

分散 MGG を用いたギブスサンプリングによる類似部分配列抽出法

河野 修久[†] 田村 慶一[†] 森 康真[†] 北上 始[†]

[†] 広島市立大学大学院情報科学研究科 〒731-3194 広島市安佐南区大塚東 3-4-1

E-mail: [†] kono@db.its.hiroshima-cu.ac.jp, {ktamura, mori, kitakami}@hiroshima-cu.ac.jp

あらまし 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、著者らは、既に、配列データベースに対してギブスサンプリング(GS)を適用する方法を提案している。しかしながら、この方法では、予め、抽出する部分文字列の長さをユーザ側で指定する必要があるほか、必ずしも精度(再現度)が良いとは限らないという問題がある。本稿では、これら 2 つの問題を解決するために、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 GS に応用した新しい類似部分配列抽出法 Gibbs-DMGG を提案する。また、この提案手法の有効性を確認するために、その方法に対する性能の評価・考察を行ったので、その結果についても報告する。

キーワード データマイニング, テキストマイニング, バイオインフォマティクス, 情報抽出

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、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 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)^[6]と、分散遺伝的アルゴリズム(島モデル)^{[7][8]}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。

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 章では提案手法について述べる。5 章では提案手法の評価を行い、6 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^{[9][10]}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^[11]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^[12]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^[6]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がある。また、最適化のための多くのパラメータ値を設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^{[7][8]}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^{[9][10]}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。

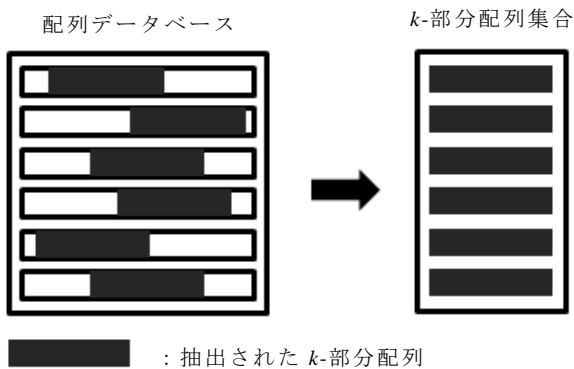


図1: 配列データベースと k -部分配列集合

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 $\Sigma=\{a_1, a_2, \dots, a_n\}$ 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、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。

- (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 a_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 $x=\langle a_1 a_2 \dots a_k \rangle$ の出現頻度 A_x は、 $A_{1,1} \times A_{2,2} \times \dots \times A_{k,k}$ 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 a_j の背景頻度は、 BS に存在する文字総数に対する文字 a_j の出現総数 P_{a_j} と定める。 k -部分配列 $x=\langle a_1 a_2 \dots a_k \rangle$ の背景頻度 P_x は、 $P_{a_1} \times P_{a_2} \times \dots \times P_{a_k}$ 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にアルゴリズムとして知られている。

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 s_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 $S=\{s_1, s_2, \dots, s_t\}$ 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 s_i 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 a_j の背景頻度 P_{a_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. $\{U_1, U_2, \dots, U_{|Z|-k+1}\}$ 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 s'_m を選び、 S を更新する($1 \leq m \leq |Z|-k+1$)。
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。

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

P_j は、文字 a_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。

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

$C_{i,j}$ 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 $f_j \times B$ により決定される。ここで、 f_j は、文字 a_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、まず、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。

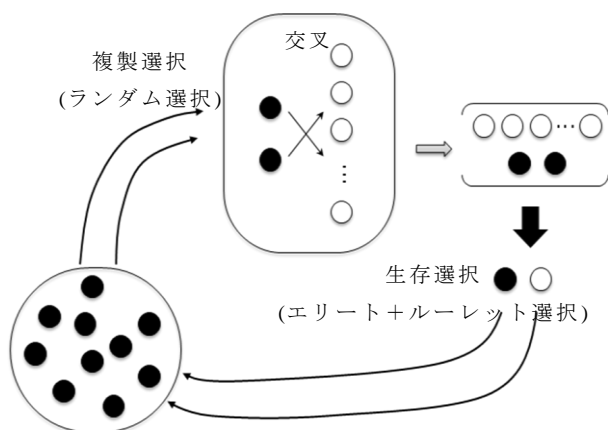


図 3 : MGG の流れ

4.1.1. MGG

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。

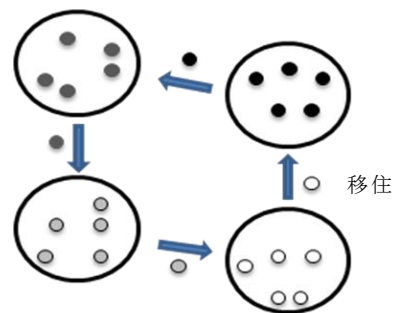


図 4 : 島モデルの概念

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。

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、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、Gibbs-MGG では、同じ個体が 2 回選

ばれることがないように変更を加えている。2つの個体から作成される子個体の例を図5に示す。

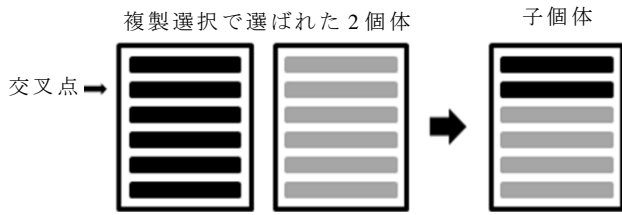


図5：交叉操作による子個体の生成方法

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

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う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、 $-k/4 \sim k/4$ の範囲で乱数を取り、部分配列の開始位置及び終了位置をそれぞれ変更する。

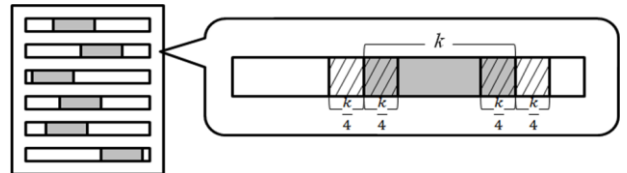


図7：配列長の変更操作

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、配列データベースに含まれる配列数 num を参考に、数世代ごとに行う。

- 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

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。

5. 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、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)-[RKSAQ]-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

5.1. 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、配列データベース内に存在するモチーフ領域の集合を $A = \{A_1, A_2, \dots, A_n\}$ と表現する ($1 \leq i \leq n$)。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 $B = \{B_1, B_2, \dots, B_n\}$ と表現する ($1 \leq i \leq n$)。 $|A_i|$ を領域 A_i 内に存在する文字数とし、 $|A|$ を $\sum |A_i| \cdot [1 \leq i \leq n]$ 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 $\{C_1, C_2, \dots, C_n\}$ 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。

モチーフ領域の文字総数 $|A|$ は $|A_1| + |A_2| + \dots + |A_5|$ 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。

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

5.2. 評価実験

提案手法の性能評価を行うために、各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% 以上となっており、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と言える。また、各島数での適合率をみると、島数 15 における結果が最もよい値となった。

続いて、表 3 において最も良い結果となった島数 15 の場合を用いて、母集団内の個体数を変化させた場合での抽出結果の差を調べた。島数、個体数、繰り返し回数以外のパラメータは表 3 の実験で用いたパラメータを使用し、母集団内の個体数は 5、10、20 と変化させてそれぞれ 10 回試行した。繰り返し回数は個体数の変化に比例して増加させた。このときの部分配列長、平均再現率、平均適合率を表 4 に示す。

表 4 より、個体数の増加が適合率の向上に比例しないことが分かった。しかし、部分文字列長の出現回数分布から、個体数を増加させると出現する部分文字列長の範囲が小さくなることが分かった。

これらより、島数 15、個体数 5 のとき、最も良い精度で類似部分配列を抽出可能となることが分かった。このとき、抽出された部分配列長はおおよそ 70、再現率 99% 以上となる。Leucine Zipper データセットにおいて、モチーフ長は 14~16、平均配列長は表 1 より約 392 であるので、元の配列データベースの 5 分の 1 以下、モチーフ長の 5 倍程度の類似部分配列となっている。元の配列データベースから考えると、かなりの不要部分の削除に成功しているといえる。また、再現率におい

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					再現率	適合率	CPU 時間 (sec)	世代数
	54~60	61~70	71~80	81~90	91~114				
6	1	2	4	2	1	99.33	21.19	600.0	4800
10	0	3	5	1	1	99.29	20.67	985.8	4800
15	1	4	4	1	0	99.28	22.62	1447.7	4800
30	0	1	8	0	1	99.33	20.62	2670.6	4800

表 4：母集団内の個体数を変化させたときの抽出性能比較

個体数	自動決定された文字列長の出現回数				再現率	適合率	CPU 時間 (sec)	世代数
	54~60	61~70	71~80	81~83				
5	1	4	4	1	99.28	22.62	1447.7	4800
10	0	5	5	0	99.26	21.64	3119.6	9600
20	0	1	8	1	99.12	20.93	6870.0	19200

でも、従来手法を用いた場合およそ 82%程度であるので、提案手法は効果的な類似部分配列抽出法ではないかといえる。

6. 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。その結果、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 5 倍程度の長さのとなっているので、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] 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.
- [7] Erick Cantu-Paz: Efficient and Accurate Parallel Genetic Algorithms, Springer, 2000.
- [8] 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.
- [9] 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.
- [10] 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.
- [11] 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.
- [12] Kazuhito Shida: Hybrid Gibbs-Sampling Algorithm for Challenging Motif Discovery: GibbsDST, Genome Informatics, Vol.17, No.2, 3-13, 2006.