2V-1

シミュレーションデータセットを用いた実験による Multi-Domain HMMsearch の評価

瀬下 真吾* 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‡ 同 ゲノム創薬研究センター§

1 はじめに

バイオインフォマティクスにおいて，タンパク質のドメインに注目して相同性検索を行うことは，遠縁な相同タンパク質の発見に有効である．ドメインとは機能の発現に必要な領域であり，進化の過程でもアミノ酸配列が保存されやすい特徴を持つ．そのため，同一ファミリーに属するタンパク質は共通のドメインを持つことが多い．また，複数のドメインを持つタンパク質が存在し，複数のドメインを持つこと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．

相同性検索ツールとしては HMMER [1] が広く利用されている．しかし，HMMER では単一の類似領域しか考慮できないため，マルチドメインタンパク質を対象とした場合，ドメイン間の配列も比較対象とされてしまい，ドメインのみに注目した検索ができないという問題がある．

そこで，我々は HMMER の問題点を解決するために，ドメインごとに類似度を計算し，それぞれのドメインの類似度から統合的な類似度を求める方法を提案した [2]．また，提案手法に基づいて相同性検索ツールを実装した．このツールを **Multi-Domain HMMsearch** と呼ぶ．

本論文では，Multi-Domain HMMsearch の検索精度を評価するために，シミュレーションデータセットを作成して行った実験の結果を示す．また，HMMER と比較を行って，マルチドメインタンパク質に対し我々の提案手法が有効であることを示す．

2 Mulit-Domain HMMsearch について

Mulit-Domain HMMsearch 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．入力は PROSITE 等のアミノ酸データベースから入手した各ドメインのアミノ酸配列群であり，これが検索のクエリーとなる．データベースはあらかじめ用意されたアミノ酸データベースの中からユーザーが検索の際に選択する．検索が実行されると，ドメインごとの類似度を HMMER によって求め，全てのドメインの類似度から統合的な評価値である**結合 E-value** を計算する．結合 E-value とは，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．ドメインの位置情報も考慮に入れた上で結合 E-value で昇順にソートし，最終的な検索結果を出力する．

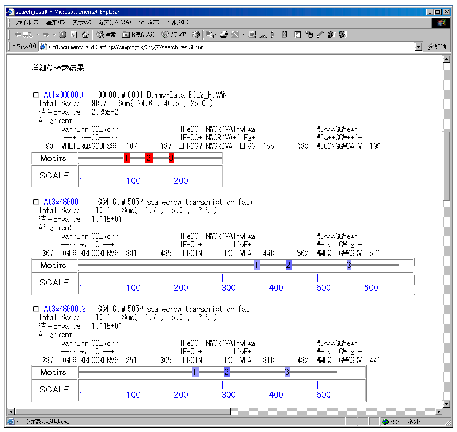

図 1 Multi-Domain HMMsearch による検索の実行例

図 1 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，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．

3 実験手法

3.1 シミュレーションデータセットの生成法

Mulit-Domain HMMsearch の検索精度を評価するために，仮想のマルチドメインタンパク質を含むシミュレーションデータセットを作製した．

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 セットにより構成されており，検索のクエリーとなるドメイン配列群は図 2 のような流れで生成した．

STEP1 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する．本実験ではドメイン数 3，各ドメインの配列長 20，ドメイン間の配

Evaluation of Multi-Domain HMMsearch by experiment that uses simulation data set

Shingo SEJIMO*, Tohgoroh MATSUI†, Hayato OHWADA†, Hidetaka KAYA§, 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 §,

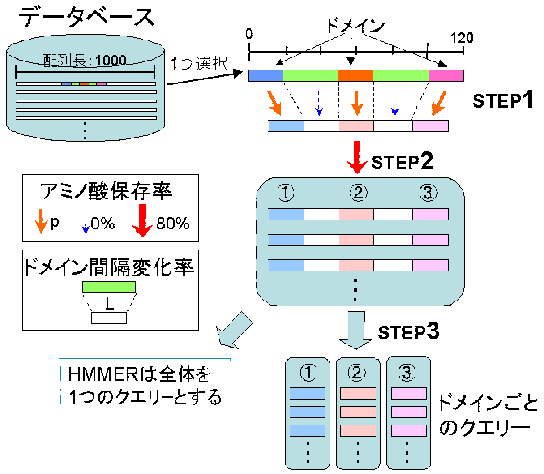

図 2　シミュレーションデータ生成の流れ

列長を 30 とする．このタンパク質が検索における「正事例」となる．選択したタンパク質のドメイン部分について保存率 p でアミノ酸を変異させ，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る．また，この時ドメイン間の間隔を変化率 L で変化させる．

STEP2　STEP1 で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る．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する．

STEP3　Multi-Domain HMMsearch はドメインごとのクエリーが必要なので，STEP2 で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 つ得る．一方，HMMER では STEP2 で生成されたドメイン配列群をそのまま 1 つのクエリーとする．

データセットにおいて，データベースは植物のシロイヌナズナに，ドメインの数とドメインの間隔はアポトーシス関連因子のひとつである Bcl-2 ファミリーに基づいている．Bcl-2 ファミリーに属するタンパク質は BH1，BH2，BH3 の 3 つのドメインを持つことで知られている．

3.2　検索精度の測定法

検索精度の測定基準には First-Hit rate を用いる．First-Hit rate はデータベースへの検索の結果，正事例タンパク質を一番目に検出することのできた割合を示す．本実験では 2 つのパラメータ p と L をそれぞれ 5 通り変化させ，合計 25 通りの組み合わせ条件で検索を行う．ランダム性を考慮し各条件において 100 回の検索を行い，検索精度の割合を計算する．

4　結果

前章で作製したシミュレーションデータセットを用いた実験結果を表 1 に示す．表からも分かるとおり実験した 25 通り全ての条件において Multi-Domain HMMsearch は HMMER よりも高い精度となった．

表 1　シミュレーションデータセットにおける検索精度（First-Hit rate）の結果

		L									
		-40%		-20%		± 0%		+20%		+40%	
		mdh	hmm	mdh	hmm	mdh	hmm	mdh	hmm	mdh	hmm
p	50%	**0.98**	0.04	**1.00**	0.47	**1.00**	**0.99**	**0.99**	0.73	**1.00**	0.18
	45%	**0.99**	0.03	**0.99**	0.23	**0.99**	**0.92**	**0.98**	0.36	**0.97**	0.00
	40%	**0.94**	0.00	0.85	0.04	0.89	0.57	0.85	0.08	**0.90**	0.00
	35%	0.52	0.00	0.39	0.00	0.58	0.24	0.58	0.01	0.48	0.00
	30%	0.13	0.00	0.13	0.00	0.13	0.12	0.16	0.00	0.12	0.00

mdh：Multi-Domain HMMsearch，hmm：HMMER

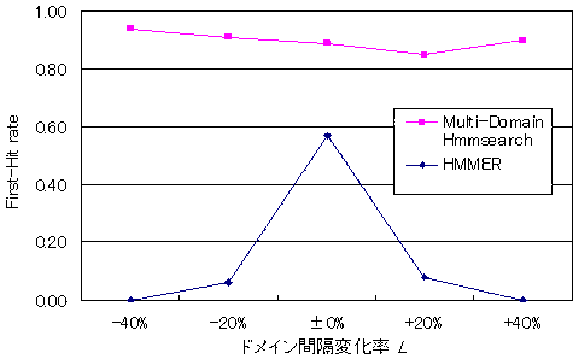

図 3　ドメイン間隔変化率 L による First-Hit rate の変化（$p = 40$）

パラメータ L が精度に及ぼす影響を見るために，$p = 40$ における First-Hit rate の変化を図 3 に示す．Multi-Domain HMMsearh では L による影響が見られないが，HMMER では $L = \pm 0$ % を中心にして，変化率が大きいほど精度が下がっていることが分かる．これはドメイン間の配列も比較対象としてしまうという HMMER の問題点によるものだと考えられる．ドメイン間の配列はランダムな配列のため，クエリーとの類似性が低くなり評価値が下がってしまうのである．この L が精度に及ぼす影響は，異なる p を基準にした場合においても同様の結果が見られた．

5　結論

本論文ではマルチドメインタンパク質に対する Multi-Domain HMMsearch の検索精度を評価するために，シミュレーションデータセットを作製し，実験を行った．

ドメイン部分の保存率 40 %，ドメイン間隔変化率 +40 %の時，HMMER では全く検出できず精度が 0 %であったのに対し，Multi-Domain HMMsearch は 94 %の精度であった．他の条件においても HMMER よりも高い精度で検索可能であることが確かめられ，マルチドメインタンパク質の検索に対し我々の提案手法が有効であることが示された．

参考文献

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

[2] 瀬下真吾. マルチドメインを持つ遠縁な相同タンパク質の検出手法. *FIT2005*, pp. 381–382, 2005.