大きく変異したマルチドメインタンパク質のための MDHMMER の改良

平田 裕和* 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

タンパク質の機能の発現に必要な領域であるドメインに注目して相同性検索を行うことは，進化的に遠縁なタンパク質の発見に有効である．タンパク質が複数のドメインによって機能を発現する時，このようなタンパク質を本論文ではマルチドメインタンパク質と呼ぶことにする．マルチドメインタンパク質を持つ遠縁の遺伝子を発見するために，ドメイン間の配列を考慮しない相同性検索手法を用いた MDHMMER [1] が提案された．しかし，MDHMMER の手法の場合，すべてのドメインを検出できなければならないという問題点があった．これは複数のドメインの中の一つが大きく変異している場合，そのドメインを発見できなければそのタンパク質を発見できないということを意味している．

そこで本論文では検索アルゴリズムを改善し，大きく変異したドメインを持つマルチドメインタンパク質に対しても有効な検索手法を提案する．

2 提案手法

2.1 MDHMMER

MDHMMER はマルチドメインタンパク質を対象とした相同性検索を行うためのツールである．マルチドメインタンパク質に対しての相同性検索を行うために，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用いている．検索の評価値として結合 E-value を定義している．これは検索に用いたデータベース中から誤って相同だと判断されるタンパク質の数の期待値あり，値が小さいほど統計的に有意である．

瀬下らの手法においては各ドメインの評価値が得られることを前提としている．そのため各ドメインにおける相同性検索の結果，大きく変異していることにより相同であると判断できなかったドメインが一つでもあった場合，そのタンパク質の結合 E-value を算出できなかった．

2.2 提案手法

本研究では HMMER [2] が発見したドメインのみからマルチドメインタンパク質全体の結合 E-value の算出を行う手法を提案する．図 1 に提案手法の一連の流れを示す．

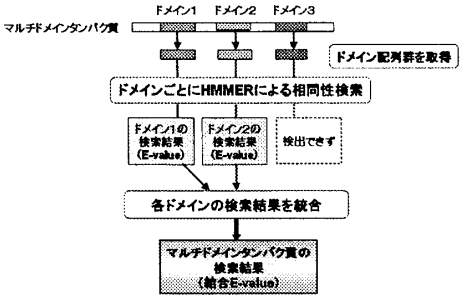

図 1 提案手法の流れ

STEP1 検索対象であるマルチドメインタンパク質に含まれるドメインの数だけドメイン配列を取得する．取得したタンパク質のアミノ酸配列からドメイン部分のアミノ酸配列を抜き出すか，ドメイン配列群を提供しているデータベースから取得する．

STEP2 ドメインごとに相同性検索を行う．相同性検索には HMMER を用いる．

STEP3 ドメインごとの検索結果を統合する．

HMMER によるドメインごとの検索結果からは，タンパク質 s のドメイン i における E-value: $E_i(s)$ が取得できる．マルチドメインタンパク質の確率的類似度である結合 E-value を算出するために，ドメインごとの E-value を用いる．

HMMER による相同性検索の結果，検出できたドメイン数が 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 $E_i(s)$ が独立の時，その同時確率 $Z_n(s)$ を次のように定義する．

$$Z_n(s) = \prod_{i=1}^{n} \frac{E_i(s)}{DBsize} \tag{1}$$

ここで $DBsize$ は検索に用いたデータベースのサイズである．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 $F_n(s)$ を次式で計算する．

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

$F_n(s)$ に $DBsize$ を掛けることで結合 E-value: $E(s)$ を求める．

$$E(s) = F_n(s) \times DBsize \tag{3}$$

HMMER における E-value と同じく，$E(s)$ は検索に用いたデータベース中から誤って相同だと判断されるタンパク質の数の期待値であり，値が小さいほど統計的に有意であるといえる．結合 E-value で昇順にソートし，最終的な検索結果を出力する．

The improvement of MDHMMER for mutated Multi-Domain proteins

Hirokazu HIRATA*, Tohgoroh MATSUI*, and Hayato OHWADA*

Department of Industrial Administration, Faculty of Science and Technology, Tokyo University of Science*

$F_n(s)$ の式は，次のように求めることができる．

$X_1,\cdots,X_n$ が $\overline{0,1}$ 上の一様分布に従う互いに独立な n 個の変数の時，積 $Z_n = X_1X_2\cdots X_n$ の分布関数を求めることを考える．$Y_i = -\ln X_i$ とおく時，$t > 0$ に対して

$$Pr\{Y_i \geq t\} = Pr\{X_i \leq e^{-t}\} = e^{-t} \tag{4}$$

となる．

ここで Feller [3] が示した次の定理を用いる．$Y_1,\cdots,Y_n$ が指数分布を持つ互いに独立な確率変数ならば，和 $S_n = Y_1 + \cdots + Y_n$ は

$$G_n(x) = 1 - e^{-x}\sum_{i=0}^{n-1}\frac{x^i}{i!}, \quad x > 0 \tag{5}$$

で与えられる分布関数 $G_n(x)$ を持つ．

$Z_n = e^{-S_n}$ の分布関数 $U_n(t)$ は，$0 < t < 1$ として $U_n(t) = 1 - G_n(-\ln t)$ で与えられる．

この定理に対し，本研究では $\frac{E_i(s)}{DBsize}$ が $\overline{0,1}$ 上の一様分布に従う変数であると仮定し，

$$Y_i = -\ln\frac{E_i(s)}{DBsize} \tag{6}$$

とおくことで $F_n(s)$ を求めている．

3 実験

3.1 実験方法

本実験ではマルチドメインタンパク質に含まれるドメインのうち，1つのドメインが大きく変異した場合の MDHMMER の検索精度を調べることを目的とした．また，人工的に生成した仮想のシミュレーションデータセットを実験に用いた．データセットの構成については以下の通りである．

3.1.1 データベース

データベースはランダムに生成される配列長 1000 のタンパク質 30000 セットにより構成した．

3.1.2 クエリー配列

まずデータベース中からタンパク質を一つ選択し，マルチドメインタンパク質であると仮定した．選択したタンパク質はドメイン数 3，各ドメインの配列長 20，ドメイン間の配列長を 30 とした．このタンパク質が検索における「正事例」となる．次に保存率に基づいてアミノ酸配列を変異させた．保存率とはアミノ酸配列のアミノ酸を変異させるとき，変異せずにそのまま保存されて残るアミノ酸の割合である．選択したタンパク質のドメイン部分のについて，2つのドメインについては保存率 50 %で変異させた．残りの 1 つのドメインを保存率 p (10 %〜50 %) でアミノ酸を変異させた．また，このドメインをドメイン X と呼ぶことにした．ドメイン間の領域については保存率 0 %とし，完全にランダムな配列に置き換えた．また，この時ドメイン間の間隔を変化率 150 %で変化させた．

生成されたタンパク質の配列全体に対して保存率 80 %でアミノ酸を変異させた．変異させる位置をランダムに変化させながら 30 回繰り返し，ドメイン配列群を生成した．

MDHMMER は生成された配列群よりドメイン領域に当たる配列を取り出し，ドメイン配列群を 3 つをクエリーとした．

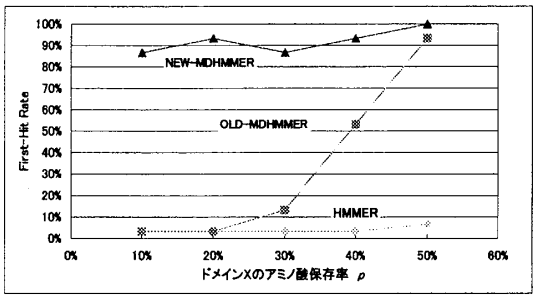

図 2 ドメイン X のアミノ酸保存率による検索精度の変化

一方，HMMER では生成されたドメイン配列群をそのまま 1 つのクエリーとした．

3.1.3 評価指標

検索精度の測定には First-Hit rate を用いた．First-Hit rate はデータベースへの検索実験を 30 回行った結果，正事例タンパク質が一番目に検出された割合を示したものである．

3.2 結果

シミュレーションデータセットを用いた実験結果を図 2 に示した．改良前の MDHMMER を「OLD-MDHMMER」，改良後の MDHMMER を「NEW-MDHMMER」とした．

4 考察

図 2 から，HMMER の検索精度が一貫して低いことがわかる．これはドメイン間の配列長が変化したことによる影響だと考えられる．よってドメイン間の配列を考慮せず，ドメイン部分のみに注目した相同性検索が可能であるという従来の MDHMMER の長所を維持できていると言える．また，ドメイン部分のアミノ酸保存率が低くなっていくと，改良前の MDHMMER は検索精度が低下しているのに対し，改良後の MDHMMER は高い水準の検索精度を保てていることが分かる．従って，改良後の MDHMMER では 1 つのドメインが大きく変異し検出できなかったとしても，残りのドメインの相同性検索結果を用いてマルチドメインタンパク質を検出することが可能となった．

しかし，ドメイン部分の検出は HMMER の検索精度に依存している．そのため大きく変異したドメインを多く持つマルチドメインタンパク質の場合，少数のドメインから結合 E-value を算出しなければならなくなり，精度の高い検索を行うことは難しくなることが考えられる．従って個々のドメインの検索精度をいかに向上させていくかが今後の課題である．

参考文献

[1] 瀬下真吾. マルチドメインを持つ遠縁なタンパク質のための相同性検索ツール. *FIT2006*, pp. 153–156, 2006.

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

[3] W.Feller. An Introduction to Probability Theory and its Applications. Vol. 2, 1957.