高速モチーフ探索回路の提案

佐藤　由香†　　田沢　純子†　　宮崎　敏明†

会津大学†　〒965-8580 福島県会津若松市一箕町鶴賀

E-mail: †{m5111110, s1140139, miyazaki}@u-aizu.ac.jp

1　はじめに

遺伝子配列モチーフとは，塩基で構成された複数の遺伝子配列に現れる局所的な類似部分配列のことである．モチーフを発見することは，機能新規予測やファミリーの同定をすることができるため，バイオインフォマティクス分野で強く求められ，幾つかの手法が提案されている．本稿では，その手法の一つとしてよく知られたGibbs Sampling 法のハードウェアによるモチーフ探索の高速化を目指し，モチーフ探索回路のアーキテクチャを提案する．また，提案回路を FPGA に実装し，その実現可能を確認する．

現在，モチーフを探索する手法が幾つか提案されている．モチーフが登録されているデータベースを用いる探索手法には FASTA[1]や BLAST[2]，ヒューリスティックなアルゴリズムには Gibbs Sampling 法[3]や Metropolis-Hasting 法[4]がよく知られている．Gibbs Sampling 法を用いてモチーフ探索を行うツールには AlignAce[5]がよく知られている．上述した手法は，AlignAce のように汎用計算機を用いたソフトウェアとして実現されているが，計算量が莫大なため，より高速なモチーフ探索法が求められている．ここでは，Gibbs Sampling 法をハードウェア実装し，モチーフ探索の高速化を目指す．

2　Gibbs Sampling 法

Gibbs Sampling 法は Markov Chain Monte Carlo 法の解法のひとつである．モチーフ探索に Gibbs Sampling 法を適用すると，いくつかの塩基配列の各配列から指定した長さ h の類似した部分配列を，アライメントを行わないで求めることができる．手順を以下に示す．

1) N本の塩基配列のモチーフ候補を決定
2) N本の塩基配列から配列Mを1本選択
3) 配列M以外のモチーフ候補の出現確率 A_K を計算
4) 各配列のモチーフ候補に含まれていない部分から，各塩基 "A", "T", "G", "C"の出現確率（背景出現確率 P_j）を計算
5) 配列Mの全ての部分配列の評価値 U_i を計算
6) 5) で求めた評価値 U_i からランダムに U_K を選択
7) 配列Mのモチーフ候補を更新．ただし，U_K はできるだけ大きな値から選ばれるとする．
8) 収束するまで，上記した 2) ～ 7) を繰り返す．

A high-speed circuit for motif detection

Yuka Sato†, Junko Tazawa†, Toshiaki Miyazaki†
†University of Aizu　Tsuruga, Ikki-machi, Aizu-wakamatsu City, Fukushima, 965-8580 Japan
{m5111110, s1140139, miyazaki}@u-aizu.ac.jp }

3　評価値の吟味

Gibbs Sampling 法では評価値を求める際に，出現頻度 A_1 と背景出現頻度 P_1 を用いて $U_1 = A_1 / P_1$ とする．これは，塩基配列の構成要素である "A", "T", "G", "C" のうち全体の出現頻度が高いだけのものが検出されないよう，モチーフ候補以外の背景出現頻度で除算することで正規化している．しかし，除算は FPGA 実装において非常に実装コストが高く，面積も実行ステップも多くなってしまう．この除算の工程を省略することができれば，それらを削減することができる．

そこで，従来から行われている背景出現頻度による正規化がどのような影響を与えているか実験を行った．具体的には，いくつかの塩基配列のすべての部分配列に対して，2つの評価値（背景出現頻度による正規化を行うものと，出現頻度のみを用いるもの）を計算し，評価値が大きくなる部分配列の開始位置を比較した．結果の一例を図1に示す．

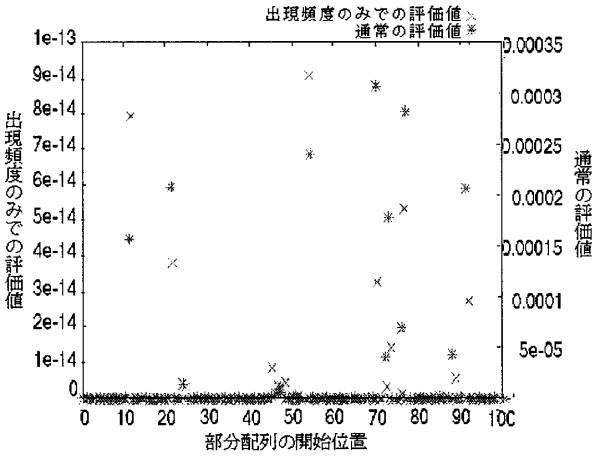

図1．2つの評価値による配列の開始位置

図1において，縦軸は評価値，横軸は部分配列の開始位置を表している．図から分かるように，2つの評価値(出現頻度を用いた評価値(×)と従来の評価値(*))の大きさは異なるが，大きくなる部分配列の開始位置に有意な違いが認められないことが分かった．他のすべての実験条件において同様の結果を得た．よって，FPGA に実装するにあたり除算の工程を省略し，出現頻度のみを用いた評価値を用いることにした．

4　回路アーキテクチャの概要

図2に提案する回路アーキテクチャの概要を示す．

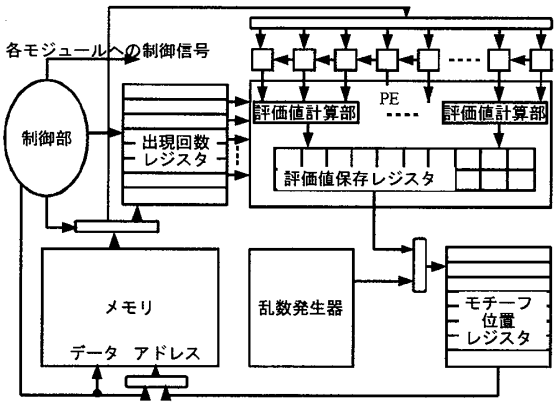

図２．回路アーキテクチャの概要

提案回路は，モチーフの候補位置の初期値を乱数によって決定する乱数発生器，決定されたモチーフの候補位置を保存するモチーフ候補位置レジスタ，モチーフ候補位置と出現回数を保存する出現回数レジスタ，評価値を計算し更新されたモチーフ候補位置を出力する PE (Processing Element)，PE の入力に接続されるシフトレジスタ，探索対象となる塩基配列を保存するメモリ，制御部からなる．PE は，シフトレジスタから配列データを入力し，並列で評価値計算を行う複数個の評価値計算部と，塩基配列のすべての位置の評価値から上位 20 個の値を保持する評価値保存レジスタからなる．制御部からの制御信号は各モジュールに接続され，各モジュールを制御する．

5　FPGA 実装

４章で述べたアーキテクチャの実現可能性を検証するために，プロトタイプを FPGA (Stratix III EP3SE110F780C4) 上に実装した．実装したプロトタイプは，N 本の塩基配列からそれぞれ長さ 8 のモチーフを探索するものである．評価値計算部は 4 個とし，乱数発生器には線型帰還シフトレジスタ （LFSR, Liner Feedback Shift Register）を用いた．LFSR は，レジスタのみで構成することが可能で，帰還関数を最適に設定すると周期の長い乱数のようなビット列を生成する．ここでは，文献[6]で示された最適な帰還関数を用いた．

表 1 はプロトタイプを FPGA 上に実現したときの FPGA 利用率を示している．本プロトタイプは 48.17MHz で動作している．

表 1．プロトタイプの FPGA 利用率

ブロック名	LC 数	DSP Element*	LC %
PE	1,845	112	2.16
シフトレジスタ	4,058	0	4.76
メモリ	13,368	0	15.69
モチーフ位置レジスタ	197	0	0.23
出現回数レジスタ	4,254	0	4.99
乱数発生器	3	0	< 0.01
データパス全体	22,699	112	26.64

* 1 DSP element = 9 bits ×9 bits 組み込み乗算器

6　評価

同じ塩基配列を使用して，プロトタイプと AlignAce のモチーフ探索速度を比較した．使用した塩基配列は，塩基をランダムに発生させて人工的に作成した長さ 108，計 20 本のデータ（１つずつ共通部分配列を含む）と，実応用データの塩基配列 (YBR018C gal7 と YBR019C gal10, および，それらの逆方向の配列，計４本）を用いた．YBR018C gal7 の塩基配列の長さは 500， YBR019C gal10 のそれは 655 である．

AlignAce の実行には，Dell Precision PWS670 Intel® Xeon™ CPU 3.00 GHz 2.99 GHz，1.00 GB RAM を用い，プロトタイプの検証は，Cadence 社 NC-verilog シミュレータを用いた．クロック数に，５章で示した動作周波数 48.17MHz を乗じて実行時間を算出した．結果を表２に示す．

表２．実行結果

データ種類	AlignAce	本アーキテクチャ
人工データ	4.15 s	0.58 ms
実応用データ	10.93 s	16.09 ms

表２から，本プロトタイプは AlignAce に比べ，人工データの場合では約 7000 倍，実応用データでは約 680 倍高速であることが分かった．プロトタイプでは，塩基配列のすべての部分配列位置の評価値の算出時間が実行時間のボトルネックとなる．よって，塩基配列長が短い方が，より高速化される．

7　まとめ

塩基配列中の類似部分配列であるモチーフを高速に検出することを目指し，PE が評価値を並列計算する回路アーキテクチャを提案した．本アーキテクチャの考え方は，他の Markov Chain Model を用いる問題などに広く適用可能である． N 本の各塩基配列から長さ 8 のモチーフを検出するプロトタイプを FPGA に実装した．本アーキテクチャは，ソフトウェアに比べて約 680 倍高速化できることがわかった．今後，ハードウェアによる Gibbs Sampling 法のさらなる高速化を検討する．

参考文献

[1] FASTA, http://fasta.genome.jp/

[2] BLAST, http://blast.ncbi.nlm.nih.gov/

[3] Lawrence, C.E., Altschul, S.F., Boguski, M.S., Liu, J.S., Neuwald, A.F., Wootton, J.C. “Detecting Subtle Sequence Signals: A Gibbs Sampling Strategy for Multiple Alignment,” Science, Vol.262, No.5131, pp. 208-214, 1993.

[4] Bailey, T. L. & Elkan, C. “Fitting a mixture model by expectation maximization to discover motifs in biopolymers,” Proc Int Conf Intell Syst Mol Biol, Vol.2, pp.28-36 , 1994.

[5] Roth, F. P., Hughes, J. D., Estep, P. W. & Church, G. M, “Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation,” Nat Biotechnol, Vol.16, No. 10, pp.939-45 , 1998.

[6] Peter Alfke, “Efficient Shift Registers, LFSR Counters, and Long Pseudo-Random Sequence Generators,” http://www.xilinx.com/bvdocs/appnotes/xapp052.pdf, 1998