情報処理学会論文誌　数理モデル化と応用　Vol.9　No.3　32–43 (Dec. 2016)

ギブスサンプラに基づく アミノ酸配列モチーフの高精度抽出法

高橋 誉文[1]　北上 始[1]　福本 翔平[1]　森 康真[1]　田村 慶一[1]

受付日 2016年6月5日，再受付日 2016年7月27日，
採録日 2016年8月19日

概要：アミノ酸配列データベースから類似部分配列を抽出することとして知られている従来のギブスサンプリング法の抽出精度を向上させるために，多重整列化に基づく新しい方法を提案する．従来のギブスサンプリング法の抽出精度は初期値に大きく左右される．この点に着目し，提案手法では，できるだけ良い初期値を計算するため，配列データセットに対して多重整列化を行い，ある幅のウインドウを多重整列上にスライドさせ，p 値が最小となるウインドウ領域（類似部分配列）を初期値として選択する．多重整列化によって挿入されるギャップについては，ランダムに文字をあてはめる場合とすべての文字が等確率に現れる場合を比較する．また，ギブスサンプリングで利用される擬似度数に進化的な知識を導入し，抽出される類似部分配列としての配列モチーフ（進化的に保存される配列パターン）の抽出精度を向上させている．

キーワード：ギブスサンプリング，アミノ酸配列，モチーフ検索，多重整列化

Method for High-precision Motif Extraction Based on Gibbs Sampler in Amino Acid Sequences

Yoshifumi Takahashi[1]　Hajime Kitakami[1]　Syouhei Fukumoto[1]　Yasuma Mori[1]
Keiichi Tamura[1]

Received: June 5, 2016, Revised: July 27, 2016,
Accepted: August 19, 2016

Abstract: In order to improve the extraction accuracy of the existing, well-known Gibbs sampling method for extracting similar subsequences from amino acid sequence databases, we propose a new extraction method based on multiple sequence alignment in the sequence dataset. The extraction accuracy of the existing Gibbs sampling method is highly dependent on the initial solution selected randomly. In focusing on this point, the proposed method performs multiple sequence alignment for the sequence dataset to calculate the best possible initial solution. After that, we slide the aligned sequences on a window of a certain width and select the window region including the set of subsequences, where p-value is minimized, as the initial solution. In order to confirm the effectiveness of the proposed method, we carried out comparative experiments with random distribution and equal distribution. Moreover, we improve the accuracy of the existing Gibbs sampling method by using an amino acid substitution matrix as the knowledge of molecular evolution for pseudocount.

Keywords: Gibbs sampling, amino acid sequence, motif search, multiple alignment

1. はじめに

アミノ酸の配列モチーフは，生物進化の過程で保存された生物学的な機能やタンパク質の立体構造と深く関係する．ギブスサンプラは，配列データベースから配列モチーフにできるだけ近い類似部分配列を抽出する方法であり，多くの研究 [1], [2], [3] が行われている．また，ギブスサンプラを用いて，配列モチーフに近い類似部分配列を抽出する Web サービス [4] も行われている．

機能が未知のアミノ酸配列から配列モチーフを特定することができれば，たとえば，その構造の推定をはじめとし

[1] 広島市立大学大学院情報科学研究科
Graduate School of Information Sciences, Hiroshima City University, Hiroshima 731–3194, Japan

て，SURFNET [5], [6] などにより，活性部位や結合ポケットなどのような機能部位の特定につなげることができるものと期待されている [7]．また，このような機能部位の解析は，医薬品の開発に重要な分子シミュレーションや分子設計などを支援するものと期待されている．

ギブスサンプラは，Geman 兄弟 [8] によって画像処理の分野で利用され，画像復元に対する有効な統計的手法として紹介された．バイオインフォマティクス分野では，ギブスサンプラは，アミノ酸配列データセットの各配列から配列モチーフに最も近い類似部分配列を 1 個のみ抽出する方法として，Lawrence ら [1] によって紹介された．本論文では，これをサイトサンプラと呼び，各配列から 0 個以上の配列モチーフに最も近い類似部分配列を抽出するモチーフサンプラと区別する [2], [9], [10]．いずれも，マルコフ連鎖モンテカルロ法 [11], [12] の 1 つとして分類されているギブスサンプラに該当する．なお，モチーフサンプラの計算では，サイトサンプラの計算結果が利用されているが，モチーフサンプラの計算精度を向上させるには，サイトサンプラの計算結果が生物進化の過程で保存された配列モチーフにできるだけ近いことが重要である．

バイオインフォマティクス分野でのギブスサンプラは，アミノ酸配列データセットから配列モチーフの抽出問題に適用する研究 [1], [2], [13], [14], [15] と，DNA 配列データセットから DNA 配列上のタンパク質結合部位（配列モチーフ）の識別問題に適用する研究 [3], [4], [7], [16], [17], [18] とに分類される．

アミノ酸配列データセットや DNA 配列データセットから配列モチーフに最も近い類似部分配列を抽出するために，Lawrence らのサイトサンプラに対して，さまざまな改良法が提案されている．これらの手法は，機械学習や統計学などを導入し，数学的な最適解の高精度な探索に努力が払われてきたが，生物進化の過程で発生するアミノ酸置換の相対頻度を表す表（アミノ酸置換行列）をサイトサンプラの計算モデルに十分に反映させていないため，従来の計算モデルから数学的に厳密な類似部分配列が得られても，その類似部分配列はバイオインフォマティクス分野の専門家にとって興味ある配列モチーフから外れていることがある．

本論文では，アミノ酸配列データベースから配列モチーフにできるだけ近い類似部分配列を抽出するために，Lawrence らによって紹介されたサイトサンプラに基づき，新しい計算モデルに基づく計算手法を提案する．

サイトサンプラ（ギブスサンプラ）は，ユーザにより与えられた配列モチーフの長さ K をもとに，N 本の配列データを含む配列データセットから配列モチーフの候補である $N \times K$ の整列行列（K-類似部分配列の集合）を出力する方法である．配列データセットに含まれる N 本の配列データの長さは同一ではないが，簡単のため L とすると，配列データセットには，$(L-K+1)^N$ 個の $N \times K$ の整列行列が存在するため，これらの中から配列モチーフに近い類似部分文字列を直接探索するのは現実的ではない．サイトサンプラでは，配列データセットに含まれる N 本の各配列データからランダムに選択された K-部分配列の集まりを候補配列集合として選択し，それらを初期値とすることにより，候補配列集合を更新しながら配列モチーフとして最も確からしい類似部分配列を見つけ出そうとする確率的アルゴリズムである．しかしながら，アミノ酸配列データセットを対象にした従来のサイトサンプラには，以下のような問題がある．

（1）初期値のランダム性

サイトサンプラは，局所最適解を回避する工夫がなされていないため，計算精度が初期値に大きく依存し，計算結果に最適解が保証されていない．初期値のランダム性を低減させるために，N 本の配列からいくつかのサンプルをランダムに選択し，それらのサンプルのみから貪欲探索により最良の初期値を取り出す方法 [2], [10] があるが，初期値のランダム性に関する本質的な解決にはなっていない．

（2）進化的知識としてのアミノ酸置換行列の欠如

配列データセットからランダムに選択された 1 本の配列データを Z とする．サイトサンプラでは，候補配列集合から Z 上に存在する部分配列 X を探し，それまでに計算された文字の出現確率を表すプロファイル行列に適合する部分配列 X' を Z 上から確率的に選択し，候補配列集合内の X を X' に置き換え，候補配列集合を更新している．しかし，アミノ酸の置換のしやすさを数値化したアミノ酸置換行列 [19] についての知識が考慮されていないため，X' が適切に選択されない．

本論文では，これらの 2 つの問題点を解決するために，Thompson ら [20] や Larkin ら [21] の文献で提案されている案内木に基づく多重整列化（マルチプルアラインメント）に加え，Henikoff らが提案した擬似度数 [22], [23] をサイトサンプラの計算モデルに組み込んだ新しい抽出法を提案する．なお，案内木は近隣結合法 [24] などで作成された分子進化系統樹（生物進化の枝分かれの様子を描画した木構造）が多く利用されている．提案手法の要点は以下のとおりである．

（1）初期値のランダム性の問題を解決するために，できるだけ良質な初期値を計算する．そのために，まず，配列データセットに対して，分子進化系統樹に基づく多重整列化 [20], [21] を利用し，$N \times L$ の整列行列を導出する．この整列行列にスライディングウインドウ法を適用し，ある幅を持つウインドウを左から右に 1 文字ずつスライドさせることにより，すべてのウインドウ領域を列挙する．これらのウインドウ領域から配列モチーフに最も近いウインドウ領域 [25] を選択する．この選択の評価尺度として p 値を用いる．ただし，配列モチーフの両端にはどちらもギャップが含まれないため，ウインドウ領域の両端の少なくとも

一方に閾値以上のギャップ数が存在する場合は，そのウインドウ領域を選択の対象から除外する．

（2）進化的知識としてのアミノ酸置換行列が欠如しているという問題に対処するために，プロファイル行列の計算式に現れる擬似度数に，アミノ酸置換行列 [19], [26] に基づく知識を組み込む [25]．アミノ酸置換行列は，進化的な知識であり，生物進化の過程で発生するアミノ酸置換の相対頻度を行列 [19] で表現したものとして知られている．

以上のような提案方法について，実験により有効性を確認するために，5 種類のデータセットを用いて，提案手法により抽出された類似部分配列がすでに登録されている配列モチーフにどれだけ近いかを評価した結果，1 件のデータセットを除き約 90%以上も近いことが分かった．

本論文の構成は以下のとおりである．2 章では，関連研究について述べる．3 章では従来手法であるサイトサンプラ（ギブスサンプラ）について述べる．その中でプロファイル行列を用いた出現頻度の計算法や背景頻度の計算方法を説明する．4 章では多重整列化に基づく提案手法について述べ，5 章では従来手法と提案手法による計算結果を比較して評価を行う．最後の 6 章では本論文のまとめと今後の課題について述べる．

2. 関連研究

アミノ酸配列や DNA 配列などの配列データセットから配列モチーフに最も近い類似部分配列を抽出する方法には，列挙法 [27], [28], [29]，隠れマルコフモデル [30]，ギブスサンプラ [1], [2], [3], [4], [7], [13], [14], [15], [16], [17], [18], [31], [32] などがある．

列挙法とは，配列モチーフ内に存在するワイルドカード（任意の文字と一致する記号）を考慮して，PrefixSpan [27] のアプローチにより頻出パターンをすべて抽出する方法である．しかし，非常に多くの頻出パターンが列挙されるため，その中から配列モチーフに最も近い頻出パターンを探し出すことが困難である．進化的知識としてアミノ酸置換行列（生物進化の過程で発生するアミノ酸置換の相対頻度を表す表）を計算モデルに組み込んでいないことも原因の 1 つと考えられる．

隠れマルコフモデルは，ユーザがあらかじめ設計したネットワーク構造に基づいて，状態遷移確率や出力確率などのモデルパラメータを計算する方法である．このネットワーク構造に対するモデルパラメータはプロファイル HMM と呼ばれている．あるヒューリスティクス [33] を用いて，長さ K に対するプロファイル HMM が決定されると，Viterbi アルゴリズムを用いて，N 本の部分配列を含む類似部分配列集合の確率が最大の組を選択する．この類似部分配列には，文字の挿入や欠失が考慮されているのに対して，ギブスサンプラでは，これらは考慮されていない．しかし，一般的なネットワーク構造が存在しないため，モデルパラメータの計算に必要なネットワーク構造は，ユーザ自身があらかじめ与える必要がある．すなわち，ユーザは，なんらかの知識をもとに，アミノ酸配列データセットに合致したネットワーク構造を設計しなければならないという煩わしさがある．また，Durbin ら，Krogh ら，Eddy ら，Hughey らが文献 [30], [34], [35], [36] で隠れマルコフモデルの計算に必要なモデルパラメータの推定の手間を削減する方法について研究している．しかし，進化的知識としてのアミノ酸置換行列が考慮されていないため，満足のいく結果が得られるとは限らない．

ギブスサンプラの研究には，与えられた配列データセットの各配列から 1 つのみの配列モチーフを探索する研究と 0 個以上の配列モチーフを探索する研究がある．これらの研究は，アミノ酸配列データセットから配列モチーフの抽出に適用する研究 [1], [2], [10], [13], [14], [15] と，DNA 配列データセットからタンパク質と結合する DNA 配列上の部位（配列モチーフ）の識別問題に適用する研究 [3], [4], [7], [16], [17], [18], [31], [32] とに分類される．本論文で着目しているギブスサンプラは，サイトサンプラと呼ばれ，アミノ酸配列データセットの各配列から配列モチーフの最も近い類似部分配列を 1 つのみ抽出する方法 [1], [2], [9], [10] である．

アミノ酸配列データセットの各配列から配列モチーフに最も近い類似部分配列を 1 つのみ抽出する研究に着目すると，これまでの研究では進化的知識としてアミノ酸配列行列を計算モデルに十分に組み込まれていない．このため，従来の計算モデルを用いる限りは，たとえ数学的な厳密解が得られても，抽出された類似部分配列がバイオインフォマティクス分野の専門家にとって興味ある配列モチーフから外れてしまうことがある．

本論文では，これらの問題を解決するために，分子進化系統樹を案内木として利用されている多重整列化および相対エントロピーを評価尺度とするスライディングウインドウ法によりサイトサンプラの初期値を計算する．次に，プロファイル行列の計算に組み込まれている擬似度数に進化的知識としてのアミノ酸置換行列を導入する．これにより，従来のサイトサンプラよりも配列モチーフにできるだけ近い類似部分配列を各配列から抽出している．

3. 従来のサイトサンプラ

サイトサンプラは，ギブスサンプラの一種であり，配列データセットの各配列から配列モチーフに最も近い類似部分配列を 1 件のみ抽出する方法である．アミノ酸配列データセットの各配列は，さまざまな長さの文字列であり，その表現には M 種類の文字が使われているとする．ここでは，この M 種類の文字からなる文字集合を Ω と表記する．

図 1 は，サイトサンプラの処理イメージを図示したものである．アミノ酸配列モチーフに最も近い K-類似部分

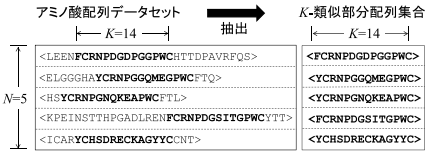

図 1　配列データセット DS と K-部分配列集合との関係
左図の太字は K-部分配列集合として抽出される文字を意味し，細字は背景配列集合と呼ばれる

Fig. 1　Relations with sequence data set DS and the K-subsequence set.

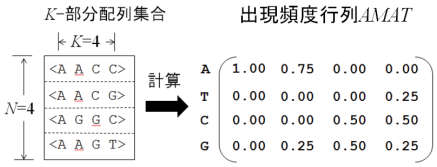

図 2　K-部分配列集合に対する出現頻度行列 $AMAT = (c_{ij})$

Fig. 2　Appearance frequency matrix $AMAT = (c_{ij})$ for K-subsequence set.

配列（長さ K の類似部分配列）を各配列から探索するために，サイトサンプラではアミノ酸配列データセット DS の各配列から長さ K の候補配列が 1 つのみ選択される．これにより選択された候補配列集合を $N \times K$ 行列と見なし，この行列からある行をランダムに削除し，削除された $(N-1) \times K$ 行列を用いて $M \times K$ プロファイル行列を計算する．次に，直前に計算されたプロファイル行列を用いて候補配列集合に含まれる 1 要素を更新する．このような処理を繰り返し，各配列から配列モチーフに最も近い K-類似部分配列が探索される．繰り返す試行回数に比例して計算時間がかかる．

本章では，まず，サイトサンプラ [1], [2], [9], [10] で中心的な役割を演じているプロファイル行列，背景頻度行列，オッズ比について説明する．次に，サイトサンプラのアルゴリズムを説明する．最後に，サイトサンプラの問題点について述べる．

3.1　プロファイル行列

K-類似部分配列の統計的な特徴を表現したものはプロファイル行列 $PMAT = (p_{ij})$ と呼ばれている．素朴なサイトサンプラでは，プロファイル行列として出現頻度行列 $AMAT = (c_{ij})$ が利用されている．出現頻度行列 $AMAT$ は，K-部分配列集合の列 j ごとに出現する文字 α_i の頻度を表現する $M \times K$ の行列 $AMAT = (c_{ij})$ である．

出現頻度行列 $AMAT$ の例を示すために，4 本の部分配列からなる 4-部分配列集合 $<AACC>, <AACG>, <AGGC>, <AAGT>$ について考えてみよう．ただし，アミノ酸配列を表現する文字集合 Ω の要素数 M は 20 であるが，この例では説明を簡単にするため，$\Omega = \{A, T, C, G\}$ としている．**図 2** にこの 4-部分配列集合に対する出現頻度行列 $AMAT$ の例を示す．図中の出現頻度行列 $AMAT$ では，どの列もすべての文字が現れる頻度を合計すると 1 となるように頻度が定められている．これにより，列ごとに現れやすい文字と現れにくい文字を表現できるようになる．

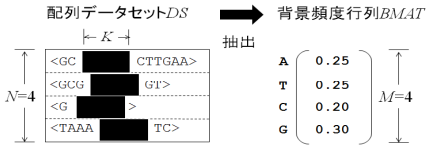

図 3　配列データセット DS と背景頻度行列 (q_i)

Fig. 3　Sequence dataset DS and Background frequency matrix (q_i).

プロファイル行列として出現頻度行列 $AMAT = (c_{ij})$ を使用してみよう．アミノ酸配列データセット DS 内に含まれる配列を Z とし，その長さを $|Z|$ と表記すると，配列 Z には $|Z|-K+1$ 個の K-部分配列が存在する $(|Z| \geq K)$．それらの中に含まれる K-部分配列の 1 つを $x = <\alpha_1\alpha_2 \ldots \alpha_K>$ と表記すると，プロファイル行列を用いて，その出現頻度（生起確率）P_x を次式のように計算することができる．

$$P_x = c_{11} \times c_{22} \times \ldots \times c_{KK}$$

ただし，x の部位に存在する文字 $\propto_i$ がプロファイル行列の i 行目に対応し，c_{ij} は j 列目における文字 $\propto_i$ の出現する頻度を意味する．これにより計算された x の出現頻度 P_x が高ければプロファイル行列の計算に用いられた K-部分配列集合に類似し，低ければ類似しないと解釈される．

さて，配列データの本数が少ないことなどが原因で，$AMAT = (c_{ij})$ の要素 c_{ij} に 0 が出現すると，他の要素の値がいくら大きくても出現頻度 P_x が 0 になってしまうことがある．これを避けるため，文献 [1], [2], [9], [10] では，ベイズ統計解析を導入し，アミノ酸配列データセット DS から配列 Z を取り除いた $N-1$ 本の K-部分配列に対する，プロファイル行列 $PMAT = (p_{ij})$ を以下のように定義している．

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

ただし，N は DS に含まれる配列総数，B は $\sqrt{N}$ と定める．$N-1$ 本の K-部分配列は，DS からある配列データ Z を除いた DS' から得られる．また，p_{ij} の i 行目に該当する文字の全配列に対する相対出現頻度を f_i とすると，p_{ij} 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，これにより分子のゼロ除算を回避することができる．

3.2　背景頻度行列

図 3 の例で示されるように，配列データセット DS から K-部分配列集合を除いた結果は背景配列集合と呼ばれる．背景頻度行列 $BMAT$ とは，背景配列集合に出現する文字 $\propto$ の頻度を表現する $M \times 1$ の行列 $BMAT = (q_i)$ である．q_i は，i 行目に対応する文字 $\propto_i$ の背景頻度を意味し，文字

$\propto_i$ の背景頻度は，背景配列集合内に存在する文字の総数に対する文字 $\propto_i$ の出現頻度（生起確率）である．

DS 内のある配列データ Z に存在する K-部分配列 $x = <\alpha_1\alpha_2 \ldots \alpha_K>$ の背景頻度 Q_x は以下のとおりである．$Q_x = q_1 \times q_2 \times \ldots \times q_K$ ただし，q_i は文字 $\propto_i$ の背景出現を意味する．これにより計算された x の背景頻度が高ければ，K-部分配列集合に非類似となり，低ければ類似していると解釈できる．

3.3　オッズ比

素朴なサイトサンプラでは，DS 内に含まれる配列 Z からプロファイル行列に適合する K-部分配列 x を計算するために出現頻度 P_x だけを用いるが，文献 [1], [2], [9], [10] では，次式で定義されるオッズ比 A_x あるいは対数オッズ比 $\log_2 A_x$ が用いられている．

$$A_x = P_x \div Q_x$$

オッズ比 A_x が高い x を配列データ Z から選択することは，K-類似部分配列集合に類似している（出現頻度 P_x が高い）と同時に背景配列集合に似ていない（背景頻度 Q_x が低い）x を配列データ Z から選択することを意味する．逆にオッズ比 A_x が低ければ，x は K-類似部分配列集合に似ていないと同時に背景配列集合に近い x を配列データ Z から選択することを意味する．

3.4　アルゴリズム

サイトサンプラは N 本の配列からなる DS からランダムに選択された配列 Z を用いることで，プロファイル行列と背景頻度行列を算出し，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) DS の各配列に対して K-部分配列の開始点 st_i をランダムに選び，それらを行列順に整列させた N 本の K-部分配列集合 $\{st_1, st_2, \ldots, st_N\}$ をサンプルとする．
(2) サンプルを 50 個作成し，最も良質なサンプルを初期値とする．
(3) DS からランダムに 1 つの配列 Z を選択する．
(4) $DS' = DS - \{Z\}$ から，$N-1$ 本の K-部分配列に対するプロファイル行列 $PMAT = (p_{ij})$ と背景配列集合に対する背景頻度行列 $BMAT = (q_i)$ を算出する．
(5) 配列 Z 内に存在する $|Z| - K + 1$ 個の K-部分配列 x から，それぞれの出現頻度 P_x および背景頻度 Q_x を計算し，類似度スコア A_x を計算する．すなわち，x のオッズ比 $A_x = P_x \div Q_x$ あるいは対数オッズ比 $\log 2A_x$ を算出する．
(6) $\{A_1, A_2, \ldots, A_{|Z|-k+1}\}$ 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-部分配列集合の新たな開始点 st'_Z を st_Z に置き換える．
(7) (2)〜(6) をユーザが定めた回数分繰り返す．繰返しが終わった $\{st_1, st_2, \ldots, st_N\}$ を結果とする．繰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．

3.5　相対エントロピー

抽出した K-部分配列集合が配列モチーフとしてどのぐらい近いかを評価するために，相対エントロピーが利用されている [1], [2], [9], [10]．相対エントロピーは，次のように定義される．

$$F = \sum_{i=1}^{K} \sum_{j=1}^{M} c_{ij} \log 2(p_{ij}/q_i) \tag{2}$$

ただし，この定義では，$M \times K$ の出現頻度行列 $AMAT = (c_{ij})$，$M \times K$ のプロファイル行列 $PMAT = (p_{ij})$，$M \times 1$ の背景頻度行列 $BMAT = (q_i)$ が用いられている．K-部分配列集合の相対エントロピーが大きければ配列モチーフに近く，小さければ配列モチーフから離れているものと判断している．

3.6　サイトサンプラの問題点

サイトサンプラのアルゴリズムは，計算精度が初期値（3.4 節の (1) で与えられる開始点）に大きく依存する．初期値による影響を少なくするため，SA 法 [37] や遺伝的アルゴリズム [38] の利用が考えられる．しかし，従来の相対エントロピーを SA 法の目的関数や遺伝的アルゴリズムの適応度に利用されるため，配列モチーフとは異なる類似部分配列が得られることがある．すなわち，従来の相対エントロピーの計算式は，K-部分配列集合内の配列どうしの類似度を表すが，K-部分配列集合内の各配列と配列モチーフとの近さを表すものではない．

サイトサンプラで計算される K-類似部分配列を配列モチーフにできるだけ近づけるためには，できるだけ品質の高い初期値を計算することや，相対エントロピーの計算式を改善することが重要となる．

4.　提案手法

提案手法では，ランダムに初期値を与えることを避けるため，多重整列を用いて初期値（K-類似部分配列集合）の探索を行う．この多重整列は，あらかじめ，分子進化系統樹を用いた多重整列化（マルチプルアラインメント）により求めたものである．従来のサンプラでは，プロファイル行列は，処理手順 (4) における出現度数をはじめとして，処理手順 (6) の相対エントロピーの計算で利用されおり，プロファイル行列の計算式には，擬似度数が組み込まれている．提案手法では，従来の擬似度数の計算式に，アミノ酸置換のしやすさを数値化したアミノ酸置換行列（進化的な知識）を導入している．以上の提案手法によるサイトサ

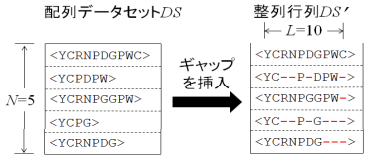

図 4　多重整列化の例
Fig. 4　Example of multiple alignment.

図 5　ギャップに対してランダムに文字を割り当てた例
Fig. 5　Example assign a character at random for a gap.

ンプラにより，配列モチーフにできるだけ近い K-類似部分配列を抽出しようとしている．

本章では，まず，多重整列化の方法について述べた後，多重整列から安定した初期値を探索する方法について述べる．次に，プロファイル行列の擬似度数の計算式にアミノ酸置換行列を導入する方法について述べる．最後に，配列モチーフとしての K-類似部分配列集合を抽出するアルゴリズムについて述べる．

4.1　多重整列化の方法

N 本の系列データ（時系列データやテキストデータなど）に対する多重整列化では，編集距離を尺度とする動的計画法 [39] が利用されている（$N \geq 3$）．この多重整列化は，非類似度スコア（編集距離）の累積が最小となるように，系列データの適当な場所にギャップを挿入し，系列データ間の文字の対応付けを行うことにより，すべての系列データの長さを揃えている．

N 本の配列データを含むアミノ酸配列データセットに対する多重整列化では，非類似度スコアの累積が最小となる方法を利用せず，類似度スコアを尺度とする動的計画法 [39] が利用されている．これにより累積類似度スコアが最大になるように，系列データの適当な場所にギャップを挿入し，配列データの長さを揃えている [40], [41]．なお，類似度スコアの計算には，生物進化の過程で発生する文字どうしの置換のしやすさを表す置換行列を利用している．また，進化的に近縁の配列データどうしを整列化した方が精度向上につながるため，分子進化系統樹を案内木とする多重整列化を行っている [20], [21]．

N 本の配列データを多重整列化することにより，配列長が L になったとしよう．以下では，多重整列化した結果を $N \times L$ 行列と見なし，これを $N \times L$ の整列行列と呼ぶ．**図 4** に多重整列化により得られる整列行列の例を示す．ギャップと呼ばれる記号（–）は，多重整列化において，文字の挿入や削除の操作によって追加される記号である．これにより，大域的に一番類似するように配列データの長さを一致させている．図では，多重整列化によりアミノ酸配列データセット DS の長さを一致させることにより得られる整列行列を DS' としている．分子進化学の専門家は，この整列行列 DS' を用いて，配列モチーフ領域を経験的に探索している [30]．

多重整列化により配列データ上に挿入されたギャップは長さ K の配列モチーフ領域上にも挿入される．このため，整列行列上に存在する配列モチーフ領域の長さ K' は一般に K 以上になってしまう．すなわち，長さ K の配列モチーフ領域において，アミノ酸の数よりもギャップの数が多く存在する列の数を K_g とすると，$K' = K + K_g$ の関係が成立する．

4.2　整列行列を用いた初期値の選択法

整列行列上のある特定領域（連続する K 個の列）に配列モチーフが多く含まれることが経験的に知られている．これをふまえ，多重整列化により得られる整列行列 DS' から良好な初期値を探索するための新しい方法を提案する．以下では，ギャップが含まれる整列行列からプロファイル行列を計算する方法，初期値を見つけ出すためのスライディングウインドウ法，スライディングウインドウ法の精度向上に重要となる非配列モチーフ領域の判定法について述べる．

(1) プロファイル行列の計算法

多重整列化によって挿入されたギャップは，配列の長さを揃えるために挿入された空白である．このため，整列行列において，ギャップが混在する列に存在する文字だけを考えて，文字の出現頻度を計算すると，他のギャップの少ない列に比べて，その値が大きくなり，あたかもモチーフ領域の一部と見なされてしまう．これを避けるために，$N \times L$ 整列行列（多重整列）上に存在するある $N \times K'$ 配列領域の $M \times K'$ プロファイル行列を算出するには，あらかじめギャップを考慮した計算方法を定める必要がある（$K' \leq L$）．

著者らは，ギャップを考慮した計算方法として 2 つの方法を提案する．まず，整列行列 DS' 内に存在する各ギャップに対して，20 種類のアミノ酸文字をランダムに割り当てる方法である．**図 5** は各ギャップに文字をランダムに割り当てた例である．左側はギャップを含む 5×10 整列行列 DS' であり，右側は各ギャップをランダムに割り当てた結果として得られる行列 DS'' である．このような行列 DS'' を用いると $M = 20$ の $M \times K'$ プロファイル行列 $PMAT = (p_{ij})$ を容易に計算することができる．以上により，ギャップが多く含まれている列にランダム性を持たせて，プロファイル行列を計算している．次に，すべての文字にギャップ文字の出現頻度を均等に分ける方法の均等割

付けである．図 5 の左図の 6 文字目を例に説明する．出現頻度は，文字 D が 2，文字 G が 1，そしてギャップ文字が 2 である．ギャップ文字の出現頻度 2 を 20 種類のアミノ酸に割り振ると，文字 D が 2.1，文字 G が 1.1，その他の文字が 0.1 となる．このように，出現頻度を均等に分けることで，ランダムに割り当てたときに偏りがないようにしている．

(2) スライディングウインドウ法

スライディングウインドウ法とは，整列行列の左端列から右端列へ向かってウインドウをスライドさせることにより，配列モチーフに近いウインドウ領域を見つけ出し，その領域の先頭位置をサイトサンプラの初期値とする方法である．

ウインドウの長さ K' はその位置によって異なり，可変である．すなわち，ウインドウ内に存在する非ギャップ列の数 K はウインドウの位置に依存せず固定だが，ギャップ列の個数 K_g はウインドウの位置に依存する（$K' = K + K_g$）．

ウインドウをスライドさせ，配列モチーフに近い領域を見つけ出す尺度として，4.3 節で述べる p 値を利用している．

(3) 非配列モチーフ領域の判定法

PROSITE などのモチーフデータベースに登録されている配列モチーフの最左端または最右端に，ギャップは存在しない．このため，ウインドウの最左端列または最右端列にギャップが存在している場合は，そのウインドウに配列モチーフが多く含まれる可能性があるかどうか吟味する必要がある．

ウインドウの最左端列または最右端列が，以下を満たすとき，そのウインドウを非配列モチーフ領域と呼ぶ．

$$N_g \geq N \times \eta \tag{3}$$

ただし，N_g は列に存在するギャップ数，η は閾値を意味する．η は値をいくつか変化させてみたところ，0.1 が一番良い結果となったので．実験ではその値を採用する．

スライディングウインドウ法では，非配列モチーフと判断されるウインドウは，初期値の探索から外される．

4.3　進化的な知識を導入した擬似度数の計算法

提案手法における出現頻度や相対エントロピーの計算で用いられるプロファイル行列では，従来のサイトサンプラにはない進化的な知識を導入する．具体的には，プロファイル行列の定義に利用されている擬似度数にアミノ酸置換のしやすさを数値化したアミノ酸置換行列（進化的な知識）を導入する．このために，式 (1) のプロファイル行列 $PMAT = (p_{ij})$ を次のように定義する [42].

$$p_{ij} = (c_{ij} + b_{ij})/(N_j + B_j) \tag{4}$$

ただし，b_{ij} は以下で定義される擬似度数，$N_j = \sum c_{ij}[1 \leq i \leq M]$ を意味する．

$$b_{ij} = B_j \times \sum (c_{kj}/N \times g_{ki}/G_i)[1 \leq k \leq M] \tag{5}$$

B は $B_j = m \times R_j$ と定義されており，R_j を 1 クラスタにおける j 列目の文字の種類数とする．m は試験的な検索実験により決定される正の数であり，$m = 5 \sim 6$ が最も有効であると報告されている．また，g_{ki} はアミノ酸 k からアミノ酸 i への置換頻度で，次式で表される．

$$g_{ki} = g_k \times g_i \times 2^{s(k,i)} \tag{6}$$

$s(k,i)$ はアミノ酸 k からアミノ酸 i への置換のしやすさを数値化した類似スコアであり，この類似スコアは，BLOSUM62 と呼ばれるアミノ酸置換行列から取得している．g_i は DS'' 内の i の出現確率で，G_i は g_{ki} の文字ごとの総和であり，$G_i = \sum g_{ki}[1 \leq k \leq 20]$ と定められている．

進化的な知識を考慮した式 (4) は，式 (2) で紹介した相対エントロピーの計算に必要になる．しかし，この相対エントロピーでは，モチーフ領域が本当に類似しているかが分かりにくいため，評価には以下の計算式に基づく p 値 [43] を用いている．

$$p - value \leq (N+1)^{K(M-1)} 2^{-N \times H} \tag{7}$$

ただし，H は以下のとおりである．

$$H = \sum_{j=1}^{k} \sum_{i=1}^{M} p_{ij} \frac{p_{ij}}{b_j} \tag{8}$$

4.4　アルゴリズム

提案手法では配列データセットを多重整列化し，それにより得られる整列行列 DS' に対してスライディングウインドウ法を適用する．これにより探索されたクラスタをサイトサンプラの初期値とする．次に，サイトサンプラを実行することにより，配列データセットの各配列から配列モチーフにできるだけ近い K-類似部分配列を抽出する．提案手法のアルゴリズムは以下のとおりである．

【入力】配列データセット DS，配列モチーフの長さ K アミノ酸配列を表現する文字集合 Ω，閾値 η，アミノ酸置換行列 $s(k,i)$，$k \in \Omega$，$i \in \Omega$

【出力】配列データセットの各配列から抽出される K-類似部分配列

(1) 分子進化系統樹を案内木として利用する多重整列化を配列データセットに適用し，整列行列 DS' を求める．

(2) 4.2 節の (1) で述べたように，DS の表現に利用されている文字の集合から文字をランダムに選び，それを整列行列 DS' 内のギャップに割り当て DS'' を得る．

(3) 4.2 節の (2) で述べたように，DS'' に対してスライディングウインドウ法を適用する．すなわち，$N \times K'$ のクラスタ（長さ K' の部分配列集合）を 1 文字ずつスライドさせ，最後のクラスタの処理が終わるまで，以

下の処理を繰り返す．

・4.2 節の (3) で述べた非配列モチーフ領域の判定法を用いて，当該クラスタが非配列モチーフ領域と判定されたときは，そのクラスタを候補配列集合から外す．

・4.3 節の式 (4) で述べた擬似度数を計算し，その結果を用いて，当該クラスタに対するプロファイル行列 $PMAT$ を算出する．

・算出されたプロファイル行列 $PMAT$ を用いて，当該クラスタの p 値を 4.3 節で述べた方法で計算する．

(4) p 値が算出されたクラスタの集合から p 値の値が最も小さいクラスタ（K-部分配列集合）の開始位置をサイトサンプラの初期値として選択する．

(5) 初期値を用いて，3.4 節のサイトサンプラを (3) から実行する．ただし，このサイトサンプラで必要となるプロファイル行列 $PMAT$ の計算では，4.3 節の式 (4) で述べた擬似度数を用いる．

5. 評価実験

提案手法の有効性を確認するために，PROSITE と呼ばれるモチーフライブラリ（モチーフデータベース）[44] を使用した．モチーフライブラリに含まれるアミノ酸配列データセットのそれぞれには，現在までに見つかっている配列モチーフとそれを含む配列データが数多く収集されている．評価実験では，モチーフライブラリの中から 5 種類のアミノ酸配列データセット [45], [46], [47], [48], [49], [50] を選んだ．これらのデータセットの選定に際しては，データ件数の多いものや少ないものをはじめとして，配列長がほぼ同じものから大きく異なるものが含まれるように配慮した．

表 1 に 5 種類のアミノ酸配列データセットの概要を示す．表中のクラスタの長さ K' は，本来の配列モチーフ長 K に，多重整列化によって挿入された配列モチーフ内のギャップ長 K_g を加えたものを意味する．なお，予備実験として，閾値 η をいくつか変化させてみたところ，0.1 が一番良い結果が出たので，評価実験では，その値を採用している．

5.1 実験方法

多重整列化を行うプログラムとして，分子進化系統樹を案内木として利用する CLUSTAL X [22], [23] を使用している．CLUSTAL X は，計算途中で配列間の位置関係が凍結されてしまうため，完全に多重整列化された結果が出力されないといわれている．これを改善するため，CLUSTAL X などにより計算出力された多重整列に対して，適当な場所にギャップをさらに追加する反復改善法 [20] が開発されている．しかし，著者らの予備実験では，CLUSTAL X の方が反復改善法よりも多重整列中にギャップ数が少なく，良好な初期値が得られている．このため，多重整列化のプログラムとして CLUSTAL X を利用している．CLUSTALX のアラインメントのパラメータは，多重整列では BLOSUM series を使用している．すなわち，CLUSTALW では案内木を葉ノードから根ノードへたどりながら多重整列が行われており，案内木の各ノードに出現する 2 つの配列あるいは配列グループ間の関係の深さに応じて，BLOSUM80～BLOSUM30 が利用されている [20]．特徴として，BLOSUM は相同性検索に広く使われている．

以下では，提案手法に有用性を確かめるために導入した尺度として，計算精度について説明する．

従来手法と提案手法の実行において，処理の繰返し回数は，10,000 回とした．さらに両手法の試行回数はそれぞれ 10 回とし，それぞれ抽出した結果の平均を精度としている．従来手法や提案手法で予測された配列モチーフ領域（抽出された K-類似配列集合）の精度（適合率）については，以下のように計算する．

$$\text{精度 (\%)} = \frac{B}{B+C} \times 100 \tag{9}$$

ただし，B はある手法で予測されたモチーフ領域（予測領域）に含まれる正解領域の文字数を意味する．C は予測領域に含まれる非正解領域の文字数，すなわち，$K-B$ を意味する．**図 6** を例としてあげると，抽出した K-部分配列集合の全文字数を分母とし，その中でマッチした配列モチーフ領域の全文字数を分子とすることで，適合率が求められる．また，抽出されなかった配列モチーフ領域を A とすると，$K = A+B = B+C$ なので $A = C$ となり，再現率とも一致する．よって式 (9) の数値が高いほど，抽出さ

表 1　実験に使用したデータセット
Table 1　Dataset of experience.

番号	モチーフ名	登録番号	クラスタの長さ ($K' = K + K_g$)	件数
1	Kringle	PS00021	14=14+0	95
2	Homeobox	PS00027	115=24+91	1321
3	PTS_EIIA	PS00372	22=17+5	51
4	HTH_ASNC	PS00519	37=27+10	43
5	HTH_DEOR	PS00894	35=35+0	82

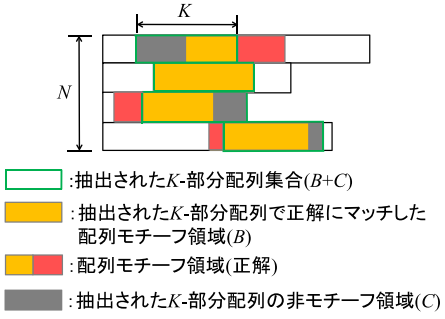

図 6　K-部分配列集合内で判断する配列モチーフ領域
Fig. 6　Sequence motif area to be determined in the K-subsequences set.

れた K-部分配列集合は配列モチーフ領域と一致している部分が多く，低いほど，配列モチーフ領域から外れていると解釈できる．

5.2　実験結果

表 2 にランダム割付けの初期値，**表 3** に BLOSUM の従来手法と提案手法であるランダム割付けと均等割付けの初期値と最終状態の精度を，**表 4** に総実行時間をそれぞれ示す．

表 2 から，ランダム割付けによって初期値が変化しないことが分かる．表 3 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，ランダム割付けより均等割付けの精度が良いことが分かる．しかし，4 番のデータセットにおいては，他のデータより配列モチーフの精度が低い．表 4 で初期値決定の時間は，従来手法が提案手法より早いことが分かる．また，サイトサンプラにかかる時間に差は見られなかった．

表 2　ランダム割付けの初期値
Table 2　Initial value of the random assignment.

実行	番号				
回数	1	2	3	4	5
1	4087	3113	889	756	337
2	4087	3113	889	756	337
3	4087	3113	889	756	337
4	4087	3113	889	756	337
5	4087	3113	889	756	337
6	4087	3113	889	756	337
7	4087	3113	889	756	337
8	4087	3113	889	756	337
9	4087	3113	889	756	337
10	4087	3113	889	756	337

5.3　考察

提案手法が効果をあげた理由としては，従来手法には考慮されていない進化的な知識を，多重整列化や p 値の計算に用いているためである．また，初期値を探索した後に適用するサイトサンプラに関しても進化的な知識を導入したプロファイルを計算しているので，抽出精度がより向上したと考えられる．

Kringle データを使用してサイトサンプラを行い反復回数 10,000 回までの $log_2(p$ 値 $)$ の変化を**図 7** に示す．これを見る限り，反復回数が 1,000 回を超えても，p 値がほとんど変化しておらず，精度面でも有意な改善が見られなかった．今回の実験では，初期値の決定で良質な解が得られた場合には，サイトサンプラで各配列に対して数回選択を行うことで，最良解を求めることができると考えられる．

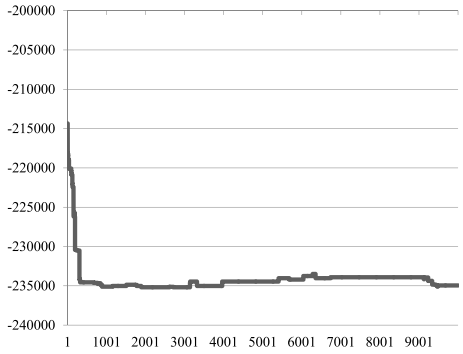

図 7　サイトサンプラの $log_2(p$ 値$)$ の変化
Fig. 7　Value of $log_2(p$-value) of the site sampler.

表 3　精度結果比較
Table 3　Accuracy results comparison.

モチーフ名	従来手法 (%)		ランダム割付け (%)		均等割付け (%)	
	初期値	最終状態	初期値	最終状態	初期値	最終状態
Kringle	1.19	64.44	56.34	88.38	56.34	88.38
Homeobox	0.64	87.75	76.69	99.84	76.69	99.84
PTS_EIIA	3.43	49.18	75.91	97.98	75.91	98.86
HTH_ASNC	2.12	41.76	51.88	53.84	53.75	59.88
HTH_DEOR	3.96	17.06	82.86	90.46	82.86	90.52

表 4　総実行時間比較（秒）
Table 4　Total execution time comparison (s).

モチーフ名	従来手法			ランダム割付け			均等割付け		
	多重整列	初期値選択	サイトサンプラ	多重整列	初期値選択	サイトサンプラ	多重整列	初期値選択	サイトサンプラ
Kringle	-	0.027	75.599	7.8	0.76	75.755	7.8	0.758	76.068
Homeobox	-	0.423	728.448	450.3	46.93	730.451	450.3	46.932	731.469
PTS_EIIA	-	0.016	39.232	12.4	0.27	38.109	12.4	0.269	38.798
HTH_ASNC	-	0.012	51.461	0.7	0.23	51.591	0.7	0.231	51.731
HTH_DEOR	-	0.023	24.132	1.7	0.19	24.917	1.7	0.194	24.565

また，ランダム割付けより均等割付けの精度が良い理由として，均等に出現頻度を分けることでギャップ以外の文字の出現頻度の順位が変化しないからだと考えられる．

初期値決定の計算時間が提案手法で多くかかった理由として，p 値の計算回数が異なることが考えられる．従来手法では 50 通りの組合せを比較しているが，提案手法ではスライディングウインドウ法を行い，すべてのクラスタを調べているので，従来手法より時間がかかっている．

また，サイトサンプラでは 3 つの手法でアルゴリズムや繰返し回数が同じなので，計算時間に差が見られなかったと考えられる．

6. まとめ

本研究では，アミノ酸配列データセットから配列モチーフに相当する類似部分配列を高精度に抽出する方法を提案した．解の品質を大きく左右する初期値をできるだけ良好なものにするために，ランダムに初期値を選択することを止め，進化的な知識が導入された多重整列化をアミノ酸配列データベースに適用し，これにより得られた整列行列から高い品質を持つ初期値を選択した．初期値の選択では，整列行列に対して，ウインドウをスライドさせるスライディングウインドウ法を適用している．スライディングウインドウ法で初期値を選択するとき，ウインドウ領域の両端の一方に，閾値 η を超えるギャップ数がある場合，そのウインドウ領域を初期値の選択から除外した．なお，初期値としてのウインドウ領域を選択するために，ギャップ文字をランダム割付けと均等割付けの 2 つの方法を用い，p 値の尺度を利用している．また，このような初期値の選択のほかに，従来の GS アルゴリズムに含まれている (1) 部分配列の出現頻度や (2) 候補配列集合の p 値の計算に進化的な知識を導入した．

提案方法の評価実験を 5 つのデータセットを用いて実施した結果，従来の方法に比べて高精度な配列モチーフとしての類似部分配列集合の抽出が可能になった．

今後の課題として，偽陽性を取り除くための方法の検討があげられる．たとえば，それぞれの文字に対する背景頻度の計算に n 次のマルコフ過程を導入する方法などがある．このほかの課題としては，サイトサンプラが前提としている配列モチーフの長さ K を遺伝的アルゴリズムや島モデルなどにより自動決定する方法の検討などがある．また，配列モチーフを抽出するだけでなく，テキストデータにおける規則的な共通部分の探索，Web 文章の履歴や顧客の購買履歴の分析の利用可能性についての検証は残されている．さらに，実験などで機能相関があることが検証された配列データセットに対して，高精度なモチーフサンプラを行う研究が重要である．

参考文献

[1] Lawrence, C.E., Altscul, S.F., Boguski, M.S., Liu, J.S., Neuwald, A.F. and Wootton, J.C.: Detecting Subtle Sequence Signals: A Gibbs Sampling Strategy for Multiple Alignment, *Science*, Vol.262, No.513, pp.208–214 (1993).

[2] Liu, L.-f. and Jiao, L.-c.: A Greedy Two-stage Gibbs Sampling Method for Motif Discovery in Biological Sequences, *Journal of Information Science and Engineering*, Vol.26, pp.2309–2318 (2010).

[3] Thompson, W., Rouchka, E.C. and Lawrence, C.E.: Gibbs Recursive Sampler: Finding transcription factor binding sites, *Nucleic Acids Research*, Vol.31, Issue 13, pp.3580–3585 (2003).

[4] The Gibbs Motif Sampler Homepage, available from ⟨http://ccmbweb.ccv.brown.edu/gibbs/gibbs.html⟩

[5] Laskowski, R.A.: SURFNET: A program for visualizing molecular surfaces, cavities and intermolecular interactions, *Journal of Molecular Graphics*, Vol.13, Issue 5, pp.323–330 (Oct. 1995).

[6] 欧州バイオインフォマティクス研究所：SURFNET, 入手先 ⟨http://www.ebi.ac.uk/thornton-srv/software/SURFNET/⟩

[7] Lee, D., Redfern, O. and Orengo, C.: Predicting protein function from sequence and structure, *Nature Reviews Molecular Cell Biology*, Vol.8, pp.995–1005 (Dec. 2007).

[8] Geman, S. and Geman, D.: Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images, *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol.6, No.6, pp.721–741 (1984).

[9] Rouchka, E.C.: A Brief Overview of Gibbs Sampling, Bioinformatics Technical Report Series, No.TR-ULBL-2008-02, University of Louisville, p.9 (Mar. 24, 2008).

[10] Liu, L.-f., Jiao, L.-c. and Huo, H.-w.: A Greedy Two-stage Gibbs Sampling Method for Motif Discovery in Biological Sequences, *2008 International Conference on BioMedical Engineering and Informatics*, pp.13–17 (2008).

[11] Hastings, W.K.: Monte Carlo Sampling Methods Using Markov Chains and their Applications, *Biometrika*, Vol.57, pp.97–109 (1970).

[12] 国友直人, 山本　拓 (監修), 北川源四郎, 竹村彰通 (編)：21 世紀の統計科学，第 III 巻 数理・計算の統計科学，第 10 章 マルコフ連鎖モンテカルロ法入門，東京大学出版会 (2008).

[13] Neuwald, A.F., Liu, J.S. and Lawrence, C.E.: Gibbs motif sampling: Detection of bacterial outer membrane protein repeats, *Protein Science*, Vol.4, pp.1618–1632, Cambridge University Press (1995).

[14] 河野修久，加藤智之，田村慶一，北上　始：配列データベースから類似部分配列を抽出するための GS 最適化手法に関する考察，電子情報通信学会第 19 回データ工学ワークショップ（DEWS2008），Online Proceedings (2008).

[15] Kono, N., Kitakami, H., Tamura, K. and Mori, Y.: Extracting Similar Subsequences by Gibbs Sampling with Distributed MGG, *Proc. 2009 International Conference on Parallel & Distributed Processing Techniques & Applications* (*PDPTA '09*), Las Vegas in USA, July 13-16, pp.669–675 (2009).

[16] Liu, J.S.: The Collapsed Gibbs Sampler in Bayesian Computations with Applications to a Gene Regulation Problem, *Journal of the American Statistical Association*, Vol.89, No.427, pp.958–966 (1994).

[17] Liu, L.-f. and Jiao, L.-c.: Moitf GibbsGA: Sampling Transcription Factor Binding Sites Coupled with PSFM

Optimization by Genetic Algorithm, *Journal of Convergence Information Technology*, Vol.5, No10, pp.141–148 (2010).

[18] Thompson, W.A., Newberg, L.A., Conlan, S., McCue, L.A. and Lawrence, C.E.: The Gibbs Centroid Sampler, *Nucleic Acids Res.*, Vol.35, Issue suppl 2, pp.W232–W237 (2007).

[19] Styczynski, M.P., Jensen, K.L., Rigoutsos, I. and Stephanopoulos, G.: BLOSUM62 miscalculations improve search performance, *Nature Biotechology*, Vol.26, No.3, pp.274–275 (2008).

[20] Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F. and Higgins, D.G.: CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice, *Nucleic Acids Research*, Vol.22, No.22, pp.4673–4680, Oxford University Press (1994).

[21] Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R., Thompson, J.D., Gibson, T.J. and Higgins, D.G.: Clustal W and Clustal X version 2.0, *Bioinformatics*, Vol.23, Issue 21, pp.2947–2948 (2007).

[22] Henikoff, S. and Henikoff, J.G.: Amino Acid Substitution Matrices from Protein Blocks, *Proc. Natural Academy of Science of the United States of America*, Vol.89, pp.10915–10919 (Nov. 1992).

[23] Henikoff, J.G. and Henikoff, S.: Using substitution probabilities to improve position-specific scoring matrices, *Computer Applications in the Biosciences*, Vol.12, No.2, pp.135–143 (Apr. 1996).

[24] Saitou, N. and Ne, M.: The Neighbor-joining Method: A New Method for Reconstructing Phylogenetic Trees, *Molecular Biology and Evolution*, Vol.4, Issue 4, pp.406–425 (1987).

[25] 福本翔平，北上　始，森　康真：多重整列に基づくモチーフの統計的抽出法，電子情報通信学会第 13 回情報科学技術フォーラム（FIT2014）論文集，D-008，pp.91–93 (2014).

[26] Henikoff, S. and Henikoff, J.G.: Amino Acid Substitution Matrices from Protein Blocks, *Proc. National Academy of Sciences of the United States of America* (*PNAS*), Vol.89, No.22, pp.10915–10919 (1992).

[27] Pei, J., Han, J., Mortazavi-Asl, B., Wang, J., Pinto, H., Chen, Q., Dayal, U. and Hsu, M.-C.: Mining Sequential Patterns by Pattern-Growth: The Prefix Span Approach, *IEEE Trans. Knowledge and Data Engineering*, Vol.16, No.11, pp.1424–1440 (2004).

[28] 加藤智之，北上　始，森　康真，田村慶一，黒木　進：極小かつ非冗長な可変長ワイルドカード領域を持つ頻出パターンの抽出，電子情報通信学会和文論文誌 D「データ工学特集号」，Vol.J90-D, No.2, pp.281–291 (2007).

[29] 加藤智之，森　康真，黒木　進，北上　始：可変長配列パターン抽出法におけるギブスサンプリングを用いた不要パターンの除去方式，日本データベース学会論文誌（DBSJ Letters），Vol.6, No.1, pp.65–68 (2007).

[30] Durbin, R., Eddy, S.R., Krogh, A. and Mitchison, G.: Chapter5 Profile HMMs for sequence families, *Biological sequence analysis — Probabilistic models of proteins and nucleic acids*, pp.100–133, Cambridge University Press (1998).

[31] Newberg, L.A. et al.: A phylogenetic Gibbs sampler that yields centroid solutions for cis-regulatory site prediction, *Bioinformatics*, Vol.23, No.14, pp.1718–1727 (July 2007).

[32] Thompson, W.A. et al.: Using the Gibbs Motif Sampler for phylogenetic footprinting, *Methods Mol. Biol.*, Vol.395, pp.403–424 (2007).

[33] 北上　始，斎藤成也，太田聡史：ビッグデータ時代のゲノミクス情報処理，コロナ社 (2014).

[34] Krogh, A., Brown, M., Mian, I.S., Sjolander, K. and Hausder, D.: Hidden Markov Models in Computational Biology Applications to Protein Modeling, *Journal of Molecular Biology*, Vol.235, pp.1501–1531 (1994).

[35] Eddy, S.R.: Multiple alignment using hidden Markov models, *Proc. International Conference on Intelligent Systems for Molecular Biology* (*ISMB-95*), pp.114–120, AAAI/MIT Press (1995).

[36] Hughey, R. and Krogh, A.: Hidden Markov models for sequence analysis: Extension and analysis of the basic method, *Computer Applications in the Biosciences* (*CABIOS*), Vol.12, No.2, pp.95–107 (1996).

[37] Kirkpatrick, S., Gelatt, C.D. and Vecchi, M.P.: Optimization by simulated annealing, *Science*, Vol.220, pp.671–680 (1983).

[38] Goldberg, D.E.: Genetic Algorithms in Search, *Optimization and Machine Learning*, Addison-Wesley Longman Publishing Co., Inc. Boston, MA, USA (1989).

[39] Bellman, R.: *Dynamic Programming*, Princeton University Press (1957).

[40] Gusfield, D.: *Algorithms on Strings, Tree, and Sequences: Computer Science and Computational Biology*, Cambridge University Press (1977).

[41] Saul, N.B. and Christian, W.D.: A general method applicable to the search for similarities in the amino acid sequence of two proteins, *Journal of Molecular Biology*, Vol.48, Issue.3, pp.443–453 (1970).

[42] 福本翔平，北上　始，森　康真：アラインメントされた配列集合からモチーフを抽出する方法，電子情報通信学会第 5 回データ工学と情報マネジメントに関するフォーラム（DEIM2013）論文集，E5-2，Online Proceedings (2013).

[43] Zia, A. and Moses, A.M.: Towards a theoretical understanding of false positives in DNA motif finding, *BMC Bioinformatics*, Vol.13, No.151, p.9 (2012).

[44] PROSITE, available from ⟨http://prosite.expasy.org/⟩

[45] Ikeo, K., Takahashi, K. and Gojobori, T.: Evolutionary origin of numerous kringles in human and simian apolipoprotein(a), *Federadon of European Biochemical Societies*, Vol.287, No.1-2, pp.146–148 (1991).

[46] Gehring, W.J. and Hiromi, Y.: Homeotic genes and the homeobox, *Annu. Rev. Genet.*, Vol.20, pp.147–173 (1986).

[47] Postma, P.W., Lengeler, J.W. and Jacobson, G.R.: Phosphoenolpyruvate: Carbohydrate phosphotransferase systems of bacteria, *Microbiology and Molecular Biolopy Reviews*, Vol.57, No.3, pp.543–594 (1993).

[48] Willins, D.A., Ryan, C.W., Platko, J.V. and Calvo, J.M.: Characterization of Lrp, and Escherichia coli regulatory protein that mediates a global response to leucine, *Journal of Biological Chemistry*, Vol.266, No.17, pp.10768–10774 (1991).

[49] von Bodman, S.B., Hayman, G.T. and Farrand, S.K.: Opine catabolism and conjugal transfer of the nopaline Ti plasmid pTiC58 are coordinately regulated by a single repressor, *Proc. National Academy of Sciences of the United States of America*, Vol.89, pp.643–647 (1992).

[50] Katoh, K. and Standley, D.M.: MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability, *Molecular Biology and Evolution*, Vol.30, Issue 4, pp.772–780 (2013).

高橋 誉文　（正会員）

広島市立大学協力研究員．博士（情報工学）．2010 年広島市立大学大学院情報科学研究科博士前期課程修了．2015 年同博士後期課程修了．同年より現職．日本データベース学会会員．

北上 始　（正会員）

広島市立大学情報科学研究科教授．博士（工学）．1976 年東北大学大学院工学研究科博士前期課程修了．同年富士通株式会社，以後，富士通研究所，新生代コンピュータ技術開発機構（ICOT）主任研究員，国立遺伝学研究所客員助教授を経て，1994 年広島市立大学情報科学部教授，2007 年より現職．筆頭著書『データベースと知識発見』，『ビッグデータ時代のゲノミクス情報処理』．1985 年情報処理学会 25 周年記念論文賞，2003 年日本工学教育協会論文・論説賞ほか．日本データベース学会（論文誌編集委員），電子情報通信学会，人工知能学会，日本バイオインフォマティクス学会，IEEE，ACM，各会員．本会シニア会員．

福本 翔平

2013 年広島市立大学情報科学部卒業．2015 年同大学大学院情報科学研究科博士前期課程修了．同年株式会社エネルギア・コミュニケーションズに入社し，現在に至る．

森 康真　（正会員）

1994 年北陸先端科学技術大学院大学情報科学研究科博士前期課程修了．現在，広島市立大学大学院情報科学研究科助教．知識情報処理の研究に従事．IEEE CS，ACM，電子情報通信学会各会員．

田村 慶一　（正会員）

広島市立大学情報科学研究科准教授．博士（情報科学）．1998 年九州大学工学部情報工学科卒業．2000 年同大学大学院システム情報科学研究科知能システム学専攻修士課程修了．2003 年同大学院システム情報化学府知能システム学専攻博士後期課程単位取得のうえ満期退学．広島市立大学情報科学部助手，広島市立大学大学院情報科学研究科助教，同講師を経て，2011 年より現職．データマイニングとその並列処理に関する研究に従事．IEEE，電子情報通信学会，日本データベース学会，人工知能学会，日本知能情報ファジイ学会各会員．