H_017

HMMプロファイルの類似性に着目した遠縁の相同体検出手法の提案
A remote homologue detection technique with similarity of HMM profile

河村 真平* 賀屋 秀隆§ 松井 藤五郎† 朽津 和幸‡§ 大和田 勇人†

1. はじめに

近年，バイオインフォマティクス（Bioinformatics：生命情報科学）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．

生物学において相同性（ホモロジー，Homology）は，ある遺伝子や形態が共通の祖先をもつことを意味し，相同性を持つタンパク質の事を相同体と呼ぶ．バイオインフォマティクスでは，タンパク質の相同性は配列類似性に基づいて判断される．例えば2つの遺伝子がほとんど同一のDNA配列をもっている場合，それらはおそらく相同であろうと考えられる．

生物学的な機能が同じで共通の祖先を持つが，進化の過程により，タンパク質のアミノ酸配列が大幅に変わってしまった相同体を遠縁な相同体と呼ぶ．遠縁の相同体を予測し，検出するには，BLAST[1]等の従来の検索手法ではアミノ酸配列が大幅に変わってしまっているため，目的のタンパク質ではなくそれより局所的に類似度が高いタンパク質が検出されてしまう．そのため，目的のタンパク質，すなわち遠縁の相同体を検出出来ないという問題点が有る．

そこで本研究では遠縁な関係に有る相同体を，直接クエリー配列をデータベースに対して検索をかけるのではなく，目的のタンパク質の生物と同じ生物のタンパク質をクエリーとして得るために，その間にHHpredによる検索を入れ，遠縁の相同体を検出する手法を提案する．

2. HMMpredによるHMMプロファイルの検索

HMMとは，隠れマルコフモデル（Hidden Markov Model）の事である．現在の状態に依存して，次の状態が決定するような確率過程をマルコフ連鎖という．あるアミノ酸が現れたとき，次にどのアミノ酸が現れるかは，状態間の遷移確率によって決まる．このことに基づき，アライメント中の各位置の状態を20種類のアミノ酸の出現確率として表現したものを，隠れマルコフモデルと呼ぶ[2].

Johannes Söding らによって開発された HHpred[3]は，手持ちのマルチプルアライメントから HMM プロファイルを作成し，その HMM プロファイルと類似度の高い HMM プロファイルを Pfam，PROSITE 等のドメイン配列群データベースから検出する事の出来るプログラムである．二本のアミノ酸配列をアライメントするように，二つの HMM プロファイルをアライメントし，スコアを算出する事によって類似度を比較するという手順を踏んでいる．[4]

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 理工学部 応用生物科学科‡
同 理工学部 ゲノム創薬研究センター§

3. 提案手法

提案手法の概要は以下の通りである．

1．機能が判明している生物Aのアミノ酸が所属するファミリーのドメイン配列群を取得する．

2．取得したドメインの配列群と，類似度が高いドメインの配列群を持つファミリーを HHpred を用いて取得する．

3．HHpred で検出したドメイン配列群からホモロジー検索の場合なら生物Aのタンパク質を，モチーフ検索の場合なら生物Aのタンパク質を含むプロファイルを探し出す．

4．手順3で取得したタンパク質，又はプロファイルを用い，それぞれ対応したツールを用い，生物Bのタンパク質データベースに対して検索を行う．

図1は以上の手順を示している．

本研究は遠縁な相同体を検索する場合に，クエリー配列やプロファイルを，直接ホモロジー検索ツール BLAST やモチーフ検索ツール HMMER を用いデータベースに対して検索をかけるのではなく，途中で HHpred を用いる事によって，より効果的に遠縁な相同体を検出する手法を提案する．遠縁な関係に有る生物Aと生物Bを例に挙げて一連の流れを説明する．まず機能が判明している生物Aのアミノ酸が所属するファミリーのドメイン配列群を取得する．ドメイン配列群は WEB 上に公開されているデータベースから取得する．次に，データベースから取得したドメインの配列群と，類似度が高いドメインの配列群を持つファミリーを HHpred を用いて取得する．そして最後に HHpred で検出したドメイン配列群からホモロジー検索の場合なら生物 A のタンパク質を，モチーフ検索の場合なら生物 A のタンパク質を含むプロファイルを探し出し，BLAST を用いて生物Bのタンパク質データベースに対して各種検索を行う．このような流れで生物Aのタンパク質，もしくはプロファイルの代わりに，生物Aの所属するドメイン配列群と類似度の高いドメイン配列群内の生物Bの配列をクエリーとして用いる事が可能となる．

生物Bのタンパク質を検索する際に遠縁の生物Aのタンパク質，もしくは生物Bのタンパク質含んでいないプロファイルをクエリとして直接用いるのではなく，生物Aのタンパク質と類似したドメインを持った生物Bのタンパク質，もしくは生物Bのタンパク質を含んだドメイン配列群を検索に用いることで，今まで検出する事の出来なかった目的のタンパク質を検出する事が可能になる．

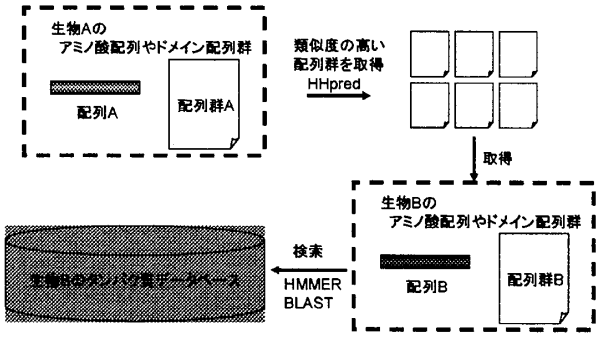

図1 提案手法の概要

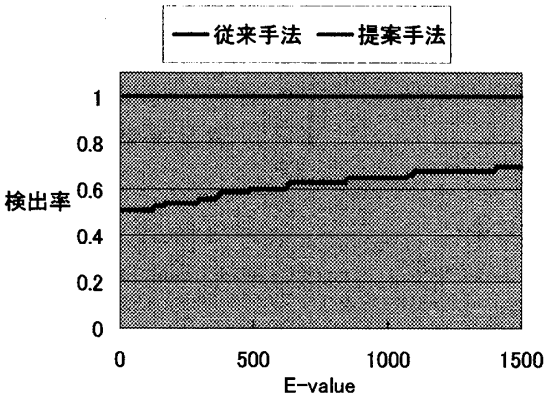

図2 E-valueと検出率の推移

4. 実験

本実験では，あらかじめ目標とするタンパク質を設定し，従来手法と提案手法ではそのタンパク質をどれだけ正確に検出出来るかを比較する．

4.1 実験手法

本論文で提案した手法を，実際にホモロジー検索ツールBLASTを使って実験を行った．本実験ではPfamに登録してあるドメイン配列群 zf-C3HC4 からシロイヌナズナのアミノ酸配列を抜き取った物を従来手法のクエリーとして用いた．

今回の実験では，抜き取ったシロイヌナズナのアミノ酸配列の内の一つである COP1_ARATH を目標のタンパク質とした．又，zf-C3HC4 を Pfam に対して HHpred にかけた時に，一番類似度が高い配列群として U-box が検出出来た．U-box 内にシロイヌナズナのタンパク質を 3 本発見する事が出来たので，これらを提案手法のクエリーとして用いた．

シロイヌナズナ以外の zf-C3HC4 内のタンパク質一本一本をクエリーとして，シロイヌナズナのタンパク質データベースに対して，BLAST で検索をかけた場合の COP1_ARATH の発見率と，U-box 内の 3 本のシロイヌナズナのタンパク質を，それぞれシロイヌナズナのタンパク質データベースに対して BLAST で検索をかけた場合の COP1_ARATH の発見率とを比較する事によって，提案手法の有効性を示す．

本実験では BLAST の E-value を 10 から 1500 まで 10 刻みで設定して実験を行った．E-value とは互いに関係がないのにも関わらずスコアの高い配列が，偶然にヒットするという可能性を数値で表したものである．検索結果の E-value が低い値であるほど，その結果は偶然でない確かな一致であるとみなすことができ，統計的に有意であると言える．E-value を上げれば検出出来る配列の数は増やす事が出来るが，上げれば上げるほど統計的な信頼性は無くなっていく．

検出率は COP1_ARATH を検出出来た配列の数を，全クエリー配列の数で割ることで算出した．例えば提案手法の場合だと zf-C3HC4 内には 65 本のタンパク質配列が含まれているので，この中の 32 本が BLAST により COP1_ARATH を検出できた場合は 32/65 より約 50%と計算する．

4.2 実験結果

実験結果は図2に示す．どの E-value でも従来手法より HH-pred を途中で用いた提案手法の方がより正確に目的のタンパク質を検出出来ている．

5. 考察

TreeView[5]を用いて，クエリーとして用いた U-box 内の 3 本のシロイヌナズナのタンパク質と，検索目標の zf-C3HC4 内のシロイヌナズナのタンパク質 COP1_ARATH を比較した．ドメイン配列群 zf-C3HC4 の中に COP1_ARATH を入れて，TreeView を用いて系統樹を作成した．その結果，COP1_ARATH と U-box 内の 3 本のシロイヌナズナのタンパク質は，COP1_ARATH が含まれているドメイン配列群 zf-C3HC4 内のどのタンパク質より，近縁であることが分かった．これより，HHpred を用いた提案手法により獲得できた 3 本のタンパク質は，COP1_ARATH を検索するためのクエリーとして有効なものであったと考えられる．

6. まとめ

本論文では遠縁の相同体を検出する為に，各種検索を直接行う前に HH-pred により類似度の高いドメイン配列群を検出し，クエリーに用いることによって従来の手法では検出することの出来なかったタンパク質を検出する手法を提案した．又，その有用性も実験により証明出来た．

本研究では HH-pred から検出したドメイン配列群に対象となる生物のタンパク質が含まれていない場合，それらの事を考慮しない方針をとった．この点については今後検討する．

参考文献

[1]Altschul S.F., Gish W., Miller W., Myers E.W.,and Lipman D.J. 1990.Basic local alignment search tool.J.Mol.Biol.215:403-410.

[2]Application of Bioinformatics. http://wwwitblpg.apr.jaeri.go.jp/itblpg/bioedu/index.html

[3]HHpred.http://protevo.eb.tuebingen.mpg.de/toolkit/index.php?view=hhpred

[4]Johannes Söding Protein homology detection by HMM-HMM comparison.Bioinformatics **21**, 951-960 2005.

[5]TreeView:http://taxonomy.zoology.gla.ac.uk/rod/treeview.html