化率により遺伝距離行列を作成
2. 遺伝距離行列からUPGMA法，または近隣結合法により案内木を作成
3. 案内木を元に累進法によりアライメントを作成

以上の操作により，複数の配列からマルチプルアライメントを構築することができる．

3 Test and Results

提案手法により作成したアライメントの精度を検証するにあたりBAliBASE 3.0を用いた [14]．BAliBASEとはマルチプルアライメントアルゴリズム評価用のアライメントデータベースである．登録されている各アライメントは蛋白配列から構成され，その立体構造を考慮し作成されている．バージョン3.0は217個のアライメントで構成されており，アライメントされる元の配列の種類に応じて５つのReferenceに分かれている．

表1　BAliBASEデータベースの構成内容

	No. of alignments	summary
Reference1	V1　38 V2　45	alignments of equidistant sequences and is divided into 2 subsets, according to two levels of sequence variability
Reference2	41	families aligned with one or more highly divergent "orphan" sequences
Reference3	30	divergent subfamilies
Reference4	48	sequences with large N/C-terminal extensions
Reference5	16	sequences with large internal insertions

Reference1V1: 各配列ペア間の残基一致率が20%以下であるデータセット．

Reference1V2: 各配列ペア間の残基一致率が20%-40%であるデータセット．

ここに登録されているアライメント（以下，リファレンスアライメント）と，我々の手法により作成したアライメント（以下，テストアライメント）を比較することで，アライメント精度を評価することができる．精

度を評価する指標としてはSPS，TCS[15]を用いた．

SPS(Sum of pairs score)　アミノ酸ペアがどの程度正しくアライメントできているかを表す指標である．N本の配列による配列長がMであるテストアライメントへの評価値として，

$$SPS = \frac{\sum_{i=1}^{N-1} \sum_{j=i+1}^{N} S_{ij}}{S_r}, \quad S_{ij} = \sum_{k=1}^{M} P_{ij}^{k}$$

と定義される．ここで，S_rはリファレンスアライメントにおける全アミノ酸ペアの総数であり，P_{ij}^{k}はテストアライメントのk列目における配列iと配列jのアミノ酸ペアが，リファレンスアライメントにもペアでアライメントされている場合には1，そうでない場合は0となる．

TCS(Total column score)　どれだけの列が正しくアライメントできたかを表す指標である．配列長がLであるリファレンスアライメントにおいて，C_iをリファレンスアライメントのi番目の列とテストアライメントの対応する列とが完全に一致した場合に1，一致しなかった場合に0としたとき，

$$TCS = \frac{\sum_{i=1}^{L} C_i}{L}$$

と定義される．

累進法を利用するプログラムであるClustalW[12]が現在一般的に用いられている．そこで今回，提案手法とClustalW1.83との精度比較による検証を行った．なお，ClustalW1.83は標準の設定で利用した．提案手法における各種パラメータは次の通りである．アライメントを行う際に利用するアミノ酸変異行列はGONNET250を，ギャップコストはアフィンギャップ法を用い開始コストを10, 伸張コストを0.1とした．また，案内木の作成法として近隣結合法を利用した．

提案手法とClustalW1.83での精度の検証結果を以下に示す．なお，表中の各値は中央値である．また，2つの手法間での結果に対しウィルコクソンの符号付順位和検定[16]を行い，有意水準5％で有意な差が見られるものについて中央値の高い方の数値に*記号を記している．

表2　Result

	Reference 1 Equidistant Sequences				Reference 2 Family with "Orphans"		Reference 3 Divergent Subfamilies		Reference 4 Large Extensions		Reference 5 Large Insertions	
	V1:<20%ID		V2:20-40%ID									
	SPS	TCS	***SPS***	***TCS***	***SPS***	***TCS***	SPS	TCS	SPS	TCS	SPS	TCS
proposal	0.46	0.16	***0.88**	***0.70**	***0.88**	***0.26**	0.68	0.20	0.73	0.21	0.65	0.14
ClustalW1.83	0.48	0.15	**0.84**	**0.67**	**0.85**	**0.25**	0.7	0.25	0.75	0.33	0.63	0.17

表中の値：　各データセットに対する指標SPS,TCSによる評価の中央値．

太字数値：　ClustalWと提案手法の間に有意差が見られるもの．具体的には，有意水準5%でのウィルコクソンの符号付順位和検定による判定．

*記号：　ClustalWと提案手法の間に有意差が見られるとき，数値が高い方に表示．

4 Discussion

累進法を用いる現在一般的なソフトである ClustalW を比較対象とし BAliBASE を用いた比較を行った結果，指標 SPS,TCS による評価により精度が向上することが確認できた．

具体的には，データセット Reference1V2,Reference2 において有意な改善がみられ，その他のデータセットに関しては有意な差は見られなかった．データセット Reference1V2 は，各配列ペア間の残基一致率が 20-40% のもので構成される．また，データセット Reference2 におけるファミリーは，残基一致率 40% 以上のもので構成される．このことから，残基一致率が 20% 以上である配列群に対するアライメントで精度が改善されたと見ることが出来る．これは，情報量の観点から距離を測ることで，より生物学的に正しい案内木が作成され，累進法の弱点であるギャップの初期段階での誤った挿入が防げたためだといえる．

現在，様々な手法において配列における各サイトは独立したものと仮定されているが，実際にはなにかしらのつながりがあると考えるのが自然である．特に，アミノ酸配列においては立体構造的な観点から考えてもサイトは独立したものとは考えにくい．エントロピー進化率はアミノ酸配列における各サイトを独立したものとは扱わず，サイト間のつながりを情報量としてとらえ考慮しているといえる．そのため，SP スコアに比べ情報量的に正しい距離を推定することができる．その結果，より生物学的に正しい案内木が作成されるため，こういった結果につながったといえる．

現在，マルチプルアライメント構築アルゴリズムは累進法を応用したものがほとんどである．本論文の結果から，これらの累進法を利用したツール全般において，エントロピー進化率を利用することで精度が改善されることが期待できるといえる．

5 参考文献

参考文献

[1] Kimura M., The evolutionary rate at the molecular level. Nature, vol. 217, pp.624-626, 1968.

[2] Information theoretical treatment of genes, Trans. IEICE, E72, No.5, 556-560, 1989.

[3] M.Ohya,Miyazaki,Sugawara, The efficiency of entropy evolution rate for construction of phylogenetic trees, Genes Genet. Syst., 71, 323-327, 1996.

[4] Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schaffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, Nucleic Acids Res. 25:3389-3402. 1997

[5] D. J. Lipman, W. R. Pearson: Rapid and sensitive protein similarity searches, Science, 227, 1435-1441, 1985

[6] Sokal R. and Sneath P.H.P., Principles of Numerical Taxonomy. Freeman&Co., New York, 1968.

[7] Saitou N. and Nei M., The neighbor-joining method: a new method for reconstructing phylogenetic trees. Molecular Biology and Evolution, vol. 4, pp.406-425, 1987.

[8] Needleman S.B., Wunsch C.D., A general method applicable to the search for similarities in the amino acid sequence of two proteins. J. mol. Biol., 48(3): 443-53, 1970.

[9] T.F.SMITH AND M.S.WATERMAN, Identification of Common Molecular Subsequences,Reprinted from J.Mol. Biol.147,195-197, 1981.

[10] M.Ohya and Y.Uesaka, Amino acid sequences and DP matching:a new method of alignment, Information Sciences,63,139-151, 1992.

[11] 大矢雅則, "情報進化論", 岩波書店, 2005

[12] Thompson, J. D., D. G. Higgins, and T. J. Gibson:CLUSTALW, Nucleic Acids Res. 22, 4673-4680, 1994

[13] Thompson,J.D., Plewniak,F. and Poch,O. BAliBASE: a benchmark alignment database for the evaluation of multiple alignment programs, Bioinformatics, 15, 87-88, 1999

[14] Thompson J.D., Koehl P., Ripp R., Poch O. BAliBASE 3.0: latest developments of the multiple sequence alignment benchmark. Proteins. 2005;61:127–136.

[15] Thompson,J.D., Plewniak,F. and Poch,O. A Comprehensive comparison of multiple sequence alignment programs, Nucleic Acids Research, 27, 2682-2690, 1999

[16] David F. Bauer, Constructing confidence sets using rank statistics, Journal of the American Statistical Association 67, 687–690, 1972