

生物学の知識の埋め込みを可能とする モチーフ検索システムの開発

吉原 久雄* 賀屋 秀隆[§] 松井 藤五郎[†] 朽津 和幸[‡] 大和田 勇人[†]

東京理科大学大学院理工学研究科経営工学専攻* 同 理工学部 経営工学科[†]
同 応用生物科学科[‡] ゲノム創薬研究センター[§]

1 序論

現在バイオインフォマティクスの分野では、タンパク質のモチーフ検索をするものとして HMMER [1] が幅広く使用されている。HMMER は、クエリー側の配列から確率・統計的な profile-hmm [1] というモデルを構築してデータベースと比較する。ここで構築されたモデルは確率・統計的に見れば確かにそのようになる。しかし生物学から見ると、「とある機能を持つタンパク質配列の 20 文字目は K でなければならない」といったものや、「ドメイン 1 とドメイン 2 の間の配列は特に類似している必要はない」といったような知識が反映されてない場合が多い。

よって本論文では、生物学者の知識が反映されるように profile-hmm を修正することによって、よりの確にタンパク質を検出する手法を提案し、それを実行するためのモチーフ検索システムを開発する。

2 従来手法の問題点

従来の手法の問題点は、一般的な HMMER の使用方法で profile-hmm を構築した場合、予想していた profile-hmm ではない、つまり生物学者の知識が反映されていない結果になっていることが多く、その結果検索にも影響してしまうこともある、という点である。機能が類似している配列から profile-hmm を構築した場合、モチーフ部分に偏りが見られるのだが、そのほかの機能に関係のない配列部分も何かしらの偏りがあるとしてモデルを構築してしまう。これでは配列全体が類似しているものでないと検出されない、

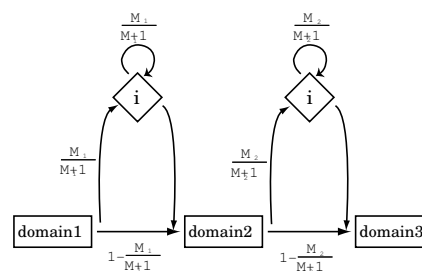


図1 編集後の profile-hmm

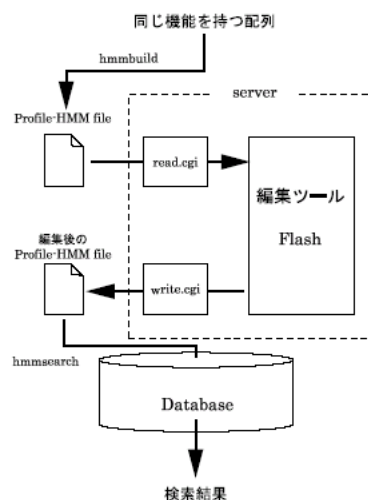


図2 システム構成図

モチーフ部分以外の配列が類似しているものも検出されてしまう、ということが起きる。これはいくつかのドメインで構成されているタンパク質に多く見られるようである。

3 提案手法

従来は構築された profile-hmm をそのままモチーフ検索に使用していたが、本研究では先ほども示したように生物学者の知識を反映させるため、構築後の profile-hmm ファイルに表記されているアミノ酸出力確率、状態遷移確率を編集する。profile-hmm ファイルに表記されている数字は

Development of motif search system that enable burial of knowledge of biology

Hisao Yoshihara*, Hidetaka KAYA[§], Tohgoroh MATSUI[†], Hayato OHWADA[†], and Kazuyuki KUCHITSU[‡]

*Department of Industrial Administration, Graduate school of Science and Technology, Tokyo University of Science, [†]Department of Industrial Administration, Faculty of Science and Technology, [‡]Department of Applied Biological Science, Faculty of Science and Technology, [§]Genome and Drug Research Center

類似度を表すスコアの計算の都合上、確率の対数を取っているため、HMMER で使用されている計算式を用い、確率と対数の両方向変換を必要に応じて行う。

いくつかのドメインで一つのモチーフが形成されている時、そのドメイン間のアミノ酸配列は必ずしも類似していなくてもよい。そのような場合、モチーフ間の配列は一般的なアミノ酸出力を用い、一つの挿入状態として表すことにする。そこで今回は挿入状態での状態遷移確率を求めるために、同じようにモチーフ検索に使用されている Meta-MEME [2] で使用している近似式を使用した。連続する挿入状態の確率を IP、モチーフ間のアミノ酸の数を M とすると、近似式は次の通りである。

$$IP = \frac{M}{M+1} \quad (1)$$

これにより、profile-hmm を自由に編集することが出来る。各々のドメイン内でも図 1 のような profile-hmm を使うことによりある程度のパターンを持たせつつ、且つ各ドメイン間でギャップ・挿入状態を明確にすることが出来る。これにより今までの profile-hmm よりも柔軟に、且つ生物学者の知識を反映させた検索ができるようになった。

今回の提案手法でのシステム構成は図 2 の通りである（破線で囲まれた所が提案した部分）。インターフェース等の大部分は Flash を使用しているが、Flash はファイルの入出力が不十分のため cgi を使ってそれを補っている。cgi 言語として Ruby を使用した。

4 実験

今回は検索するデータベースは nr、profile-hmm は Pfam から複数のドメインを持つスーパーファミリーと、それを編集したものを使用した。pfam から選び出した配列を表 1 に示す。これらは Pfam、PROSITE 両方に登録してあるスーパーファミリーである。提案手法では、ドメイン毎の情報が詳しく示されている PROSITE を参考にし、Pfam に登録されている profile-hmm のアミノ酸出力確率と状態遷移確率を編集した。

nr データベースの中にはここで選ばれた各々のスーパーファミリー内のタンパク質はほぼ全て含まれている。Pfam で登録されている profile-hmm と編集した profile-hmm で、各々のスーパーファミリーに属するタンパク質をどれだけ誤検出を少なくして検出出来るかを比較した。図 3 が実験結果を ROC 曲線で表したものである。ROC 曲線とは縦軸に感度を、横軸に（1-特異度）をとってプロットしてできる曲線である。ROC 曲線において左上方にある曲線ほど有用であるといえる。

編集したファイルの検索結果は、ドメイン部分のみ類似

表 1 使用したスーパーファミリー

Aldehyde dehydrogenase family
Arginosuccinate synthase
Carboxylesterase
Chitinase class I
Glycoprotein hormone
Glycosyl hydrolase family
GMC oxidoreductase
Histidine acid phosphatase
NAD-dependent DNA ligase adenylation domain
Phosphotriesterase family
POT family
Proliferating cell nuclear antigen, C-terminal domain
Sodium:solute symporter family
Sulfatase

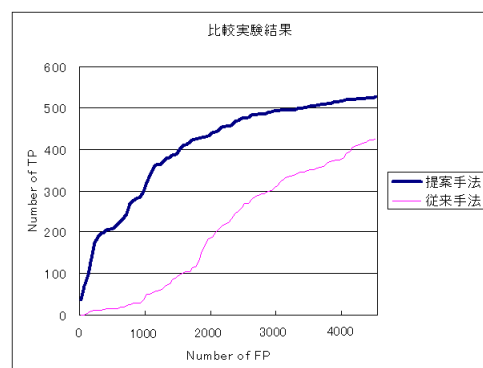


図 3 実験結果

しているものを出力しており、ドメイン部分以外の類似の結果、検出されたタンパク質は全く含まれていなかった。又、提案手法の方が従来の方法で検索したものよりも誤検出が著しく少ない。これにより、ドメイン部分に重点を置いた比較がされていることが分かる。しかし、正しく検出されたタンパク質の数を比べると、従来手法に比べ本研究での提案手法が少なくなっている。これはドメイン部分のアミノ酸出力確率も編集したために、ドメインの柔軟性が多少低くなってしまったのが原因であると考えられる。

5 結論

今回は HMMER で学習し終わったファイルを編集することにより生物学者の知識を反映させ、誤検出を 74.4% 減少させることが出来た。ただし、ドメイン部分のアミノ酸出力確率も編集すると、ドメイン部分の配列パターンの柔軟性が弱くなり、検出出来るタンパク質の数は減少してしまう。

参考文献

- [1] Sean R.Eddy. Profile hidden markov models. *BIOINFORMATICS REVIEW*, Vol. 4, pp. 755–763, 1998.
- [2] Meta-MEME. <http://metameme.sdsc.edu/>.