1W-3

相同性検索手法の組み合わせによる検索精度向上

滝沢雅俊† 山田真介‡ 山名早人†

†早稲田大学理工学部情報学科 ‡早稲田大学大学院理工学研究科

１．はじめに

タンパク質のアミノ酸配列比較の際、相同性は重要な基準となる。相同性とは、共通祖先に由来する子孫間の類似性を指す。共通祖先から分岐した相同タンパク質の間では、類似した構造や機能を有していることが多い。また、機能や構造未知のアミノ酸配列を問い合わせ配列とし、データベース中の既知のアミノ酸配列から、問い合わせ配列と相同なアミノ酸配列を収集する方法を相同性検索とよぶ。現在までに、様々な相同性検索手法が開発されてきたが、相同配列の検出精度を更に向上させることが求められている。

そこで、本研究では、BLAST、FASTA、SSEARCH、WU-BLAST、SCANPS、PatternHunter といった合計６つの相同性検索手法について２手法を組み合わせることにより、検索精度の向上を図る。

２．組み合わせについての概要

２．１．相同についての定義

本研究では、タンパク質構造分類データベース SCOP[1]の分類に従い、相同性を定義する。SCOP ではタンパク質の立体構造に対し、構造的、進化的類似を描写するために、ファミリー、スーパーファミリー、フォールド、クラスといった階層的な分類を用いている。同一ファミリーとして分類されるタンパク質群は配列一致度が高く、機能や立体構造が良く保存されている。また、同一スーパーファミリーとして分類されるタンパク質群は配列一致度が低いが、構造と機能から考えると相同である。

本研究では、同一ファミリーであるか、ファミリーが異なるもののスーパーファミリーが同一であるタンパク質ペアを相同であると定義する。

Increased sensitivity and specificity obtained by combination of methods for protein sequence database searching

Masatoshi Takizawa†, Shinsuke Yamada‡, Hayato Yamana†

†Department of Information and Computer Science, School of Science and Engineering, Waseda Univ.

‡Graduate School of Science and Engineering, Waseda Univ.

２．２．相同性検索手法

BLAST、FASTA、WU-BLAST、PatternHunter、SCANPS、SSEARCH といった合計６つの相同性検索手法をデフォルトのパラメータで用いることによって組み合わせを行う。ただし、検出精度を上げるために、FASTA の k-tuple を 1 に、SCANPS のギャップペナルティをアフィンギャップに、PatternHunter の spaced seed を 4 に設定したうえで組み合わせを行う。

２．３．手法の組み合わせ方法

相同性検索を行い、タンパク質ペアの E-value について閾値を定め、閾値以下の E-value のタンパク質ペアを出力する。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る。

本研究では、２手法を組み合わせることを考えているが、union と intersection という２つの操作が基本となる[2]。union 操作とは、各手法において閾値以上の配列ペア全てを出力する操作を指す。union 操作によって、specificity（相同性検索手法が出力した全てのペアに対する相同性検索手法が出力した相同なペアの比率）は低下する可能性があるが、sensitivity（データセットに含まれる全ての相同なペアに対する相同性検索手法が出力した相同なペアの比率）が向上する。また、intersection 操作とは、両手法ともに閾値以上の配列ペアに限って出力する操作を指す。intersection 操作によって、sensitivity は低下するが、specificity が向上する。

２．４．提案手法

本提案では、タンパク質ペアの E-value に応じて union、intersection を使い分け、sensitivity、specificity 両方の向上を目指す。まず、各相同性検索手法の specificity が高い、E-value が 1.0*E-3 以下のペア同士を組み合わせる場合に、union 操作を行う。これによって、specificity の低下を抑え、sensitivity の向上を図る。さらに、各手法の specificity が低い、E-value が 1.0*E-3 よりも大きいペア同士を組み合わせる場合に、intersection 操作を行う。これによって、specificity 低下の抑制を図る。

3．実験結果

3．1．データセット

タンパク質構造分類データベース ASTRAL[3]から入手した、sequence identity が 40%未満の配列データセットを用いて実験を行った。このデータセットには 5674 本の配列が含まれている。その全配列ペアのうち相同な配列ペアは 58853 ペアである。

3．2．手法単独で用いた結果

図１は、E-value 閾値が 1.0*E-3 において相同性検索手法単独で用いた結果である。E-value 閾値が 1.0*E-3 においては、各手法の specificity が非常に高いことが確認された。

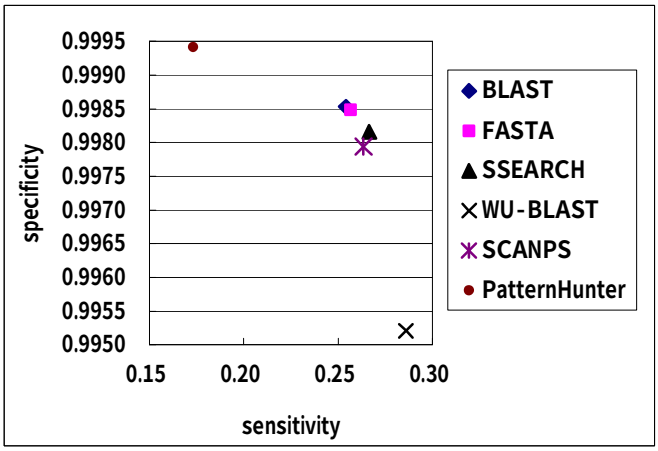


図１ 手法単独で用いた結果（E-value 閾値 1.0*E-3）

3．2．相同性検索手法の組み合わせ結果

組み合わせを行った結果、E-value 閾値が 1.0*E-3 以上の近い範囲では sensitivity、specificity ともに親の手法に比べて向上させる組み合わせの存在が確認できた。表１は E-value 閾値 3.0*E-3 において、sensitivity、specificity 共に向上した組み合わせ結果である。E-value 閾値 3.0*E-3 の場合、BLAST と FASTA の組み合わせで sensitivity、specificity 共に両方の親の手法に比べて向上させることができた。また、表 1 に示す 3 通りの組み合わせで sensitivity、specificity ともに片方の親の手法に比べて向上させることができた。

表１ sensitivity、specificity 共に向上した組み合わせ結果（E-value 閾値 3.0*E-3）

Method A	Method B	Sensitivity %increase over parent methods	Specificity %increase over parent methods
BLAST	FASTA	3.0 / 0.47	0.052 / 0.0011
BLAST	SSEARCH	4.4 / -1.6	0.043 / 0.014
BLAST	SCANPS	4.4 / -0.86	0.018 / 0.096
FASTA	SSEARCH	1.7 / -1.8	0.0038 / 0.026

また、図２は E-value 閾値 3.0*E-3 における BLAST と FASTA の組み合わせ結果である。

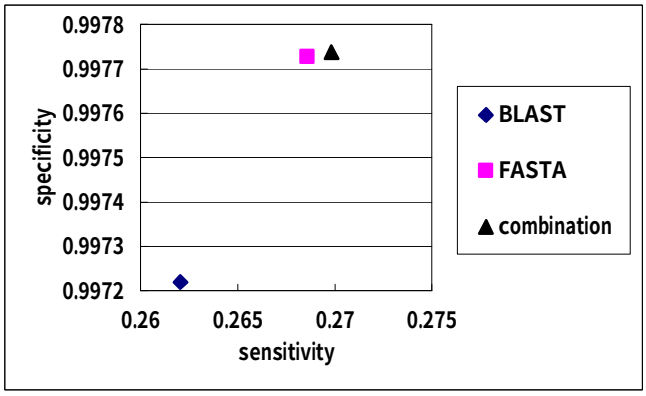


図２ BLAST と FASTA の組み合わせ結果（E-value 閾値が 3.0*E-3）

4．考察

手法単独で用いる場合と比較して sensitivity や specificity が向上した組み合わせが存在する背景には、タンパク質のアミノ酸配列のアラインメントを行うアルゴリズムや E-value の計算方法が各手法で異なるということが考えられる。また、組み合わせを行っても、sensitivity を大幅に向上するには至らなかったことから、各手法が拾ってくる相同なタンパク質ペアは似通っていると考えられ、ペアワイズアラインメントで相同性検索を行う点に限界を感じる。

今後は、手法の組み合わせによって通常の Gapped-BLAST より多くの配列を拾うことで、PSI-BLAST のプロファイルの質や PSI-BLAST の精度にどのような影響を及ぼすのか検証を行う予定である。

参考文献

[1] Murzin A.G., Brenner S.E., Hubbard T., Chothia C.: SCOP: a structural classification of proteins database for the investigation of sequences and structures. J. Mol. Biol., Vol.247, pp.536-540(1995)

[2] Caleb Webber, Geoffrey J.Barton: Increased coverage obtained by combination of methods for protein sequence database searching, Bioinfomatics, vol.19, pp.1397-1403(2003)

[3] Brenner SE, Koehl P, Levitt M.: The ASTRAL compendium for sequence and structure analysis. Nucleic Acids Research , vol.28, pp.254-256 (2000)