1E5-1

Repulsive Parallel MCMC アルゴリズムによる塩基配列のモチーフ探索

Repulsive Parallel MCMC algorithm for DNA motif discovery

池端 久貴[*1*3] 吉田 亮[*1*2*3]
Ikebata Hisaki Yoshida Ryo

[*1] 総合研究大学院大学 複合科学研究科 統計科学専攻
The Graduate University for Advanced Studies, School of Multidisciplinary Sciences, Department of Statistical Science

[*2] 情報・システム研究機構 統計数理研究所
Research Organization of Information and Systems, The Institute of Statistical Mathematics

[*3] 独立行政法人 科学技術振興機構 CREST
JST, CREST

A conventional MCMC sampler tends to get stuck in a local mode of a multimodal distribution, and encounters the difficulty in moving to different modes within a finite time of simulation runs. To reach to many existing modes, the sampling is often repeated several times with different initial states. However, the problem remains unresolved; the different chains tend to get trapped in a particular mode having a much higher probability mass. To overcome the difficulty, we propose a new parallel MCMC algorithm. The idea is rather simple: During a parallel run of several MCMC simulations, a repulsive effect is added on each pair of the trajectories. Then the different samplers can explore different regions. After describing the methodology, an application to a pattern mining problem for DNA sequences, called the motif finding problem, is demonstrated.

1. モチーフ発見問題

遺伝子発現の制御機構を理解する上で，転写因子結合部位 (TFBSs：transcription factor binding sites) を予測することは非常に重要である．転写因子と呼ばれるタンパク質は，遺伝子上流に位置する TFBS の特異的な塩基配列を認識し，そこに結合することで，遺伝子の発現を制御している．一般に，転写因子の認識配列の長さは 10 塩基ほどである．ある転写因子に特異的な配列パターンは，複数の遺伝子の上流に埋め込まれている．したがって，一つの転写因子が制御する遺伝子は複数存在し，それらの標的遺伝子の多くは共発現する．そこで，共発現する遺伝子集合を同定した後，それらの上流配列に存在する短い共通配列のパターン，すなわちモチーフを発見することで，TFBS の予測が可能になる．

モチーフを探索する手法はいくつかに大別できるが，ここでは，Lawrence らによって提案されたギブス・サンプリングに基づくアルゴリズム [Lawrence 93] について考える．これまでに多くの拡張アルゴリズムが提案されてきたが，本稿で示されるように，依然として重大な問題点が残されている．遺伝子の上流配列には，多様なモチーフが複数存在するが，既存手法はそれらのごく一部しか検出できない．とりわけ，ギブス・サンプリングの場合，局所解へのトラップが問題となる．本研究では，複数のモチーフを網羅的かつ効率良く列挙するための並列型 MCMC アルゴリズムを提案する．

連絡先：池端久貴，総合研究大学院大学，東京都立川市緑町 10−3，ikebata.hisaki@ism.ac.jp

2. モデルとアルゴリズム

長さ L の n 本の配列 $\boldsymbol{S} = (S_1, \cdots, S_n)$ が与えられる．各配列は，開始位置 $\boldsymbol{u} = \{u_i\}_{i=1,\cdots,n}$ $(u_i \in \{1, \cdots, L-K+1\})$，長さ$K$ のモチーフ配列を持つと仮定する．単純に全ての位置を探索すると解候補の総数は $(L-K+1)^n$ となり，n が大きくなると現実的に取り扱える計算量ではなくなる．そこでパラメータ $\boldsymbol{u}, \boldsymbol{\theta}, \boldsymbol{\theta}_0$ によって特徴付けられる塩基配列の確率モデルを導入し，$\boldsymbol{S}$ で条件付けられた事後確率分布 $\pi(\boldsymbol{u}, \boldsymbol{\theta}, \boldsymbol{\theta}_0 \mid \boldsymbol{S})$ を評価し，モチーフの位置を予測する．

2.1 モチーフモデル (従来手法)

モチーフの構造を表すパラメータとして，サイズ $|\Sigma| \times K$ の Position Weight Matrix (PWM) $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_K)$ を導入する．モチーフの k 番目の文字は $\boldsymbol{\theta}_k$ をパラメータとする多項分布に従うと仮定する．Σ は文字集合であり，DNA 配列の場合，$\Sigma = \{\mathrm{A}, \mathrm{G}, \mathrm{C}, \mathrm{T}\}$となる．$|\Sigma|$ は Σ の要素数，$\boldsymbol{\theta}_k = \left(\theta_{k,\sigma}\right)_{\sigma \in \Sigma}$ は確率ベクトル $\left(\sum_{\sigma \in \Sigma} \theta_{k,\sigma} = 1, (k = 1, \cdots, K)\right)$ である．また，$\boldsymbol{\theta}_0$ を長さ $|\Sigma|$ の確率ベクトルとし，配列中のモチーフ以外の文字（バックグラウンド）はパラメータ $\boldsymbol{\theta}_0$ の独立な多項分布に従うと仮定する．ここ

で，n 本の配列 $\boldsymbol{S}$ が観測されたもとで，尤度関数は

$$P(\boldsymbol{S}|\boldsymbol{u},\boldsymbol{\theta},\boldsymbol{\theta}_0) = \prod_{\sigma\in\Sigma}\prod_{l=1}^{L}\theta_{0,\sigma}^{\sum_{i=1}^{n} I(s_{i,l}=\sigma)}\prod_{k=1}^{K}\left(\frac{\theta_{k,\sigma}}{\theta_{0,\sigma}}\right)^{\sum_{i=1}^{n} I\left(s_{i,u_i+k-1}=\sigma\right)} \quad (1)$$

となる（$s_{i,l}$ は配列 i の l 番目の文字を表す）．$\boldsymbol{\theta}$, $\boldsymbol{\theta}_0$ の事前分布には以下のようなディリクレ分布を仮定する．

$$P(\boldsymbol{\theta}_k|\,\boldsymbol{\beta}_k) = Z^{-1}(\boldsymbol{\beta}_k)\prod_{\sigma\in\Sigma}\left(\theta_{k,\sigma}\right)^{\beta_{\sigma,k}-1} \quad (2)$$

$$P(\boldsymbol{\theta}_0|\boldsymbol{\alpha}) = Z^{-1}(\boldsymbol{\alpha})\prod_{\sigma\in\Sigma}\left(\theta_{0,\sigma}\right)^{\alpha_\sigma-1} \quad (3)$$

ここで，$\boldsymbol{\beta}_k = \left(\beta_{k,\sigma}\right)_{\sigma\in\Sigma}$，$\boldsymbol{\alpha} = \left(\alpha_{\sigma,0}\right)_{\sigma\in\Sigma}$，$Z^{-1}$ は正規化定数である．

2.2 Markov Chain Monte Carlo 法による事後分布からのサンプリング

推定すべきパラメータは $\boldsymbol{u}$, $\boldsymbol{\theta}$, $\boldsymbol{\theta}_0$ である。ここでは、ギブス・サンプリングを用いて強い多峰性を持つ事後分布 $\pi(\boldsymbol{u},\boldsymbol{\theta},\boldsymbol{\theta}_0\,|\,\boldsymbol{S})$ から効率的にランダムサンプリングを実現する方法を提案する．

前述の従来のモチーフモデルでは，モチーフ長は固定長 Kとしていたが，実際にはさまざまな長さのモチーフが存在する．これに対して，本稿ではモチーフの長さ K をパラメータに追加する．まず，K の事前分布 $p(K)$ を以下のように与える．

$$p(K=k) = \begin{cases} p_k & : k = K_{min}, K_{min}+1, \cdots, K_{max} \\ 0 & : otherwise \end{cases} \quad (4)$$

ここで，$p_k > 0\ (k = K_{min}, K_{min}+1, \cdots, K_{max})$, $\sum_{k=K_{min}}^{K_{max}} p_k = 1$である．$K_{min}$，$K_{max}$ は仮定する最小，最大モチーフ長である．モチーフの長さ K を変えることは，モデルパラメータ $\boldsymbol{\theta}$ の列のサイズを変えることに相当する．そこで，$\boldsymbol{\theta}$ の列サイズに対する MCMC の遷移ルールを設計する．可変サイズの $\boldsymbol{\theta}$ に対して，MCMC の詳細釣り合い条件を課すために，Reversible Jump MCMC のアイデアを用いる [Green 95]．ある時点のモチーフ長 $K = k$ から $K^* \in \{k-1,\ k,\ k+1\}$ への遷移は，以下の離散分布に従うと仮定する．

1. $\Pr(K^* = k|\,K = k) = f_1$
2. $\Pr(K^* = k+1, \text{Extension at the left of } \boldsymbol{\theta}\,|\,K = k) = f_2$
3. $\Pr(K^* = k+1, \text{Extension at the right of } \boldsymbol{\theta}\,|\,K = k) = f_3$
4. $\Pr(K^* = k-1, \text{Contraction at the left of } \boldsymbol{\theta}\,|\,K = k) = f_4$
5. $\Pr(K^* = k-1, \text{Contraction at the right of } \boldsymbol{\theta}\,|\,K = k) = f_5$

ここで 3.の場合 (現在の PWM の右端に列を一つ追加する) を考える．追加する列の要素を以下のように与える．

$${\theta^*}_{k+1,\sigma} = \sum_{i=1}^{n} I\left(s_{i,u_i+k}=\sigma\right) \Big/ \sum_{\sigma\in\Sigma}\sum_{i=1}^{n} I\left(s_{i,u_i+k}=\sigma\right)$$

3.の遷移により $\boldsymbol{\theta}$ から $\boldsymbol{\theta}^* = (\boldsymbol{\theta}, {\boldsymbol{\theta}^*}_{k+1})$ になるとする．このとき，$\boldsymbol{\theta}^*$の採択確率を

$$\alpha(\boldsymbol{\theta},\boldsymbol{\theta}^*) = \min\left(1, \frac{p(\boldsymbol{\theta}^*|\,\boldsymbol{S},\boldsymbol{u})p(K^*=k+1)f_5}{p(\boldsymbol{\theta}|\,\boldsymbol{S},\boldsymbol{u})p(K=k)f_3}\right) \quad (5)$$

とする．次に，5.の場合（現在の PWM の右端の列を一つ削除する）を考える．$\boldsymbol{\Theta} = (\boldsymbol{\theta}_{-k}, \boldsymbol{\theta}_k)$ から$\boldsymbol{\theta}^* = \boldsymbol{\theta}_{-k}$ に遷移する．このとき，$\boldsymbol{\theta}^*$の採択確率を

$$\alpha(\boldsymbol{\theta},\boldsymbol{\theta}^*) = \min\left(1, \frac{p(\boldsymbol{\theta}^*|\,\boldsymbol{S},\boldsymbol{u})p(K^*=k-1)f_3}{p(\boldsymbol{\theta}|\,\boldsymbol{S},\boldsymbol{u})p(K=k)f_5}\right) \quad (6)$$

とする．1. の場合，採択確率は常に 1 となる．2. 4. については，3. 5. と同様に採択確率を導くことができる．

このようにモチーフ長をモデルパラメータとすることで従来手法よりも柔軟なモデルを構築することができる．アルゴリズムは以下のようになる．

・アルゴリズム

1. $\boldsymbol{u},\boldsymbol{\theta},\boldsymbol{\theta}_0$を初期化する．
2. $\boldsymbol{\theta}_{new} \sim \pi(\boldsymbol{\theta}|\boldsymbol{u},\boldsymbol{\theta}_0,\boldsymbol{S})$ により，$\boldsymbol{\theta}$を更新する．
3. ${\boldsymbol{\theta}_0}_{new} \sim \pi(\boldsymbol{\theta}_0|\boldsymbol{u},\ \boldsymbol{\theta},\boldsymbol{S})$により，$\boldsymbol{\theta}_0$を更新する．
4. $\boldsymbol{\theta}$の列サイズを Reversible Jump MCMC により更新する．
5. $\boldsymbol{u}_{new} \sim \pi(\boldsymbol{u}|\boldsymbol{\theta},\boldsymbol{\theta}_0,\boldsymbol{S})$により，$\boldsymbol{u}$を更新する．
6. 停止条件を満たさなければ，2.に戻る

しかしながら，後の数値実験で示されるように，このアルゴリズムをそのまま適用するだけでは局所解へのトラップにより複数のモチーフを検出することは難しい．

2.3 Repulsive Parallel MCMC

従来の MCMC 法は確率分布が多峰性を持つ場合，サンプル列が局所領域にトラップされ，その領域から抜け出すことが困難になる傾向がある．事後分布の複数のピーク付近からサンプルを得るために，初期値を変えて MCMC を複数回繰り返す方法が考えられるが，極端に高い確率を持つ領域が存在する場合，何度試行してもその領域に捉われてしまい，根本的な解決にならない．

本研究では，この問題を克服するために，複数の MCMC を並列実行しながらサンプラー間に反発効果を与え，それぞれのサンプラーが異なる領域を探索するように作用させる（図1）．アイデアは単純である．$\boldsymbol{\omega} = (\boldsymbol{u},\boldsymbol{\theta},\boldsymbol{\theta}_0)$

をパラメータセットとする．配列 $\boldsymbol{S}$ が観測されたもとで，事後分布のレプリカ $\boldsymbol{\pi}(\boldsymbol{\omega}_i|\boldsymbol{S})\ (i=1,\cdots,M)$ を M 個作り，更にレプリカ間に相互作用を与える関数 $\boldsymbol{\psi}(\boldsymbol{\theta}_1,\ \cdots,\boldsymbol{\theta}_M)$ を用いて，以下のような拡大事後分布を構成する．

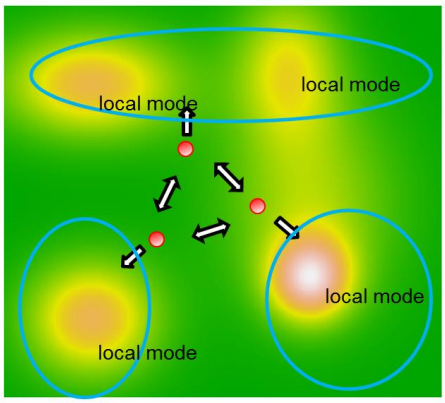

図 1: Repulsive Parallel MCMC.

$$\boldsymbol{\pi}(\boldsymbol{\omega}_1,\cdots,\boldsymbol{\omega}_M|\boldsymbol{S}) \propto \prod_{i=1}^{M}\boldsymbol{\pi}(\boldsymbol{\omega}_i|\boldsymbol{S})\boldsymbol{\psi}(\boldsymbol{\theta}_1,\ \cdots,\boldsymbol{\theta}_M)^T \quad (7)$$

この目標分布から $\boldsymbol{\omega}_1,\cdots,\boldsymbol{\omega}_M$ を同時にサンプリングする．ここで，T は温度パラメータで正の定数である．$\boldsymbol{\psi}(\boldsymbol{\theta}_1,\cdots,\boldsymbol{\theta}_M)^T$ は以下のようにあたえられる．

$$\boldsymbol{\psi}(\boldsymbol{\theta}_1,\cdots,\boldsymbol{\theta}_M)^T = \prod_{i=1}^{M} min_{j\neq i}\, exp\{-A(\boldsymbol{\theta}_i,\ \boldsymbol{\theta}_j)\}^T \quad (8)$$

ここで $A(\cdot,\cdot)$ は異なる列サイズの行列の類似度を評価する関数で，以下のように定義される．

$$A(\boldsymbol{\theta}_n,\ \boldsymbol{\theta}_m) = \max_{j\in\{1,\cdots,s\}}\sum_{i=1}^{p}\left(1-\frac{\|\boldsymbol{\theta}_{n,i}-\boldsymbol{\theta}_{m,i+j}\|_1}{2}\right)-gap\times s \quad (9)$$

ただし，$\boldsymbol{\theta}_n=(\boldsymbol{\theta}_{n,1},\ \cdots,\boldsymbol{\theta}_{n,p}),\boldsymbol{\theta}_m=(\boldsymbol{\theta}_{m,1},\ \cdots,\boldsymbol{\theta}_{m,(p+s)})$ $(s\geq 0)$ で，gap は比較する行列の列数の差に対するペナルティである．

レプリカ間の類似度の減少関数を導入することで，$\boldsymbol{\theta}_1,\cdots,\boldsymbol{\theta}_M$ のシミュレーション軌道に反発が起きる．その結果，M 本のサンプル列が異なるモードに到達する可能性が高まり，一回の計算でより多様なモチーフが検出されることが期待される．

3. 数値実験

既存手法と提案手法を比較するために数値実験を行った．TRANSFAC (http://www.generegulation.com/pub/databases.html) に登録されている 10 種類のモチーフ配列に揺らぎを与え，Tompa らにより使用されたデータ (hm01r.fasta: 2000bp, 18 本)[Tompa 05]の各配列に埋め込んだ．したがって，各配列は少なくとも 10 本のモチーフを持つことになる．MCMC を実行する上で，5 つのレプリカを作り，各レプリカから 5000 個のサンプルを生成した．性能評価指標として

$$\frac{\text{(予測モチーフ位置} \cap \text{真のモチーフ位置) の文字数}}{\text{真のモチーフ位置 の文字数}}$$

を用いた．

結果を図 2 に示す．反発を加えない場合 $(T=0)$ と比べて，反発を加えることによってモチーフの検出性能が向上していることが分かる $(T=30,50)$．しかしながら，更に反発を強くすると $(T=70,90)$，検出性能が劣化することが確認された．これはデータセットに対して，最適な温度パラメータTを設定することの重要性を示唆するものである．単純に初期状態をランダムに置き換えて複数回実行する従来の方法では $(T=0)$，ある限られた領域へのトラップや，複数のサンプラーが重複して同じ領域を探索するなど，探索の効率が極めて悪いことが確認できる（図 3）．

一方、サンプラー間に反発を与えた場合 $(T=50)$，5 つサンプラーがそれぞれ別の領域を探索するようになり，従来の方法では発見できなかったモチーフの近傍にも到達できている（図 4，図 5）．

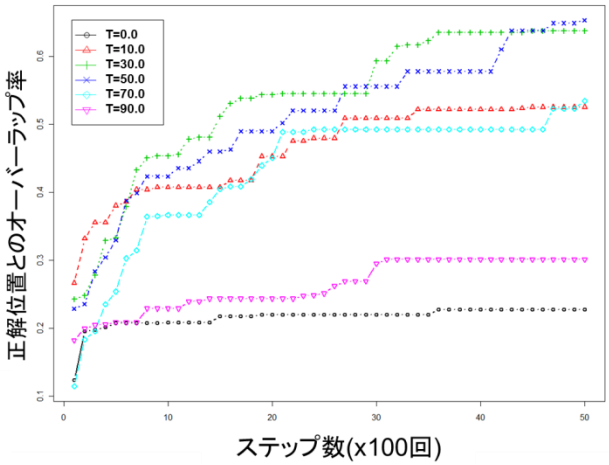

図 2: 反発の強さの違いによる性能比較

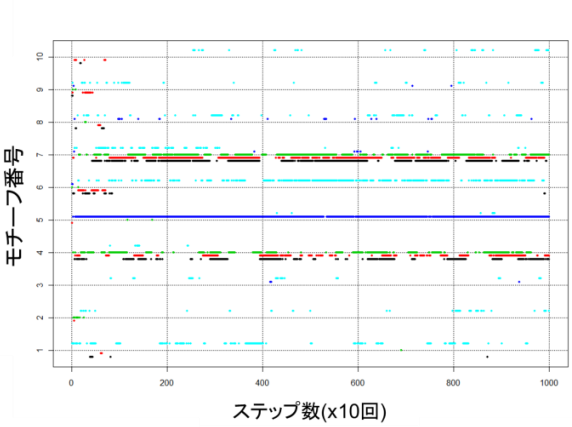

図 3: 反発なし ($T = 0$) の MCMC の遷移の様子. 色づけされた各点(黒，赤，緑，青，水色) は 5 つのレプリカのサンプル列に対応する．横軸は MCMC のステップ数，縦軸は正解モチーフの番号を表す．各サンプル列はアライメントスコアが最も高いモチーフ番号に割り付けられている.

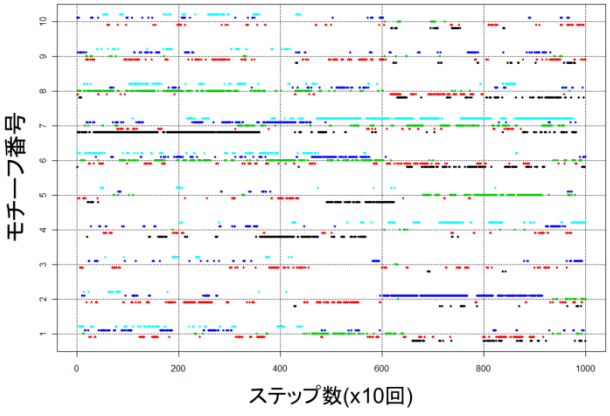

図 4: 反発あり($T = 50$)の場合の各サンプル列の最近傍モチーフ番号.

4. 今後の課題

本研究では，モチーフ発見問題における並列型ギブス・サンプリングの新しい手法を提案した．複数のレプリカサンプラーを生成し，それらの軌道に反発作用を加えることで，従来法では検出できなかった多様なモチーフ配列を検出することに成功した．現時点における未解決課題は，反発の強さを制御する温度パラメータの設定方法である.

本稿では，18 本の長さ 2000bp という小規模なデータセットを用いて提案手法の検証を行ったが，より実践的なアプリケーションでは，より大きなデータセットへ適用が求められる．データの規模が大きくなると，モチーフ開始位置の条件付き確率の計算，尤度評価における文字カウントの操作が計算速度のボトルネックになることが予想される．計算量の軽減と高並列計算機への実装は今後の検討課題である.

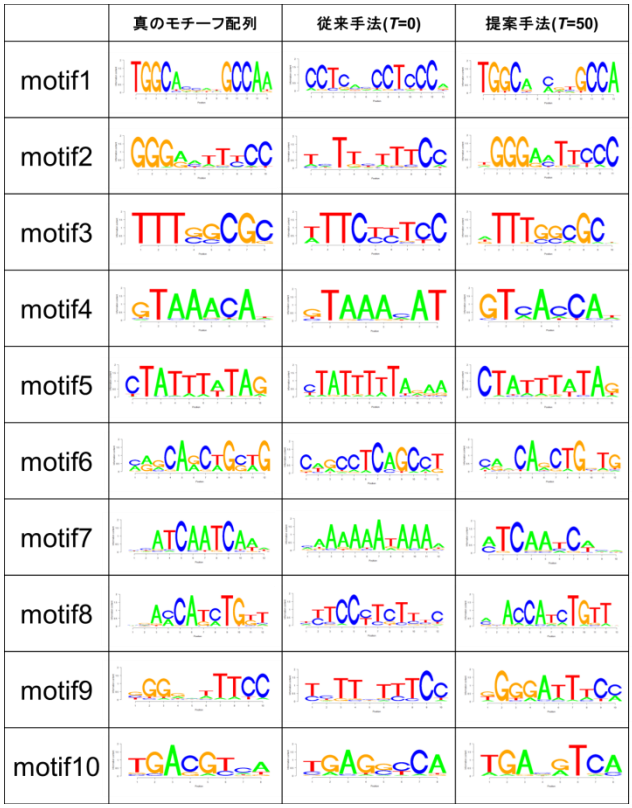

図 5: 従来法とサンプラーに反発を与えた場合の比較．第二列 (T=0)，第三列 (T=50) は各正解モチーフに最も近づいたサンプルの Sequence Logo を表す.

参考文献

[Lawrence 93] Lawrence CE, Altschul SF, Boguski MS, Liu JS, Neuwald AF, Wootton JC: Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment, *Science*, 262(5131), 208-214, 1993

[Green 95] Green PJ: Reversible jump Markov chain Monte Carlo computation and Bayesian model determination, *Biometrica,* 82(4), 711-732, 1995

[Tompa 05] Tompa M, Li N, Bailey TL, Church GM, De Moor B, et al.: Assessing computational tools for the discovery of transcription factor binding sites, *Nat. Biotechnol*, 23(1), 137-144, 2005