D-009

多重整列に基づくモチーフの統計的抽出法

Statistical method for extracting motifs based on multiple alignments

福本 翔平† 北上 始† 森 康真†

Shouhei Fukumoto Hajime Kitakami Yasuma Mori

1. はじめに

配列データベースから類似部分の多い部分配列，すなわち，モチーフを抽出する方法は，アミノ酸などの分子配列データの抽出法としても数多く提案されている．アミノ酸は 20 種類存在し，それぞれアルファベット 1 文字を対応させて表現している．代表的なモチーフ抽出法の 1 つとして，ギブスサンプリング(GS) 法[1]が知られている．しかしながら，その方法は，初期解（ある部分配列集合）からスタートし，部分配列データ集合のプロファイル（位置依存スコア行列）に最も近い類似部分配列の位置を配列データごとに確率探索している．このため，全域的な最適解に収束することが保障されておらず，局所的な最適解に落ちてしまうという危険性がある．

本稿では，配列データベースを予め多重整列化した上で，高い出現頻度を持つ類似部分配列集合を抽出する手法を提案する．この類似部分配列集合を抽出するために，多重整列の左から右に向かって，ある長さを持つスライディングウィンドウを移動させ，そのウィンドウの相対エントロピーが最大になる領域を見つけ出している．また，提案手法の有効性を検証するために，GS 法との比較実験を行う。

2. プロファイル

N 本の配列からなる配列データベース *DB* があるとしよう。*DB* の各配列から長さ *K* の部分配列をとりだし，それらを *N*×*K* 行列とみなしたものを整列行列と呼ぶ．整列行列の統計的な特徴を表現したものはプロファイルと呼ばれている。代表的なプロファイルには、出現頻度行列や位置依存スコア行列などが知られている。出現頻度行列 *FMAT* は，整列行列の列ごとに出現する文字 α の頻度を表現する *M*×*K* の行列 *FMAT* = $(p_{i,j})$ であり， 位置依存スコア行列 *PSSM* は，*FMAT* の各要素をオッズ比に置き換えた *M*×*K* の行列 *PSSM* = $(r_{i,j})$ である．q_i を *i* 行目に対応する文字∝が整数行列を除く配列領域に出現する確率（背景頻度）とすると，$r_{i,j}=p_{i,j}\div q_i$ が成立する．

3. ギブスサンプリング法

従来手法として知られているギブスサンプリング法（GS 法）の主な目的は図 1 のように，*M* 種類のアルファベットと配列数 *N* 本によって定義される配列データベース *DB* から，ユーザが定めた長さ *K* の部分配列集合（以下では *K*-部分配列と表記）の集合）を取り出し，できるだけ類似した部分配列集合となるように計算していくものである．*K*-部分配列集合とは，*N* 本の配列からなる配列データベースの各配列から取り出される長さ *K* の部分文字列集合の事である．

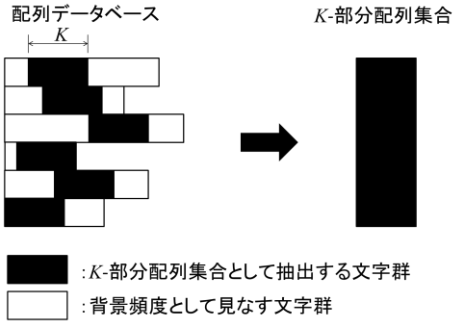

図 1 配列データベースと *K*-部分配列集合

3.1 アルゴリズム

GS 法は *N* 件の配列からなる *DB* からランダムに選択された配列 *Z* を用いる事で，プロファイル（位置依存スコア行列）を算出し，出現頻度が高くかつ背景頻度の低い *K*-部分文字列集合を抽出する処理を行っている．そのアルゴリズムは以下のとおりである．

① *DB* から *K*-部分配列集合を取り出し，同様に DB からランダムに一つの配列 *Z* を選択する．

② 配列 *Z* を除く *K*-部分配列集合－{*Z*}から，*K* 個の列ごとに *M* 個の各文字に対する *M*×*K* の出現頻度行列 *FMAT*= $(p_{i,j})$ および *M*×*K* の位置依存スコア行列 *PSSM* = $(r_{i,j})$ 算出する．

③ 配列 *Z* 内に存在する $(|Z|-K+1)$ 個の *K*-部分配列からそれぞれの類似度スコア A_x を計算する．類似度スコア A_x は，*M*×*K* の位置依存スコア行列 *PSSM* = $(r_{i,j})$ を用い，*K*-部分配列を構成する各文字 α_i のオッズ比 r_{ij} をすべて掛け合わせることにより計算する．

④ $\{A_1, A_2, \ldots, A_{|Z|-k+1}\}$ となった各値から，比例した確率で A_r を選択し，A_r に対応する *K*-部分配列を新たな Z として更新する．

⑤ 結果が収束するまで，①～④を繰り返す．繰り返し回数は多いほど良い結果が出力されるが，その分計算時間が大幅に増加する．相対エントロピーを計算し，解の最適性を評価する．

3.2 相対エントロピー

K-部分配列集合を評価する方法として，相対エントロピーと呼ばれる評価関数を用いる．先ず，ベイズ統計解析を考慮し，出現頻度行列 *FMAT*=(p_{ij}) を式(1)のように定義する．

$$p_{ij} = \frac{(c_{ij} + b_i)}{((N-1)+B)} \qquad (1)$$

c_{ij} とは，*FMAT* の *i* 行目に該当する文字が *j* 列目に現れる数である．*N* は配列総数，*B* は $\sqrt{N}$ と定める．また，*FMAT* の *i* 行目に該当する文字の疑似度数 b_i は $f_i \times B$ としており，式 (1) によって算出される出現頻度行列 *FMAT*=(p_{ij}) の相対

†広島市立大学

エントロピーFは以下の式(2)となる．

$$F = \sum_{i=1}^{K}\sum_{j=1}^{M} C_{ij} log\left(\frac{p_{ij}}{q_i}\right) \qquad (2)$$

ただし，アミノ酸配列を扱う場合，Mは 20 である．この式を K-部分配列集合に当てはめる事によって，得られた値が 0 よりもプラス側に遠ざかれば類似部分配列として近似し，0 に近ければ類似していないものと判断できる．

4. 提案手法

提案手法の主な目的は，*DB* を多重整列化（マルチプルアラインメント）した *DB'*から新しいプロファイル計算法を用いて，ユーザが定めた K 値分の K-部分配列集合を取り出し，相対エントロピーを求めるものである．多重整列化とは，ギャップと呼ばれる記号（-）を配列データの各文字を類似した部分で特定できるように挿入し，配列データベースの長さを統一するものである．

4.1 多重整列化の方法

多重整列化を行うプログラムとして，本稿で ClustalX[2]と呼ばれている系統解析用のプログラムを使用した．扱うデータに関しては PROSITE[3]から抽出した特定のアミノ酸データを用いる．動作を行うためには先ず，ClustalX に読み込ませるデータを FASTA 形式に変換する必要がある．

FASTA 形式とは，塩基配列やアミノ酸配列をアラインメントする為に用いられる表現方法であり，1 行目はシーケンスデータの詳細，2 行目以降は実際のデータの文字列で構成されている．これによって多重整列化されたデータを文字列集合 *DB'*として扱い，提案手法に導入することで類似部分を抽出していく．ただし，問題となる部分もあり，ClustalX による多重整列化にはノイズも少なからず関与しているので，完全に多重整列化された結果が出力される訳ではない．

4.2 新しいプロファイル計算法

提案手法の計算を行うためには，多重整列化された *DB'* を考慮した新しいプロファイル計算法を用いる．その過程を以下のように定めている．

(1)　ギャップの置き換え

多重整列化によって挿入されたギャップは，アミノ酸のアルファベットではないため，GS 法のような正確なプロファイルを算出する事ができない．そのためギャップを別の文字に置き換える必要がある．本稿では，*DB'*内に存在する M 種類のアルファベットを全てのギャップに対してランダムに置き換え，新たなプロファイルである $M \times K$ の位置依存スコア行列 *FMAT*=（$p_{i,j}$）を算出する．これにより，ギャップが多く含まれている部分にはランダム性があり，多重整列化により整列された部分には類似性がある事を表現している．

(2)　スライディングウィンドウ法

提案手法では，多重整列化とそれに伴うギャップの置き換えにより，新たな K-部分配列集合の抽出とプロファイルの計算法を行う．先ず多重整列化によって統一された *DB'* 全体の長さを L とすると，１つの K-部分配列集合の長さが K である事から，図 2 のように$L - K + 1$個の K-部分配列集合が作られる．*DB'*を矩形の $N' \times L$ 行列とみなし，各集合（$N' \times K$ 行列）を特にクラスタと呼ぶ．ただし，$N' = |DB'|$とする．最後にクラスタごとのプロファイル$(g_{i,j})$と相対エントロピーを算出する．そして相対エントロピーの値が最も高いクラスタを類似部分配列集合と見なして抽出する．　1 番目から$L - K + 1$番目のクラスタをスライドさせながら計算を進めることから，本稿ではこの処理をスライディングウィンドウ法と呼ぶ．

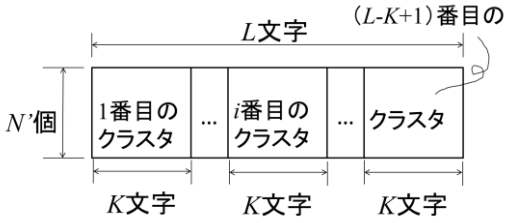

図 2　*DB'* 内に存在するクラスタ

(3)　ギャップを考慮したクラスタの削除

あるクラスタにおける文字列集合の開始位置と終了位置にギャップが存在している場合は，それらの位置のギャップ数をカウントし，両者が次式を満たす場合に限り，解の候補とし，相対エントロピーによる評価を行う．

$$\text{列に存在するギャップ数} \leq N' \times \text{閾値} \qquad (3)$$

理由の 1 つとして，モチーフを表現する正規表現において，開始位置や終了位置にギャップは存在しないことが挙げられる．よって，そのような K-部分配列集合を解の候補から除外するための処置として行っている．これにより，不要なクラスタの計算を削除することでアルゴリズム全体の計算時間の短縮も見込める．しかし，このような操作を行っても，始点と終点にギャップが含まれている文字列は完全には除去できないので，そのようなギャップ入りの K-部分配列集合を抽出する場合は，その K-部分文字列だけを削除して抽出する操作を行っている．

4.3 提案手法のアルゴリズム

提案手法ではアラインメントされた *DB'*を用いて，スライディングウィンドウ法により評価関数が最も大きいクラスタを抽出し，類似部分配列として表現する．提案手法で利用するアルゴリズムは以下のようになる．

① ClustalX にてアラインメントされた *DB'*から，長さ K の値分の部分配列集合（$N' \times K$）を 1 クラスタと見なす．

② GS 法のプロファイル計算法から 1 クラスタ分の位置依存スコア行列を算出する．
位置依存スコア行列を基に相対エントロピーによる評価を行う．

③ プロファイル$(g_{i,j})$を基に相対エントロピーの評価関数を算出する．

④ スライディングウィンドウ法により，$L-K+1$ 個分のクラスタが終了するまで，同じ処理を繰り返す．

⑤ $L-K+1$ 個分の相対エントロピーから最も値の大きい K-部分配列集合を呼び出し，K-類似部分配列集合として出力する．

4.4 相対エントロピーの新しい計算法

提案手法における相対エントロピーでは，疑似度数を考慮した位置特異スコア行列として，式（1）の疑似度数b_iを次式[4]のように定義する．

$$b_{aj} = B_j \times \sum \left(\frac{c_{ij}}{N'} \times \frac{w_{ia}}{W_a} \right) \quad [1 \le i \le 20] \qquad (4)$$

同様に式（1）の疑似度数のBは$B_j = e \times s_j$と定義し，s_jを 1 クラスタにおける j 列目の文字の種類数とする．e は試験的な検索実験により決定される正の数であり，位置特異スコア行列に用いる際，e=5~6 が最も有効であると報告されている．また，w_{ia}はアミノ酸の文字 i から文字 a への置換頻度で，BLOSUM62[5]を $s(i,a)$として利用すると，$w_{ia} = w_i \times w_a \times 2^{s(i,a)}$となる．$w_i$は DB'内の i の出現確率で，W_aはw_{ia}の文字ごとの総和であり，$W_a = \sum w_{ia}$と定められている．

5. 評価実験

本章では，閾値を 0.1 として，提案手法の評価実験を行う．使用した配列データベースは，PROSITE 内に登録されているアミノ酸データセットを 5 つ用いており，詳細を表１に示す．モチーフの長さに関しては，多重整列化によって挿入されたモチーフ内のギャップも考慮して示している．

表１：実験に使用したデータセット

番号	モチーフ名	登録番号	長さ	件数
1	Kringle	PS00021	14	95
2	Homeobox	PS00027	115	1321
3	PTS_EIIA	PS00372	22	51
4	HTH_ASNC	PS00519	37	43
5	HTH_DEOR	PS00894	35	82

従来手法の実行において，処理の繰り返し回数は，データセットごとの DB 内の文字数とした．提案手法では試行回数を 100 回とし，それぞれ抽出した結果の平均を精度とする．

また，従来手法と提案手法との性能を比較するために，以下で定義される精度の式(5)を利用する．

$$\text{精度}(\%) = \frac{B}{B + C} \times 100 \qquad (5)$$

検索で合致した範囲を B として，ノイズの部分を C とする．数値が高い程一致している部分が多いと見なす．

更に，提案手法の初期に実施される多重整列化の精度を確認する．その計算は以下の式(6)を利用する．この精度が高い程，多重整列上のモチーフが同じ位置にある事を示す．

$$\text{精度}(\%) = \frac{1}{{}_{N'}C_2} \sum \sum ST_{ij} \; [1 \le i < j \le N] \qquad (6)$$

ST_iを i 番目の配列のモチーフ位置，N を配列の総本数とする．ただし，$ST_{ij} = 1 - \{f(ST_i, ST_j) \div K\}$とし，$K$ をモチーフ長，$f(ST_i, ST_j)$については以下の式(7)のように定義する．

$$f(ST_i, ST_j) = \begin{cases} |ST_i - ST_j| & for\ |ST_i - ST_j| < K \\ K & for\ |ST_i - ST_j| \ge K \end{cases} \qquad (7)$$

表 2 に従来手法と提案手法の精度を示す。また，この表に， 5 つのデータセットのそれぞれに対する多重整列化の精度も示す． なお，この精度については，式（6）（7）を利用して計算している．

表２：提案手法と従来手法との精度結果比較

番号	提案手法(%)	従来手法(%)	多重整列化の精度
1	77.22	64.44	70.10
2	91.37	87.75	94.73
3	75.58	49.18	72.25
4	59.27	41.76	93.49
5	89.79	17.06	98.68

表 2 を見る限りでは，提案方式が従来手法よりも優れていることが分かる．しかし，多重整列化の精度と結果を見比べてみると，4 番においては，多重整列の精度が良いにも関わらず抽出されたモチーフの精度が低い．その原因は，多重整列化によってモチーフが存在するクラスタよりも更に類似している別のクラスタが現れ，そちらを抽出してしまうことにある．

以上により，モチーフ抽出の精度がデータセットごとに悪化する原因には，この多重整列化が一部関係しているのではないかと考えられる．

6. まとめ

本研究では，疑似度数を考慮した評価関数を新たに挿入してモチーフを抽出する方法を提案した．評価実験による考察で，多重整列化の精度を考慮した新たな改良（逐次改善法の導入など）を加える必要がある事が分かった．

また，提案手法によって抽出された K-類似部分配列集合の位置を GS 法の初期値とすれば，ランダムに初期値を設定するよりは，安定した精度で K-類似部分配列集合を抽出できるのではないかと思われる．

参考文献

[1] Charles E. Lawrence, Stephen F. Altscul, Mark S. Boguski, Jun S. Liu, Andrew F. Neuwald, and John C. Wootton: Detecting Subtle Sequence Signals: A Gibbs Sampling Strategy for Multiple Alignment, Science, Vol. 262, No. 513, pp.208-214, October 1993.

[2] M. A. Larkin, G. Blackshields, N. P. Brown, R. Chenna, P. A. McGettigan, H. McWilliam, F. Valentin, I. M. Wallace, A. Wilm, R. Lopez, J. D. Thompson, T. J. Gibson and D. G. Higgins: Clustal W and Clustal X version 2.0, Bioinformatics, Applications Note, Vol.23 No.21, pp.2947-2948, 2007.

[3] PROSITE：http://prosite.expasy.org/

[4] Jorjia G. Henikoff and Steven Henikoff: Using Substitution Probabilities to Improve Position-specific Scoring Matrices, Computer Applications in the Biosciences, Vol.12,No.2, pp135-143, Aplil 1996.

[5] Steven Henikoff and Jorja G.Henikoff: Amino Acid Substitution Matrices from Protein Blocks, Proceedings of Natural Academy of Science of the United States of America, Vol.89, pp10915-10919, November 1992.