Vol.2009-MPS-76 No.46
Vol.2009-BIO-19 No.46
2009/12/18

ギブスサンプリングとアラインメント処理に基づく類似部分配列の抽出方式

河野修久† 田村慶一† 森康真† 北上始†

配列データマイニング処理では，配列データベースから非常に多くの頻出配列パターンが抽出される．この頻出配列パターンを大幅に削減するために，本論文では，遺伝的アルゴリズムの世代交代モデルの 1 つである Minimal Generation Gap (MGG)と分散遺伝的アルゴリズム(島モデル)の考え方をギブスサンプリングに応用し，抽出する部分文字列長を自動決定することが可能である類似部分配列抽出法 Gibbs-DMGG を提案する．また，Gibbs-DMGG により抽出された類似部分配列集合に対しマルチプルアラインメントを実施し，その実施結果として得られる集合から類似部分配列を抽出する．評価実験により，Gibbs-DMGG が高い再現率を持つと同時にマルチプルアラインメントの実施が Gibbs-DMGG の適合率を 4〜5%程度向上させることを示す.

A Method for Extracting Similar Subsequences based on Gibbs Sampling and Multiple Alignment

NOBUHISA KONO† KEIICHI TAMURA†
YASUMA MORI† HAJIME KITAKAMI†

In the field of sequence data mining, a large quantity of frequent sequential patterns are extracted from a sequence database. In order to significantly reduce these frequent sequential patterns, we propose in this paper a new similar subsequence extraction method having the capacity to automatically determine the length of a similar subsequence called Gibbs-DMGG. This method applies Minimal Generation Gap (MGG), a generation alternation model of the genetic algorithm, and a distributed population scheme called the Island model. Moreover, we execute a multiple alignment for similar subsequence set extracted by Gibbs-DMGG. After that, we extract similar subsequences using Gibbs-DMGG from the set provided as the result of the multiple alignment. We show that the use of the multiple alignment not only improves Precision of Gibbs-DMGG around 5% but also keeps Gibbs-DMGG with a high Recall in an evaluation experiment.

1. はじめに

配列データベースから頻出なパターンを抽出する手法は，テキストデータに含まれる規則性抽出を初めとして, DNA やアミノ酸などの分子配列データのモチーフ抽出といった多くの問題解決に有用であるといわれている．モチーフとは，PROSITE[1]や Pfam[2]などで見られる生物学的に重要な機能を持つ特徴的なパターンである．自然界のモチーフには曖昧性が含まれるため，モチーフの表現手段として，正規表現や確率的表現が利用されている．テキストデータやアミノ酸配列データなどを含むデータベースに対する配列データマイニングでは，正規表現によるパターン抽出問題に焦点が当てられ，固定長や可変長のワイルドカード領域を持つ頻出配列パターンを抽出する手法の研究が進められてきた[3]．しかしながら，配列データベース全体から頻出パターンを抽出しようとすると，明らかに不要と思われるパターンが大量に抽出される．従って，著者らは，抽出される配列パターン数を削減する方法の研究に着目している[4].

ギブスサンプリングアルゴリズム GS は，配列パターン数を削減するための 1 つの方法である．GS は，配列データベースに含まれる各配列データから類似部分配列(類似部分文字列)を取り出す方法である．この方法を用いると，配列データマイニング処理に入力する配列データのサイズを小さくすることができる[4]. しかしながら, GS は，抽出する部分配列の長さをユーザ側で予め指定する必要があるほか，類似部分配列の抽出精度が安定しておらず, 必ずしも良いとは限らないという問題点を持っている[5].

本論文では，この問題点を解決するために，類似部分配列長の自動決定が可能でかつ類似部分配列の安定した精度を提供する新しい類似部分配列抽出法 Gibbs-DMGG[6]を提案する. 具体的には, 遺伝的アルゴリズムの世代交代モデルの 1 つである Minimal Generation Gap (MGG)[7]と, 分散遺伝的アルゴリズム(島モデル)[8][9]を用いて GS の最適化問題を解く．MGG は，世代間での個体分布の差異を最小化することが望ましいとの考えに基づき，探索序盤における選択圧をできるだけ下げて初期収束を回避するとともに，探索の後半においても集団内に多種多様な個体を生存させやすくして進化的停滞を抑制することを意図した優れた世代交代モデルとして知られている．しかしながら，MGG だけを利用するのでは，安定した精度が確保できないため，分散遺伝的アルゴリズムを用いて，安定した精度の確保を行う.

また，Gibbs-DMGG により抽出された類似部分配列集合に対してマルチプルアラインメント処理を実施する．さらに，この実施結果として得られるギャップ記号を含む新たな配列集合から類似部分配列集合を抽出することにより，Gibbs-DMGG に対するアラインメント処理の影響について考察する.

† 広島市立大学情報科学研究科
Graduate School of Information Science, Hiroshima City University

Vol.2009-MPS-76 No.46
Vol.2009-BIO-19 No.46
2009/12/18

以下，本稿の構成を示す．2 章では類似部分配列抽出に関する関連研究について述べる．3 章は従来手法であるギブスサンプリング法について，4 章では提案手法 Gibbs-DMGG について述べる．5 章では実験の処理手順について説明し，6 章では提案手法の評価を行い，7 章ではまとめと今後の課題を述べる．

2. 関連研究

配列データベースから類似部分配列を抽出する方法には，Lawrence らが提案したギブスサンプリング[10][11]がある．初期のギブスサンプリングは 1 本の配列からは 1 つのモチーフしか抽出できなかったが，Liu らは，複数のモチーフ抽出に対応させた手法として，Greedy Two-stage Gibbs Sampling[12]を提案している．また，抽出される類似部分配列の精度向上のため，最適化手法であるシミュレーテッドアニーリングの一種である Simulated Tempering を用いた GibbsDST[13]を初めとして，遺伝的アルゴリズムの世代交代モデルの 1 つである MGG[7]を用いた手法[5]などが提案されている．

以上の手法はいずれも抽出する類似部分配列の長さを予め指定しなければならないという問題や，安定した抽出精度が確保できないという問題があった．

本稿においては，MGG を用いたギブスサンプリングを改良し，さらに分散遺伝的アルゴリズム[8][9]の考え方を導入することにより，抽出される部分配列長を自動決定すると同時に安定的な抽出精度を確保する方法 Gibbs-DMGG を提案している．

3. 従来の抽出手法

本章では，従来のギブスサンプリング[10][11]を用いた類似部分配列抽出法と，抽出した類似部分配列(*k*-部分配列集合)の評価法について説明する．*k*-部分配列とは，ギブスサンプリング処理で，配列データベースの各配列から取り出される指定長 *k* の部分文字列をさす．また，各配列から抽出される *k*-部分配列の集合を，*k*-部分配列集合と呼ぶ．配列データベースと *k*-部分配列集合の関係を図 1 に表す．

3.1 ギブスサンプリング

配列データベース *DB* の各配列が *n* 個の文字から成る文字集合$\sum=\{\alpha_1, \alpha_2, \ldots, \alpha_n\}$上で定義されているとする．また，*DB* は *t* 本の配列から成るとする．ギブスサンプリングは，配列データベース *DB* の各配列から，ユーザが予め定めた指定長 *k* を持つと同時に互いにできるだけ類似した部分配列を見つけ出す方法である．ギブスサンプリングでは，できるだけ類似した *k*-部分配列集合を見つけ出すために，解候補としての *k*-部分配列を評価する尺度が必要になる．この尺度を計算するためには，スコア行列，出現頻度，背景頻度の 3 種類の統計的確率量が採用されている．ある *k*-部分配列集合に対して，これら 3 種類の統計的確率量は以下のように定められている．

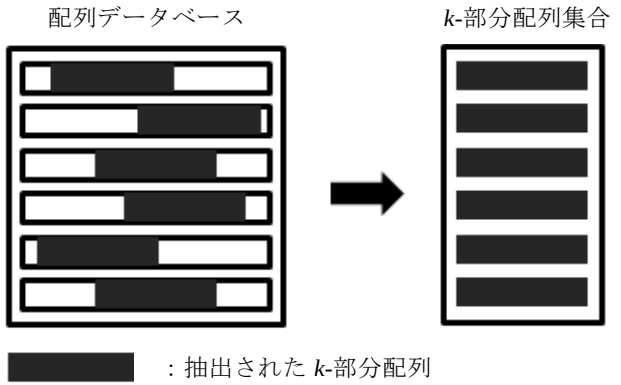

図 1　配列データベースと *k*-部分配列集合

(1) スコア行列
k-部分配列集合のスコア行列 $A = (A_{i,j})$はプロファイルとも呼ばれ，行列要素 $A_{i,j}$ は，*k*-部分配列集合を用いて，文字 α_j が位置 *i* に出現する頻度を計算することにより定められる．

(2) 出現頻度
k-部分配列 $x=<\alpha_1\alpha_2\ldots\alpha_k>$の出現頻度 A_x は，$A_{1,1}\times A_{2,2}\times\ldots\times A_{k,k}$ を計算することにより定める．これは，*k*-部分配列 *x* の確率が高ければ，*x* はスコア行列の定義に利用された *k*-部分配列集合の総意に類似し，低ければ類似しないことを意味する．

(3) 背景頻度
DB から解候補のすべての部位を除去した部分配列の集合を *BS* とするとき，文字 α_j の背景頻度は，*BS* に存在する文字総数に対する文字 α_j の出現総数 P_{α_j} と定める．*k*-部分配列 $x = <\alpha_1\alpha_2\ldots\alpha_k>$の背景頻度 P_x は，$P_{\alpha_1}\times P_{\alpha_2}\times\ldots\times P_{\alpha_k}$ を計算することにより定める．

ギブスサンプリングは，*DB* からランダムに選択された配列 *Z* から出現頻度が高くかつ背景頻度の低い *k*-部分文字列を候補解として見つけ出す処理を基本にしており，図 2 に示すアルゴリズムとして知られている．

Vol.2009-MPS-76 No.46
Vol.2009-BIO-19 No.46
2009/12/18

1. t 本の配列をもつ DB 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，それらを配列順に並べた k-部分配列集合を $S=\{s_1,s_2,\ldots,s_t\}$とする．
2. DB からランダムに 1 つの配列 Z を選択する．$t-1$ 本の配列から成る配列データベースを $DB'=DB-\{Z\}$とする．また，上記 1 で選択された開始点 st をもとに配列 Z から選択されている k-部分配列を Z'とし，$S'=S-\{Z'\}$とする．
3. $t-1$本の配列から成る k-部分配列集合 S'からスコア行列 $A=(A_{i,j})$を計算する．
4. DB' から S'を削除した集合 BS'を用いて，各文字 α_j の背景頻度 $P\alpha_j$ を計算する．
5. 配列 Z 上で開始位置 i に存在する$|Z|-k+1$ 個の k-部分配列 x のそれぞれに対して，評価値 $U_x=A_x\div P_x$ を計算する．
6. $\{U_1, U_2,\ldots, U_{|Z|-k+1}\}$の各評価値の中から，評価値に比例した確率でランダムに U_m を選択し，U_m に対応する k-部分配列の配列上の新しい開始点 st'_m を選び，S を更新する（$1\le m\le|Z|-k+1$）．
7. 収束するまで 2～6 を繰り返す．

図 2　ギブスサンプリングアルゴリズム

3.2 部分配列の評価

配列データベース DB から抽出された類似部分配列を評価するために，F 値と呼ばれる相対エントロピーを用いている．相対エントロピーは，DB から抽出された類似 k-部分配列の部分の分布と，DB から抽出されないで残った非類似部分の分布の差により計算される．従って，F 値の計算は以下のとおりである．

$$F=\sum_{i=1}^{k}\sum_{j=1}^{20}C_{i,j}\log\frac{Q_{i,j}}{P_j} \qquad (1)$$

P_j は，文字 α_j の背景頻度 P_j と同様のものである．$Q_{i,j}$ について式(2)に示す．

$$Q_{i,j}=\frac{C_{i,j}+b_j}{N-1+B} \qquad (2)$$

$C_{i,j}$ は，DB から抽出された類似 k-部分配列の部分に対するスコア行列要素であり，位置 i に存在する文字 j の個数を意味する．b_j は擬似度数であり，$C_{i,j}$ が 0 となるときに，$Q_{i,j}$ が 0 となることを防ぐために用いられ，各 b_j は $f_i\times B$ により決定される．ここで，f_i は，文字 α_j の全配列に対する相対出現頻度により決定される．また，N を配列数とすると，B は経験的に $\sqrt{N}$ としてよいことがわかっている．

4. 新しい類似部分配列抽出法

本章では，部分配列長 k の自動決定及び，類似部分配列の安定的な抽出精度確保のため，ギブスサンプリングの最適化手法として MGG と島モデルを用いた新しい類似部分配列抽出法 Gibbs-DMGG を提案する．そのために，先ず，MGG と島モデルの 2 つの最適化手法について紹介した後，MGG を用いたギブスサンプリング手法(以後，Gibbs-MGG と呼ぶ)について述べる．そして，部分配列長 k の自動決定をするために，Gibbs-MGG を改良する方法を提案する．以後，この改良された Gibbs-MGG を改良版 Gibbs-MGG と呼ぶ．さらに，安定した抽出精度を確保するために，改良版 Gibbs-MGG に島モデルを導入することにより，Gibbs-DMGG を提案する．

4.1 最適化手法

本節では，ギブスサンプリングの最適化手法として利用されている MGG と島モデルについて紹介する．

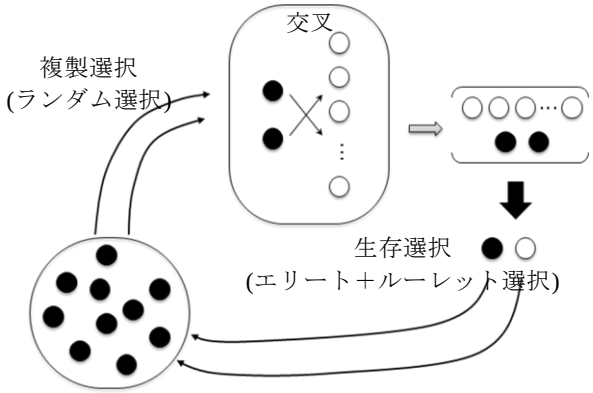

図 3　MGG の流れ

4.1.1 MGG

最適化問題の多くに有効な手法の一つとして，GA がある．GA では，適応度の高い個体（データ）を優先的に選択して交叉や突然変異を行うことで最適解を求める世代交代モデルであり，最適化問題だけでなく，NP 困難な問題など，様々な問題で使用

Vol.2009-MPS-76 No.46
Vol.2009-BIO-19 No.46
2009/12/18

される．しかしながら，探索の初期段階において，他の個体に比べて極端に適応度の大きい個体が生成されたとき，その個体が爆発的に増え，探索がかなり早い段階で収束し，局所解に陥ってしまう可能性がある．MGG は，図 3 のように，母集団として個体を複数生成し，複製選択，交叉操作，突然変異操作，生存選択を繰り返す GA の世代交代モデルの 1 つである．MGG では，世代交代の対象としてランダムに選ばれた 2 個体を用いるため，解の精度低下の原因となる初期収束が起こりにくい．また，生存選択時には最良(エリート)選択とルーレット選択を組み合わせることにより，適合度の分散をできるだけ維持できるようにして進化的停滞を抑制できるようにしている．

4.1.2 島モデル

島モデルは，GA の分散モデルの 1 つであり，個体の母集団を複数のサブ母集団(島)に分割し，島ごとに独立して遺伝的操作を行う．さらに，数世代に一度，各島内で選ばれた 1 つまたは複数個の個体を別の島の個体と交換する移住という操作を行う．島モデル固有のパラメータとして，移住を行う世代間隔を定める移住間隔と，移住する個体の割合を決定する移住率が存在する．図 4 に島モデルの概念を示す．

島モデルでは，それぞれの島で独立に探索が進むため，各島で個体は大きく異なり，各島が独自の領域を探索することが可能となる．このため単一母集団と比較して多様性が大きくなるので，安定的な精度向上を期待できる．

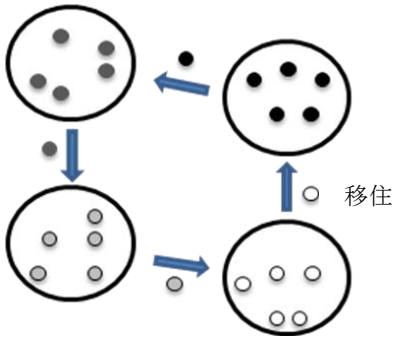

図 4　島モデルの概念

4.2 Gibbs-MGG

Gibbs-MGG[5]は，GS が持つ抽出精度の問題を改善するために提案された手法である．MGG で使用する個体は *k*-部分配列集合により定義される．個体の適応度には式(1)の *F* 値を用いる．*N* 個用意される個体に対して，複製選択操作，交叉操作，生存選択操作，突然変異操作を繰り返すことにより計算が進められる．

交叉操作では，ユーザが予め指定した生成個体数 *M* に対して，*M* 個の交叉点をランダムに決め，複製選択で選ばれた 2 個体から *M* 個の子個体を生成する．通常の複製選択操作では母集団から同じ個体が 2 回選ばれることがあるが，Gibb-MGG では，同じ個体が 2 回選ばれることがないよう変更を加えている． 2 つの個体から作成される子個体の例を図 5 に示す．

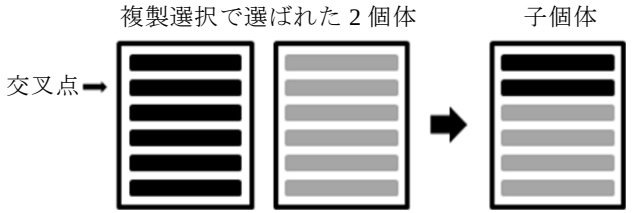

図 5　交叉操作による子個体の生成方法

生存選択操作では，複製選択で選ばれた 2 個体と作成した *M* 個の子個体から，エリート個体と，ルーレット選択で選ばれた個体を母集団に戻す．生存選択終了後に突然変異処理を行う．

突然変異操作では，ユーザが選択した *L* 個の個体を母集団からランダムに選択し，GS 処理における繰り返し処理を 1 回だけ行う．つまり，選ばれた *L* 個の個体それぞれに対して, 図 2 に示したギブスサンプリングアルゴリズムの 2~6 の操作が行われる．Gibbs-MGG のアルゴリズムを図 6 に示す．

1. 初期状態母集団として，*N* 個の個体(*k*-部分配列集合)をランダムに生成する．
2. 複製選択操作 : 母集団の中から 2 つの個体をランダムに選択する．
3. 交叉操作 : 選ばれた 2 個体で交叉を行い，*M* 個の子個体を作成する．
4. 生存選択操作 : 作成されたすべての子個体と，母集団から選ばれた 2 個体の *F* 値を計算し，最良個体 1 個体と確率的に選んだ 1 個体を母集団に戻す．
5. 突然変異操作 : 母集団の中から *L* 個の個体をランダムに選択し，突然変異を行う．
6. 指定した回数 2〜5 を繰り返す．

図 6　Gibbs-MGG のアルゴリズム

Vol.2009-MPS-76 No.46
Vol.2009-BIO-19 No.46
2009/12/18

4.3 提案手法

ギブスサンプリングでは抽出する部分配列長 k をユーザ側で指定する必要がある．部分配列長を自動決定するために，前節で説明した Gibbs-MGG を改良し，さらに島モデルの考え方を用いた新たな手法を提案する．

Gibbs-MGG では，母集団内に存在する個体の部分配列長は全て同じものとなっており，処理中に変化することはない．これに対して，改良版 Gibbs-MGG では，母集団内に様々な部分配列長の個体が存在できるようにし，交叉操作と突然変異操作を一部変更している．

交叉操作では，複製選択で選ばれた 2 個体の部分配列長が異なる場合，図 5 のように単純に 2 個体の一部を入れ替えると作成された子個体の部分配列長が途中から変化してしまう．そこで，図 5 の交叉操作を行う前に，選ばれた 2 個体の部分配列長を，どちらか一方の長さに統一する．このとき，2 個体のうちどちらの部分配列長を採用するかはランダムに決定する．

突然変異操作の終了後に，ランダムに選ばれた個体に対して部分配列長の変更操作を追加する．部分配列長の変更操作を図 7 に示す．部分配列長変更時に大幅に長さを増加減少させると，部分配列長が一定の値に収束しづらくなるのではないかと考えられる．そこで，部分配列長の変更は，変更前の部分配列長を k とするとき，変更後の部分配列長が元の長さの最大 3/2，最小 1/2 までの範囲で変わるように，$-k/4 \sim k/4$ の範囲で乱数をとり，部分配列の開始位置及び終了位置をそれぞれ変更する．

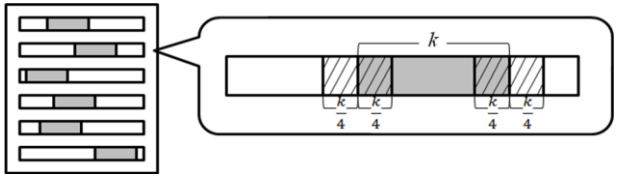

図 7　配列長の変更操作

図 7 の例では，部分配列の開始位置及び終了位置はそれぞれ斜線内のいずれかの部分に決定される．部分配列長の変更は，突然変異操作の中で毎回行うわけではない．部分配列長の変更は，配列データベース内に含まれる配列数 *num* を参考に，数世代ごとに行う．

部分配列長を自動決定する場合，部分配列長の初期値が非常に重要となる．初期値が大きすぎる場合，部分配列長が初期値以下にならず，逆に小さすぎる値の場合，極端に部分配列の抽出精度が悪くなる．この問題を解決するために，改良版 Gibbs-MGG に島モデルの考え方を用いる．Gibbs-DMGG では，母集団を複数用意し，各母集団で部分配列長の初期値を異なるものにすることで，抽出精度の悪化を防ぐ．各母集団は基本的に独立して処理が行われるが，数世代に一度，移住操作と呼ばれる母集団同士での個体移動を行う． Gibbs-DMGG のアルゴリズムを図 8 に示す．

1　初期状態の母集団群として，N 個の個体を含む I 個の島(母集団)をランダムに生成する．
2　以下の作業をそれぞれの母集団で行う．
　2.1　複製選択操作：母集団の中から 2 つの個体をランダムに選択する．
　2.2　交叉操作：選ばれた 2 個体をどちらかの長さに合わせ交叉を行い，M 個の子個体を生成する．
　2.3　生存選択操作：生成された全ての子個体と，母集団から選ばれた 2 個体の F 値を計算し，最良個体と確率的に選んだ 1 個体を母集団に戻す．
　2.4　突然変異操作：母集団の中から L 個の個体をランダムに選択し，GS 処理を行う．
　2.5　配列長の変更操作：数世代に 1 回，部分配列長の変更操作を行う．
3　移住操作：数世代に 1 回移住処理を行う．
4　指定した回数 2,3 を繰り返す．

図 8　Gibbs-DMGG のアルゴリズム

5. 処理手順

本章では実験の処理手順について述べる．まず，PROSITE から得た配列データに対し，Gibbs-DMGG を適用して類似部分配列を抽出する．次に，抽出された類似部分配列集合に対しマルチプルアラインメント処理を行う．最後に，マルチプルアラインメント処理を行った配列集合を入力データとして Gibbs-DMGG を適用し，新たに類似部分配列を抽出する．

Gibbs-DMGG を適用し類似部分配列を抽出する際は，いずれも抽出する配列長は指定せずに処理を行い，配列長を自動決定しながら類似部分配列集合を抽出する．

6. 性能評価

本章では, 提案手法の評価を行う. このために利用した計算機環境は, Intel® Core™2 Quad CPU@2.66GHz，メモリ 2.0GB：，SWAP メモリ 2.0GB：，HDD：227GB，

Vol.2009-MPS-76 No.46
Vol.2009-BIO-19 No.46
2009/12/18

OS:Fedora9(Sulphur)である. 性能評価のために使用した配列データベースは, PROSITE に含まれる Leucine Zipper データセット(登録番号 PS00036)を用いた．Leucine Zipper モチーフの形式は，**<[*KR*]-x(1,3)-[*RKSA Q*]-*N*-*x*(2)-[*SAQ*](2)-*x*-[*RKTAENQ*]-*x*-*R*-*x*-[*RK*]>** であり，最大 16 文字で形成される．データセットの詳細を表 1 に示す.

表 1　Leucine Zipper データセット

配列数	最大長	最小長	総長
188	1383	125	73673

6.1　評価の指標

ここでは，Gibbs-DMGG の抽出精度を評価するために，評価の指標となる再現率(recall)と適合率(precision)について説明する.

配列データベースに含まれる配列データの数を n としよう．A_i を識別子 *sid* の値が i の配列データに現れるモチーフ領域とするとき，配列データベース内に存在するモチーフ領域の集合を $A = \{A_1, A_2, \ldots, A_n\}$ と表現する$(1 \leqq i \leqq n)$. Gibbs-DMGG により識別子 *sid* の値が i の配列データから抽出される領域を B_i とするとき, 配列データベースから抽出される領域の集合を $B = \{B_1, B_2, \ldots, B_n\}$ と表現する$(1 \leqq i \leqq n)$. $|A_i|$を領域 A_i 内に存在する文字数とし，$|A|$を$\sum|A_i| \cdot [1 \leqq i \leqq n]$と定義する．さらに，$C_i = A_i \cap B_i$ を領域 A_i と領域 B_i との共通領域とし, $A \cap B$ を$\{C_1, C_2, \ldots, C_n\}$と定義する. このとき, 再現率を$|C|/|A|$, 適合率を$|C|/|B|$と定義する.

以下では, 例として表 2 を用いて再現率と適合率を求める. ただし, 部分配列長 k=4, モチーフを<*ATF*>とし，数値は小数第 2 位を四捨五入したものである．表中で表記されている部分配列は抽出された部分配列を意味する.

表 2　配列データベースと部分配列

sid	配列データ	部分配列
1	TATKFATFKT	ATFK
2	KATFAFTFAF	FAFT
3	AAKAKATFTK	AKAK
4	FAKATATFAA	ATFA
5	AATFTKFTTF	AATF

モチーフ領域の文字総数$|A|$は$|A_1|+|A_2|+\ldots+|A_5|$と計算することができ，$|A|$=15 となる．抽出領域の文字総数$|B|$も同様に計算できるので，k=4 より$|B|$=20 となる．*sid*=1 の部分配列にはモチーフがすべて(3 文字)含まれているので，$|C_1|$=3 となる．*sid*=2 の配列の部分配列には，モチーフ<*ATF*>のうち“*F*”しか含まれていないので$|C_2|$=1 となる．以下同様に，$|C_3|$=0，$|C_4|$=3，$|C_5|$=3 となるので，$|C|$=10 と計算できる．よって，この部分配列の再現率と適合率は，再現率=10/15=66.7%，適合率=10/20=50.0%と求められる.

6.2　評価実験

提案手法 Gibbs-DMGG の性能評価を行うために, パラメータを変更しながら実験を行った．表 3 は，母集団内の個体数を 5，作成する子個体数 5，突然変異数 1，繰り返し回数(世代数)4800 に固定し，作成する島の数を 6，10，15，30 と変化させ，それぞれ 10 回試行した時の抽出した部分配列長，平均再現率，平均適合率を示している．表 3 より，いずれの島数においても，平均再現率が 99%以上の値をとっており，各試行での再現率も安定して高く, 抽出精度の面においてはかなり良い手法であると言える．図 9 より，ギブスサンプリングのみを用いた場合での再現率はおよそ 85%程度であることが分かる．また，Gibbs-DMGG により抽出された類似部分配列長は約 70 であり，ギブスサンプリングを用いてユーザが抽出配列長を指定した場合に再現率が最大となる長さとほぼ同等である．これより，本手法で抽出される類似部分配列長は適切であると考えられる.

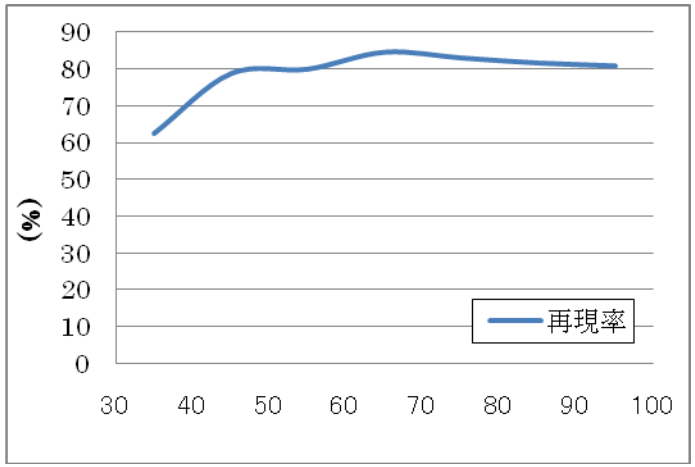

図 9　ギブスサンプリングの抽出性能

続いて，アラインメント処理の影響を調べるため，Gibbs-DMGG により抽出された類似部分配列集合に対し解析ツール ClustalX[13]を用いてマルチプルアラインメント処理を行い，得られた配列集合を入力データとして新たに類似部分配列抽出を行った.

Vol.2009-MPS-76 No.46
Vol.2009-BIO-19 No.46
2009/12/18

表 3　島の数を変化させたときの抽出性能比較

島の数	自動決定された文字列長の出現回数					平均長	再現率	適合率	世代数
	54～60	61～70	71～80	81～90	91～114				
6	1	2	4	2	1	74.7	99.33	21.19	4800
10	0	3	5	1	1	76.2	99.29	20.67	4800
15	1	4	4	1	0	69.4	99.28	22.62	4800
30	0	1	8	0	1	75.6	99.33	20.62	4800

表 4　アラインメント処理の影響

入力データの配列長	自動決定された文字列長の出現回数			平均長	再現率 (%)	適合率 (%)
	51~55	56~60	61~65			
69	2	8	0	57.8	99.00	26.86
72	1	5	4	59.0	98.50	26.16
74	0	3	7	61.8	98.50	24.97

マルチプルアラインメント処理を行うことにより，配列内の挿入や削除を考慮することができるので，抽出性能が向上するのではないかと考えられる．表 4 に，前述の実験で得られたデータの中で，任意に選んだ配列長 69，72，74 の類似部分配列集合を用いてそれぞれ 10 回実験を試行した時の抽出した配列長，平均再現率，平均適合率を示す．

表 4 より，どの入力データでもアラインメント処理を適用することで配列長が 10 程度短くなり，表 1 での結果よりもおよそ 5%適合率を改善させられることが分かった．これは，アラインメント処理により配列データに文字の挿入・欠失が考慮され，モチーフ部分が整列化された効果であると考えられる．

7. まとめ

本論文では，ギブスサンプリング法に MGG と島モデルを用いた新しい類似部分配列抽出法 Gibbs-DMGG を提案し，実装，実験を行った．また，Gibbs-DMGG に対するアラインメント処理の影響について検討した．その結果，Gibbs-DMGG は抽出する部分配列の長さを指定することなく，元の配列データベースの 5 分の 1 程度の類似部分配列を再現率 99%以上で抽出することに成功し，さらにアラインメント処理を導入することにより 5%程度適合率を上昇させることができた．

しかし，必要な部分であるモチーフ長からみると，抽出した類似部分配列はまだ充分な長さであるとは言えないので，さらなるモチーフ長に近い長さの類似部分配列の抽出が今後の課題としてあげられる．

謝辞　本研究の一部は，日本学術振興会，科学研究費補助金(基盤研究(C)，課題番号：20500137)の支援により行われた．

参考文献

1)　PROSITE. http:// kr.expasy.org/prosite.
2)　Pfam.　http://www.sanger.ac.uk/Software/Pfam.
3)　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, 電子情報通信学会論文誌 D「データ工学特集号」, Vol.J90-D, No.2, pp.281-291，2007 年 2 月.
4)　加藤 智之，森 康真，荒木 康太郎，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 年 6 月.
5)　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 (DEWS 2008)論文集, E7-2, Online Proceedings，2008.
6) Nobuhisa Kono, Hajime Kitakami, Keiichi Tamura, and Yasuma Mori:Extracting Similar Subsequences by Gibbs Sampling with Distributed MGG, Proceedings of the 2009 International Conference on Parallel & Distributed Processing Techniques & Applications (PDPTA'09), pp.669-675, Las Vegas in USA, July 13-16 in 2009.
7)　佐藤 博, 小野 功, 小林 重信 : 遺伝的アルゴリズムにおける世代交代モデルの提案と評価, 人工知能学会誌,Vol.12,No.5,pp.734-743,1997.

Vol.2009-MPS-76 No.46
Vol.2009-BIO-19 No.46
2009/12/18

8) Erick Cantu-Paz: Efficient and Accurate Parallel Genetic Algorithms, Springer, 2000.
9) 廣安知之, 三木光範, 上浦二郎: 実験計画法を用いた分散遺伝的アルゴリズムのパラメータ推定, 情報処理学会論文誌：数理モデル化と応用, Vol.43, No.SIG10(TOM7), 199-217, 2002.
10) Lawrence C.E., Altschul,S.F., Boguski,M.S., Liu,J.S., Neuwald,A.N. and Wotton,J.: Detecting subtle sequence signals: A Gibbs Sampling Strategy for Multiple Alignment, Science,263,208-214, 1993.
11) Liu,J.S., Neuwald, A.N. and Lawrence,C.E.: Bayesian Models for Multiple Local Sequence Alignment and Gibbs Sampling Strategies, JASA, 90, 1156-1170, 1995.
12) Li-fang Liu, Li-cheng Jiao, Hong-wei Huo: A Greedy Two-stage Gibbs Sampling Method for Motif Discovery in Biological Sequences, 2008 International Conference on BioMedical Engineering and Informatics, Vol.1, 13-17, IEEE Computer Society Press, 2008.
13) Kazuhito Shida: Hybrid Gibbs-Sampling Algorithm for Challenging Motif Discovery: GibbsDST, Genome Informatics, Vol.17, No.2, 3-13, 2006.
14) 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: ClustalW and Clustal X version 2.0, Bioinformatics, 23, 2947-2948, 2007.