1W-5

マルチドメインに注目したモチーフ検索

瀬下 真吾　松井 藤五郎　大和田 勇人

東京理科大学 理工学部 経営工学科

1 はじめに

近年，バイオインフォマティクスの分野ではタンパク質のアミノ酸配列に関する情報をコンピュータにより解析して，タンパク質の機能を予測する研究が進められている．

複数のタンパク質が生物学的に類似した機能を持つ場合，それらはタンパク質ファミリーとしてまとめられる．モチーフ検索は機能未知のタンパク質とファミリーに属するタンパク質の配列的特徴の比較を行うことで機能の予測を行う手法である．モチーフ検索を行うためのアプリケーションとして HMMER [1] が広く利用されている．

本研究では，マルチドメインが存在するタンパク質ファミリーを対象にしたモチーフ検索を行う．ドメインとはタンパク質ファミリー内で進化的に保存されており，機能的に重要なアミノ酸配列の領域である．複数のドメインによって機能を果たす時，それらをマルチドメインと呼ぶ．

マルチドメインを持つタンパク質を対象にした場合，HMMER を用いた一般的な検索では各ドメインを考慮することができず，マルチドメインを持たないタンパク質を検出してしまうことがある．そこで本研究では，ドメインごとに検索した結果を統合することによりマルチドメインを持つタンパク質を検索する手法を提案する．

2 HMMER によるモチーフ検索

2.1 HMMER とは

HMMER は HMM（隠れマルコフモデル：Hidden Markov Model）を用いてドメイン配列群のモデル化・データベースへの検索等を行うためのツール群である．HMMER によるモチーフ検索では，まず，ファミリーに属するタンパク質からドメイン配列群（ドメインにあたる配列を集めたもの）を取り出す．次に hmmbuild プログラムによりドメイン配列群の各位置でどのようなアミノ酸が出現するかをモデル化する．このモデルをプロファイル HMM と呼ぶ．ドメイン配列群には共通して出現するアミノ酸パターン（モチーフ）が存在するため，構築されたモデルはそのドメイン配列群の特徴を表すことができる．このモデルをクエリーとしてタンパク質データベースへの検索を hmmsearch プログラムにより行い，モデルとの一致度が高い配列を持つものを検出する．ただし，hmmsearch はローカルアライメント（モデルと最も良く一致した部分領域）を求めるため，検出されたタンパク質中にモデル全長に渡っての一致領域があるとは限らない．

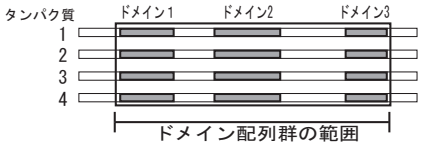

図 1　マルチドメインを持つタンパク質

検出されたタンパク質には Score と E-value の評価値が与えられる．

Score　モデルとの適合を示すスコア．よく適合しているほど値が大きくなる．

E-value　検索に用いたデータベース中で，上記の Score 以上の類似領域を持つタンパク質数の期待値．E-value が低い値であるほど、そのタンパク質が検出されたことは偶然でない確かな一致であるとみなすことができる．

2.2 HMMER の問題点

図 1 に示すようなマルチドメインを持つタンパク質からプロファイル HMM を構築する際，一般的には全てのドメインを含む範囲を 1 つのドメイン配列群として hmmbuild への入力とする．

こうして構築したプロファイル HMM を用いてデータベースへ検索を行った場合，マルチドメインを持たないタンパク質が検出されてしまうという問題がある．この原因は，hmmbuild ではマルチドメインを持つモデルを構築できないことにある．また，hmmsearch の検索アルゴリズムはモデルとの部分的な類似配列を持つものを検出してしまうため，モデル中にマルチドメインが存在しても，マルチドメイン含んだタンパク質が検出されるとは限らないのである．

3 マルチドメインに注目した検索手法

3.1 提案手法の概要

上で述べた問題点を解消するために，本研究では，マルチドメインを含むひとつの範囲からモデルの構築と検索を行う替わりに，ドメインごとにモデルの構築と検索をし，その結果を統合する手法を提案する．概要を図 2 に示す．

まず，Web 上で公開されているドメインデータベースからドメインごとのアミノ酸配列群が記述されたファイルを取得する．図中の D_i がこのファイルを表している．次に，ファイルごとに hmmbuild を用いてプロファイル HMM を構築する．続いてプロファイル HMM ごとに hmmsearch を用いてタンパク質データベースへの検索を行う．そして，ドメインごとに

Motif Search which pays attention to Multi-Domain
Shingo SEJIMO, Tohgoroh MATSUI, Hayato OHWADA
Department of Industrial Administration, Faculty of Science and Technology, Tokyo University of Science

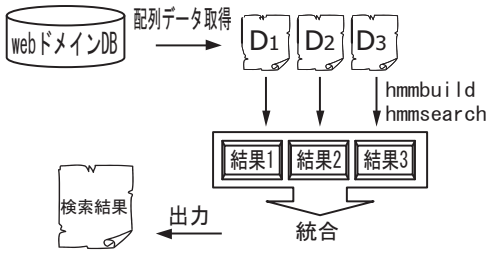

図 2　提案手法の概要

得られた検索結果を統合する．

検索結果の統合にはマルチドメインを持つタンパク質を絞り込む過程と，絞り込まれたタンパク質の確率的類似度を計算し再評価する過程がある．

3.2　マルチドメインを持つタンパク質の抽出

絞込みの第一段階として，ドメインごとの検索結果からタンパク質に付けられた固有の識別番号を取り出し，共通要素を求めることで全てのドメインで検出されたタンパク質の集合を得る．

第二段階では，集合に含まれる個々のタンパク質に対し，モデルと類似している領域の先頭位値と最後尾位置の情報を用いてドメインが正しい順序で並んでいるかを判定する．これにより，マルチドメインが正しい並び順で全て存在するタンパク質を得ることができる．

3.3　結合 E-value を用いた確率的類似度の再評価

複数のドメインが同時に一致する確率を個々の確率的類似度から求めるために本論文では結合 E-value という新しい評価値を提案する．

結合 E-value を求めるためには Bailey [2] らの提案した結合 P-value を用いる．タンパク質 s のドメイン d における Score が x_d であり，ローカルアライメントの長さが l であるときに，長さ l の部分配列の Score が x_d 以上になる確率を $P_d(s)$ とする．ドメイン数が n のマルチドメインタンパク質においてドメインごとの $P_d(s)$ が独立の時，その同時確率 $Z_n(s)$ を次のように定義する．

$$Z_n(s) = \prod_{i=1}^{n} P_i(s) \tag{1}$$

ここで $P_i(s)$ はドメイン d_i における $P_{d_i}(s)$ を短く表したものである．

この時，ドメイン数 n の同時確率 Z_n がとりうる値の中で，$Z_n(s)$ 以下の値をとる確率を結合 P-value と呼び，次式で計算する．ただし，$Z_n(s)$ は p として表示する．

$$F_n(p) = p \sum_{i=0}^{n-1} \frac{(-\ln p)^i}{i!} \tag{2}$$

本研究では $P_i(s)$ を求める方法として (2) 式の代わりに E-value から計算する手法を用いる．E-value は P-value と *DBsize*（検索に用いたデータベースの大きさ）の積によって計算されているので，式変換を行い次式を得る．

$$P_i(s) = \frac{Evalue_i(s)}{DBsize} \tag{3}$$

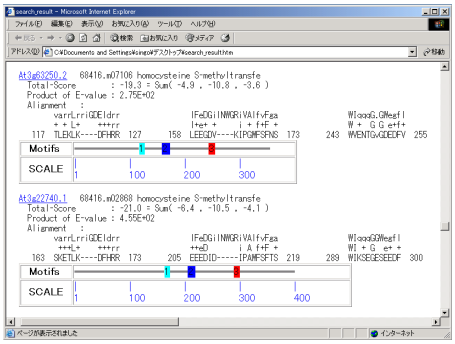

図 3　検索結果の例

これにより $P_i(s)$ を容易に計算することができる．

また，式 (4) によって得られた $F_n(p)$ に *DBsize* を掛けることで結合 E-value：$E(s)$ を求める．

$$E(s) = F_n(p) \times DBsize \tag{4}$$

HMMER における E-value と同じく $E(s)$ は数値が小さいほど統計的に良い．そこで，検出されたタンパク質の $E(s)$ を昇順に並び替えて検索結果を出力する．

4　実行例

本研究の提案手法に基づいてモチーフ検索システムを作成した．図 3 は 3 つのドメインをもつタンパク質ファミリーを検索した結果の例である．検出されたタンパク質それぞれについて，登録番号，結合 E-value，各ドメインと一致した領域のアミノ酸配列を表示している．また，ドメインの位置情報をグラフで表示することで，マルチドメインが全て存在し正しい並び順であることを視覚的に確認できるようになっている．

5　まとめ

本論文ではマルチドメインに注目したモチーフ検索を行うために，ドメインごとの検索結果を統合し，結合 E-value により確率的類似度を評価する手法を提案した．

提案手法によってマルチドメインを持つタンパク質の検索が可能となった．しかし，検出されたタンパク質とファミリーに属するタンパク質を比べると，ドメイン間のアミノ酸配列の長さが異なっていた．長い進化の過程ではアミノ酸の大きな置換・挿入・欠失が起きることがあるため，本論文では強く保存されているドメイン領域以外は配列長を考慮しない方針をとったが，これについては今後検討する．

参考文献

[1] S.R. Eddy. Multiple alignment using hidden markov models. *Ismb*, Vol. 3, pp. 114–120, 1995.

[2] T.L. Bailey. and M. Gribskov. Methods and statistics for combining motif match scores. *J. Comput. Biol.*, Vol. 5, pp. 211–221, 1998.