社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2008－MPS－72 (33)
2008－BIO－15 (33)
2008／12／18

エントロピー進化率を用いたマルチプルアライメントの精度改善
More accurate multiple alignment method 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構築時に用いる目的関数として，情報量を下に定義された遺伝距離であるエントロピー進化率を利用する手法を開発した．アライメントの精度を検証するためのデータベースを用い提案手法の検証を行った結果，現在最高水準の精度を有する T-Coffee にくらべ有意な結果が得られた．この結果から，配列間の遺伝距離を用いる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 multiple 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 T-Coffee. Therefore we claim that the entropy evolution rate is useful as the genetic distance for the multiple alignment.

1 Introduction

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][4] 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．具体的には現在最高水準の精度を有すると言われているマルチプルアライメント作成プログラムである T-Coffee[12] を下にし，遺伝距離 (genetic distance) として配列一致率 (Sequence Identity) の 替わりにエントロピー進化率を利用することを考える．BAliBASE3.0[9] による検証の結果，精度の改善を確認することができた．

2 Materials and methods

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の対応をそろえる作業のことをアライメント (alignment) と呼ぶ．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ぶことにする．２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本以上の場合をマルチプルアライメント (multiple alignment) と呼ぶ．配列は進化とともに挿入 (insertion)・欠損 (deletion)・置換 (mutation) といった変化を蓄積してきたため，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．アライメントを行う手法として DP(Dynamic Programming) を用いた手法が一般的に用いられる [5][6][7][8]．この手法では配列間の距離を最小にするアライメント結果を得ることができるが，アライメントの際の目的関数（配列間距離）の定義の定義の仕方によって得られる結果は異なったものとなる．したがって目的関数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．

ここで，マルチプルアライメント時の目的関数として配列一致率を用いた手法である T-Coffee[12] について説明する．T-Coffee とは累進法を利用したマルチプルアライメント構築のための手法及びプログラムの名称である．現在，累進法に分類されるアルゴリズムの中では最高水準の精度を有すると言われている．T-Coffee に特徴的なことは，与えられた入力配列に対しすべての配列ペアにおけるペアワイズアライメントを求めた後，それを下に各文字ペアの重み (Weight) を計算することである．この重みを用い累進法によりマルチプルアライメントを構築する．彼らの手法では文字ペアの重みを，それぞれの文字の属する配列間の一致率 (Sequence Identity) により定義している．二本の配列 A, B におけるペアワイズアライメントにおいて，ギャップも含めた全体の長さを t，異なっている箇所の個数を a，挿入・欠損が起こっているところ (ギャップ) の個数を d とするとき，配列一致率 $\sigma(A,B)$ は以下のように定義される．

$$\sigma(A,B) = \frac{t-d-a}{t-d}$$

配列一致率に代表されるように，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．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．配列間の遺伝距離を情報量を下に定義したものとして，Ohya により開発されたエントロピー進化率 [2] があげられる．そこで我々は T-Coffee によるアルゴリズムを下にし，エントロピー進化率 [2] を用いる手法を開発した．この手法の大きな流れとしては文字ペア間の重みを計算しそれを用いて累進法を行う流れとなる．その際に文字ペア間の重みとしてエントロピー進化率を利用する．手順は以下の通りである．

1. 全ての配列ペアに対してペアワイズアライメント (DP 法) を行い，各ペアにおけるエントロピー進化率を求める．
2. 全ての文字ペアを重みとセットにし記憶する．重みとしては，それぞれの文字ペアの属するペアワイズアライメントにおけるエントロピー進化率の値を 1 から引いたものを用いる．
3. 各文字ペアの重みを用い累進法によりマルチプルアライメントを構築する．記憶に存在しない文字ペアの重みは 0 と考える．

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．エントロピー進化率とは以下のようなものである．

2.1 エントロピー進化率 (Entropy Evolution Rate; EER)

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情報量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．相互情報量は 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，この相互情報量を用いて生物間の類縁度を測ることができる．ここで，A,B 間のエントロピー進化率 $\rho(A,B)$ を，

$$\rho(A,B)=1-\frac{I(A,B)}{S(A)+S(B)-I(A,B)} \quad (0\leq\rho(A,B)\leq 1)$$

と定める．

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 [9]．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 つのデータセット (Reference) に分かれている．

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[10] 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，

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

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．

提案手法 (EER を用いる) と T-Coffee(Sequence Identity を用いる) での精度の検証結果を表 1 に示す．なお，表中の各値は中央値 (median) 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 [11] 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 1　Result

	Reference 1 Equidistant Sequences V1:<20%ID	Reference 1 Equidistant Sequences V2:20-40%ID	Reference 2 Family with "Orphans"	Reference 3 Divergent Subfamilies	Reference 4 Large Extensions	Reference 5 Large Insertions
Entropy Evolution Rate	***0.476**	***0.878**	**0.864**	**0.735**	0.818	**0.725**
Sequence Identity	0.471	0.876	0.863	0.733	0.818	0.717

表中の値：　各データセットに対する指標 SPS による評価の中央値 (median).

太字数値：　各データセットにおける数値の高い方の手法を太字表示.

*記号：　提案手法 (Entropy Evolution Rate) と T-Coffee(Sequence Identity) の間に有意差が見られるとき，数値の高い方に表示．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.

Ref.1: alignments of equidistant sequences and is divided into 2 subsets, according to two levels of sequence variability.

Ref.2: families aligned with one or more highly divergent "orphan" sequences.

Ref.3: divergent subfamilies.

Ref.4: sequences with large N/C-terminal extensions.

Ref.5: sequences with large internal insertions.

4 Discussion

現在最高水準の精度を持つとされるプログラムである T-Coffee を比較対象とし BAliBASE3.0 を用いた比較を行った結果，指標 SPS 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 Reference1V1,V2 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若干改善しているように見える結果が得られた．データセット Reference1 は，各配列ペア間の残基一致率が 40% 以下のもので構成されており，このことから一見対応が分からないような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．

現在，様々な手法において配列における各サイト (配列における文字あるいはアライメントにおける列) 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各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，単純な文字の一致率である配列一致率 (Sequence Identity) に比べ情報量的に正しく距離を推定することができる．その結果，より生物学的な配列進化を正しく把握することができたため，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法は配列間の遺伝距離を用いる手法がほとんどである．本論文の結果から，これらの遺伝距離を利用した手法及びプログラム全般において，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．

参考文献

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

[2] M.Ohya, 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] T.Hara, K.Sato, M.Ohya, Multiple alignment algorithm with the entropy evolution rate, IPSJ SIG Technical Report, 2007-BIO-10, pp.65-71 (2007)

[5] 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.

[6] T.F.Smith, M.S.Waterman, Identification of Common Molecular Subsequences,Reprinted from J.Mol. Biol.147,195-197, 1981.

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

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

[9] 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.

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

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

[12] C. Notredame, D. Higgins, J. Heringa, T-Coffee: A Novel Method ofr Fast and Accurate Multiple Sequence Alignment, J. Mol. Biol., 302, 205-217, 2000