論文内容の要旨

論文題目
転写因子結合部位モチーフにおける最適な疑似度数とその検出限界

氏名　　西田　圭伸

序論

転写因子は特定のゲノム配列に結合することで，遺伝子の転写制御をすることが知られている．この転写因子結合配列は，完全に同一の配列ではなく，ある程度の塩基ゆらぎをもって観測されることがほとんどである．そのため，転写因子結合部位は，モチーフと言われる配列に共通するパターンとして取り扱われる．そして，パターンのような規則性で表現することで，配列やモチーフ自体に確率・統計的な指標を与えることができる．しかし，モチーフ表現を統計的な指標として使うには，実用上問題が生じる場合がある．その中で，本論文は 2 つの問題に着目し，研究を行った．1 つ目は，「最適な疑似度数の探索」である．これは，転写因子結合部位配列からモチーフ表現を作る際に生じる，ゼロ頻度問題を回避するために使われる疑似度数に焦点を当てた研究である．2 つ目は，「モチーフ発見ソフトウェアの検出限界」である．これは，複数の配列から共通する転写因子結合部位配列を見つけるモチーフ発見ソフトウェアが，既知の転写因子結合モチーフのうち，どこまでが検出可能なのだろうかという疑問に答えるための研究である．

最適な疑似度数の探索

モチーフ表現として最もよく使われているのは，行列モデルを用いたものである．Position Frequency Matrix（PFM）は，転写因子結合部位配列の各位置での塩基の出現頻度を表したもので，行に塩基の種類，列に配列の位置を示す行列の形となっており，各位置に，どれだけ塩基が出現したかを表現することができる．そして，Position Probability Matrix（PPM）は，PFM の各位置での塩基の出現頻度を，出現確率に変換したものである．この PPM は，転写因子結合部位に確率的な評価基準を与えるための基本となる．PPM からさらに別の指標に変換したモチーフ

表現の 1 つとして，Position Weight Matrix（PWM）が挙げられる．これは，PPM の各要素を配列のモチーフ以外の部分（バックグラウンド）での塩基の出現確率で割り，各要素をさらに対数変換して作られる．つまり，PWM の各要素の値はバックグラウンドに比べて，どれだけ高い比でモチーフの塩基が現れたかを対数で表すのである．しかし，PFM が少ない配列から生成された場合には，多くのゼロの要素を含むことがある．その場合，ゼロの要素は対数変換の際に負の無限大になるため，そのゼロとなる塩基を持つ配列は予測から外れることになる．しかし，真のモチーフでは現れるはずなのに，少ないサンプルのためにゼロの観測となった場合，このゼロは予測に大きな影響を与える．この問題を避けるため，疑似度数と言われる値を足す手法が用いられる．この疑似度数は経験的に用いられており，0.01 や 4 などの定数，元の配列数の平方根など多岐にわたっている．そこで本研究は，JASPAR に登録されているモチーフを，真のモチーフの出現確率とし，その確率から複数の転写因子結合配列を生成する．そこで，疑似度数が足された疑似度数と元のモチーフの類似度を判断基準とし，疑似度数の体系的な評価を行った．また，その結果から最適な疑似度数の提唱を行った．

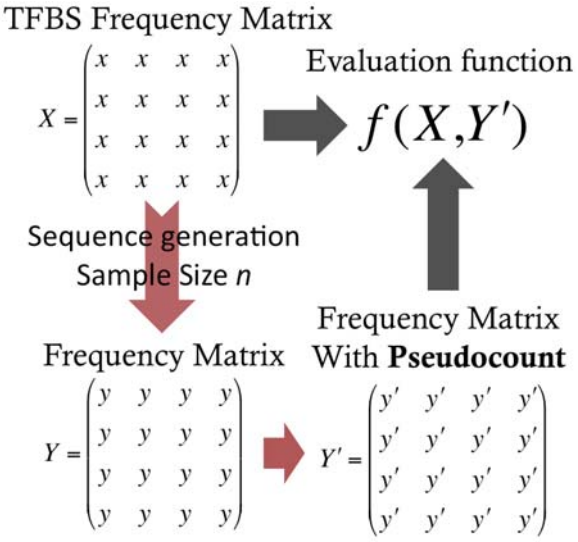

図 1　手順の概要

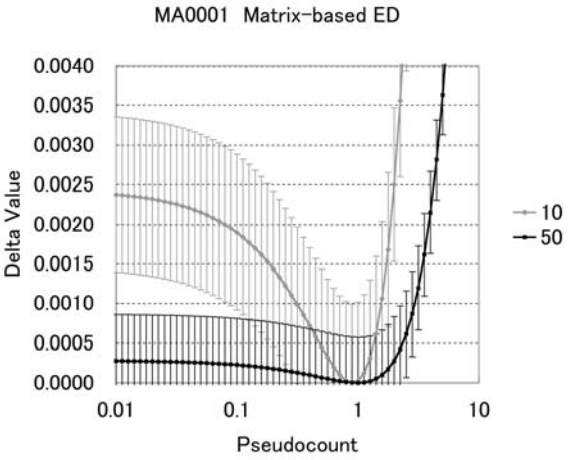

図 2　疑似度数と距離の関係

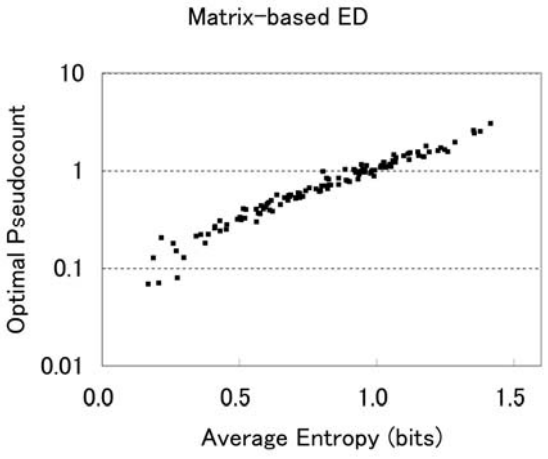

図 3　最適な疑似度数とエントロピー

方法

JASPAR に登録されている PFM を PPM に変換し，各塩基の出現確率を利用して新たな配列を生成する．その生成された配列を集めて PFM を作り，0.01 から 10 までのさまざまな疑似度数を足す. 生成された PPM と元の PPM を比較し，疑似度数の効果を評価関数にて評価する(図 1)．評価にはユークリッド距離や順位相関係数など 7 種類の方法を試した．この手順を，ひとつの条件につき 100 回の繰り返しを行い，その平均値を

評価値として用いる．

結果

図 2 は，AGL3 (MA0001)の評価法を PPM の要素に対するユークリッド距離を求めた場合の結果である．横軸には疑似度数，縦軸には評価値が表してある．そして，最低値が 0 になるように値の差分のみを表示した．生成した配列数は図の右に示してある．ここから，元の PPM と生成 PPM の距離が縮まる，最適な疑似度数があることがわかる．また，生成配列数が違っても極小となる位置が同じであることから，最適な疑似度数は生成配列のサイズに依らないこともわかる．モチーフの最適疑似度数とモチーフの位置ごとの平均エントロピーの関係を図 3 に示す．ここから，モチーフの最適な疑似度数と平均エントロピーには強い相関関係があることがわかる．

考察

モチーフ生成による解析の結果，最適な疑似度数はモチーフのエントロピーによって決まることがわかった．しかし，実用上は生成されたモチーフのみ知ることができ，元のモチーフを知ることはできない. そこで次の点に着目して, 最適な疑似度数を提唱することにする. まず, JASPAR のモチーフの結果は，1 近辺に最適な疑似度数が集中することと（図 3），小さい疑似度数を加えた場合は，大きなものを加えた場合よりも好ましい結果であることから（図 2），0.8 を JASPAR の転写因子結合モチーフに対する推奨疑似度数とする．

転写因子結合モチーフの検出限界の探索

転写因子の結合モチーフを新たに見つけ出すには，実験的手法と計算機的手法の両方が必要になる場合がある．なぜなら，実験的手法によっては，転写因子結合部位を含む近傍のゲノム配列までしか絞り込むことができないためである．この問題は，各配列に頻出する類似度の高い配列を探し出す問題と捉えることができる．しかしながら，統計的に最適なモチーフは現実的な計算時間で検出することができない．つまり，現在利用できるモチーフ発見ソフトウェアは，最適解である保証が無い答えを返すのである．そのため，多種多様なソフトウェアが開発され，アルゴリズムの違いなどからそれぞれ異なる性能を発揮している．しかし，どのような入力配列があればどの程度の有意なモチーフを探す事ができるのか，といった性能の限界についてはよくわかっていない．そこで本研究は，モチーフの検出限界を探索することを目的とし，JASPAR のモチーフを様々なバックグラウンド配列に埋め込むことで，性能評価用のデータセットを作成した．このデータセットを用いることでモチーフ探索ソフトウェアの検出限界を知ることができる．またこの結果は，実験手法での探索範囲の絞り込みのための，重要なガイドラインになると考えている．

方法

バックグラウンド配列として，長さ 1600，800，400，200，100bp の配列を，それぞれ 128，64，32，16，8 本用意する．配列の生成方法としては，ヒトの遺伝子の上流配列 1600bp を無作為に選択したものを用いる．1600bp 以下の配列長は無作為に切り出した配列を用いる．また，繰り返し実験のため，これらの条件の全組み合わせを 100 セット用意する．人工の転写因子結合配列は，JASPAR の PFM を PPM に変換し，その塩基の出現確率を用いて生成する．そして，それぞれの配列のデータセットの配列数の 3/4 に，生成した転写因子結合配列を埋め込む．モチーフ発見ソフトウェアは，Gibbs Sampler，MEME，Weeder，Seeder を用いる．予測した部分か埋め込んだ配列の，どちらか短い方の長さの 2/3 が重なっていた場合は，その予測が正しいものと判断する．Gibbs Sampler，Weeder，Seeder は，実行に必須なパラメータとして予測するモチーフの長さが必要になる．Gibbs Sampler と Seeder は，6，8，10，12 塩基長の 4 種類を設定し，Weeder は 6，8，10 塩基長の 3 種類を設定する．その中で，最もよい性能を示したものをそのソフトウェアの結果とする．

結果

まず，5 つのモチーフと，バックグラウンド配列の全部の条件の結果から（図 4），Weeder が比較的安定してよい結果となることがわかった．Weeder 以外のソフトウェアがよりよい精度となる場合でも，Weeder が僅差の精度を示すので，Weeder を用いることで検出限界を定めることができると考えられる．そこで，Weeder を用いて全モチーフの探索を行った．例として，モチーフの検出感度（Sensitivity）が 0.5，PPV（Positive Predictive Value）が 0.4 をこえるものを，予測可能なモチーフとした場合の結果を表 1 に示した．

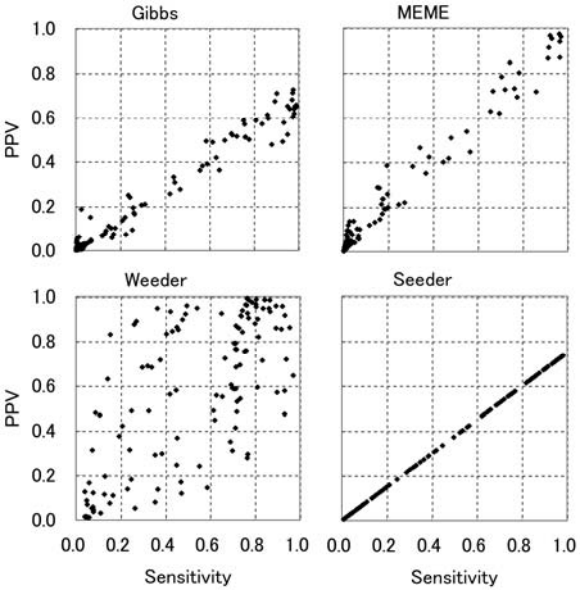

図 4　ソフトウェアごとの予測精度

Sequence Length	Number of Sequence				
	8	16	32	64	128
100	34.4%	54.1%	52.5%	43.4%	41.8%
200	11.5%	39.3%	39.3%	35.2%	32.8%
400	4.9%	20.5%	27.9%	30.3%	27.0%
800	4.9%	11.5%	18.0%	20.5%	22.1%
1600	0.8%	4.1%	9.0%	13.1%	12.3%

表 1

考察

表 1 に示すように，よい条件でも既存のモチーフの 50%程度しか発見することができないことがわかった．また，入力の配列の数が 10 以下の少ない場合や，100 以上の多い場合には，予測できるモチーフが減ってしまうことがわかる．そこで，効率的にモチーフを探索するために，入力配列数をコントロールがすることが有効であること考えられる．