論文の内容の要旨

論文題目

Development of methods for large-scale bioinformatics analysis
of protein sequences and structures and their applications
（蛋白質配列及び立体構造情報の大規模解析手法の開発とその応用）

氏名　　中村　司

本研究は，蛋白質の配列情報及び構造情報から，蛋白質の機能を明らかにすることを目的として行った．

1. 蛋白質複合体予測手法の高度化とその効果の検証

いわゆる次世代シーケンサーの登場を背景とし，蛋白質配列情報は爆発的な速度で増大している．一方，蛋白質立体構造情報は，生物学的実験を通しての決定の困難さや技術的限界などにより，その進展速度は大きく乖離しているのが現状である．このギャップを埋める手立てとして，蛋白質立体構造の予測は非常に重要であると考えられている(Levitt *PNAS*, 2009)．しかし蛋白質立体構造の予測は，単体構造の予測においても未だ発展途上であり，複合体の予測に関してはその問題の複雑性から更に困難である．その一方で，蛋白質は実際に機能するとき，しばしば生物学的集合体と呼ばれる，特定の数と配置を持った複合体をとると考えられている．そのため，これらの蛋白質の機能に対する理解を深めるためには，生物学的集合体構造の解明が必要となる．

既知の蛋白質構造をテンプレート（立体構造モデルの鋳型）として利用する，テンプレートベースの立体構造予測手法は，構造予測において幅広く用いられてきている手法である．また近年，配列情報，構造情報の増大とともに，より有用なアプローチとなってきている．生物学的集合体構造の予測に関しても，Protein Data Bank に登録されている複合体構造数の増加を踏まえると同様の状況であると考えられるものの，生物学的集合体構造の予測そのものは自明な問題ではない．テンプレートベースの立体構造予測手法の，生物学的集合体構造の予測における有用性を検証するために，この手法を用いて世界的な立体構造予測実験（出題される蛋白質に対する立体構造予測の精度を測定し予測技術の向上をはかる実験）である CASP に参加した．2014 年に複合体予測で優秀な成績を収めた[4]のに続き，2016 年の CASP12 では，複合体構造予測のカテゴリで，全参加チーム中 1 位の正確性を達成するという成績を収めた[2].

さらに本研究では，手法の有効性について遡及的な解析を行い，プロファイルプロファイルアラインメント(Tomii and Akiyama *Bioinformatics*, 2004)を行う際に使用するプロファイルを複数の異なる作成手法で

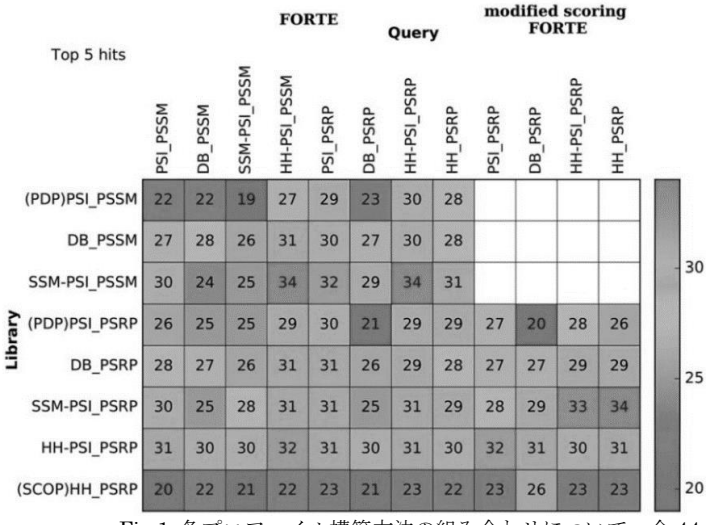

TABLE 1　Summary of methods used for profile construction

Abbreviations	Profile construction (DB, # iterations)
PSI_PSSM	(TM-align *only for library* +) PSI-BLASTexB (nr, 5)
DB_PSSM	DELTA-BLAST (CDD, 1)
SSM-PSI_PSSM	SSEARCH (nr) + MAFFT + PSI-BLASTexB (nr, 1)
HH-PSI_PSSM	HHblits (up20, 3) + PSI-BLASTexB (nr, 1)
PSI_PSRP	(TM-align *only for library* +) PSI-BLASTexB (nr, 5)
DB_PSRP	DELTA-BLAST (CDD, 1)
SSM-PSI_PSRP	SSEARCH (nr) + MAFFT + PSI-BLASTexB (nr, 1)
HH-PSI_PSRP	HHblits (up20, 3) + PSI-BLASTexB (nr, 1)
HH_PSRP	HHblits (up20, 3)

Fig.1 各プロファイル構築方法の組み合わせについて，全 44 ドメイン中何ドメインについて
”正しい”テンプレートを提示可能であったかを示した図.

構築した．それらすべての組み合わせについて，プロファイルライブラリ中に存在する可能性がある，テンプレートとなる既知立体構造を提示可能であったかを検証した．その結果，テンプレートを提示可能であったと考えられるほとんどすべてのターゲットについて，実際に提示可能であることが確認された．

2. MSA 構築法の大規模並列化と他手法との性能比較

蛋白質立体構造予測の高度化に関連して，マルチプルシーケンスアラインメント(MSA)の精度向上は非常に重要な課題である．例えば，適切なテンプレートの検索に用いるプロファイルの性能に，MSA の精度は非常に密接に関わっている．また，二次構造予測やコンタクト予測などの機械学習手法の入力としても MSA は非常に重要な役割を担っている． 世界的に有名な MSA 作成ソフトウェアの一つである MAFFT (Katoh and Toh *Bioinformatics*, 2010)の G-INS-1 オプションは，入力の配列本数が多い場合でも高い精度で MSA を作成できることが以前より報告されていた．しかし G-INS-1 はスレッド並列のみにしか対応しておらず，全入力配列対全入力配列のペアワイズアラインメントの計算に伴う時間計算量や，その後プログレッシブアラインメントで使用する結果の保持のために必要な空間計算量のために，現実的な解析には非実用的なものとなっていた．そのため，MSA の入力配列数が多い場合や配列長が長い場合には，精度を犠牲にした他のモードを用い現実的な時間で終わらせざるをえなかった．この問題点を克服するため，MPI を利用してクラスタ並列計算機に対応した MAFFT のオプション G-large-INS-1 を開発した．

G-large-INS-1 では，全入力配列対全入力配列のペアワイズアラインメントをクラスタ計算機上の使用可能な計算資源に逐次的に割り振り，計算上のこのステップが現実的な時間で実行可能になるようにした．また，その結果の保持のために，のちに使用される順序で共有ファイルシステムに書き出すことで，実メモリの使用量を減らすことで現実的なメモリ空間での計算が可能になるようにした．そして，二次構造予測手法 JPred4 (Drozdetskiy *et al. NAR*, 2015)及び QuanTest (Le *et al. Bioinformatics,* 2017)のデータセットを用いて，二次構造予測を通じた MSA の精度評価及び，計算時間のベンチマークを行い目的が達成されていることを確認した[1].

さらなる解析として，200 本以上の配列を用いて作成した MSA から，ランダムに選択された 200 本を複数回抜き出し（手法間では固定），その MSA から二次構造予測を行った結果について比較を行った(Fig. 2). その結果，Clustal Omega では MSA の精度そのものは配列本数の増加に伴って減少していることが確認された．これはプログレッシブツリーの作成に近似法を使っている MAFFT の FFT-NS-2 や Clustal Omega

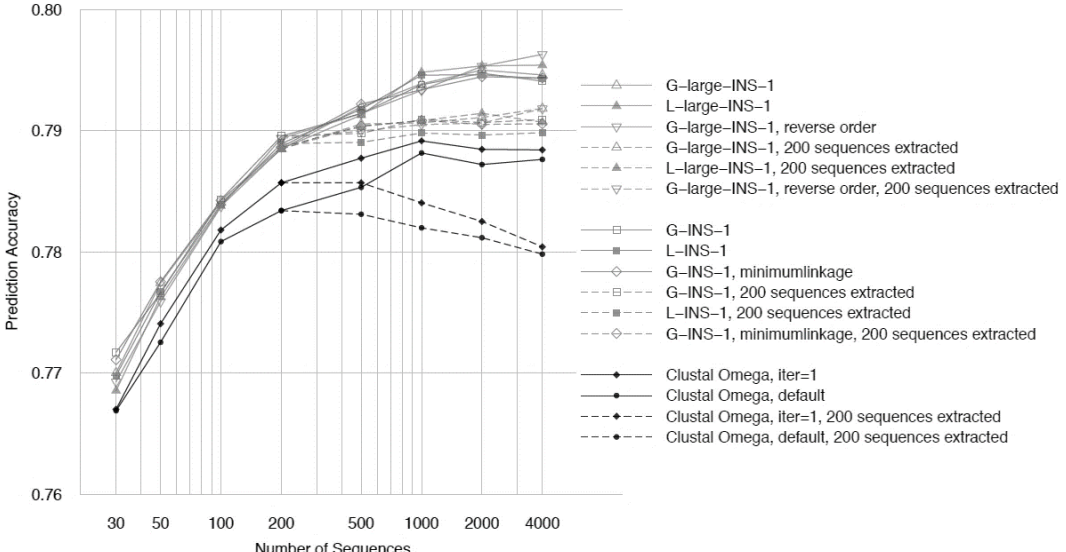

Fig.2 MSA 構築時の入力本数を変化させた時の各手法の二次構造予測精度．入力本数がアラインメント精度にのみ与える影響（点線)と，アラインメント精度と二次構造予測精度の両方に与える影響（実線）.

に共通して存在する問題であると(Sievers *et al. Bioinformatics*, 2013)などで以前より報告されていた問題点である．一方 G-large-INS-1 では配列本数の増加に伴う精度の劣化は起きていないことが確認された．これにより MSA 作成時にどの程度の本数を用いるべきかという問題から解放され，計算資源の上限まで配列を用いることでより良い MSA が作成可能であることが確認された．

3. 基質結合ポケットの粗視化表現としての，高性能なベクトル表現法の開発

蛋白質立体構造中におけるポケット、ここでは特に基質結合部位の情報を網羅的に解析し，その特性を明らかにすることで，結合基質の予測や創薬研究の基盤となる技術の開発が可能である．蛋白質の既知構造のデータベースである Protein Data Bank には，結合する低分子化合物が既知である，基質結合部位が約 35 万個登録されているとともに，それら蛋白質の既知構造に対して，ポケット同定プログラムを用いることにより約 620 万のポケット形状が推定可能であると報告されている(Ito *et al. NAR, 2015*)．これらを網羅的に解析し，基質結合部位に対する新たな知見を得るためには，高速に基質結合部位を比較するためのアルゴリズムが必要である．高速に基質結合部位を比較するために，基質結合部位を Cα 原子を頂点とする三角形の集合に分解し，事前に用意しておいた各三角形パターンの出現頻度を基質結合部位のベクトル表現とする方法がいくつか提案されている．しかしこれらの既存の方法には，

1. 出現頻度をカウントする三角形のパターン数を増加させた場合，頻度ベクトルがスパースになり，適切な類似度の計算が困難になるため，パターン数をあまり増やせず，表現力に欠ける．
2. 三角形のパターンに離散化する際, bin を超えて三角形パターン間の類似度を考慮しないため形状変形への対応が限られ，類似度を取り落とす可能性が存在する．

という問題点が存在する．本研究ではこれらの問題点を解決することで，既存の方法に比べて，より高性能な基質結合部位のベクトル表現法となるという仮説のもと，新規手法を開発した．

既存の問題点を解決するため，理論的に出現しうる全ての三角形パターンの間に類似度を定義し，この類似度行列を用いる形でポケット間の類似度を定義した．これにより，出現頻度をカウントする三角形のパタ

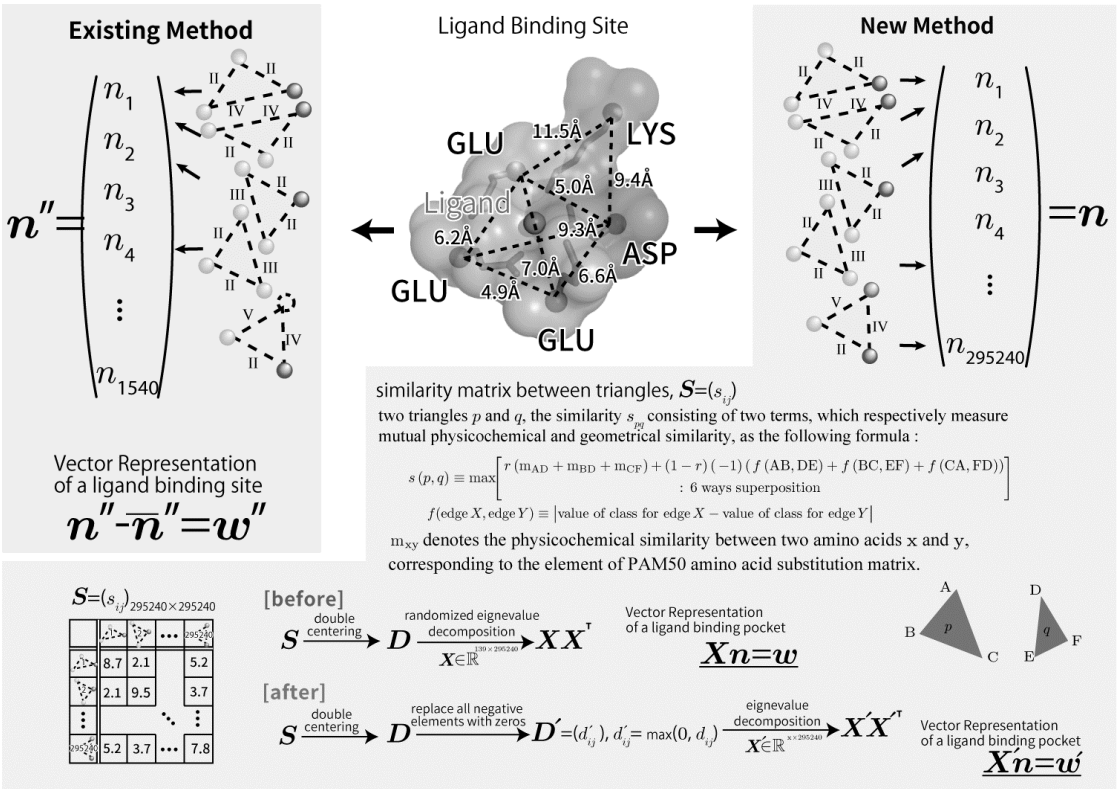

Fig.3 新規手法の概要.

ーン数を増加させた場合でも，頻度ベクトルがスパースになることによっておこる，どのポケットペアの類似度もほぼ 0 となる現象を抑えることが出来，パターン数を増加させることが可能となった．具体的には，既存手法では 1,540 種類(Ito *el al. Proteins*, 2012)でしか三角形パターンを扱えなかったのに対して，本研究では 295,240 種類で扱うことを可能にした（cf. 問題点 1). また，三角形パターン間の類似度を考慮して各三角形パターンの座標を得ることで，三角形間の類似度も考慮した（cf. 問題点 2）.

具体的には，三角形パターンの間の類似度を，アミノ酸間の類似度と，三角形同士の図形としての類似度を足し合わせた形の関数で定義した．アミノ酸間の類似度としては，一般に基質結合部位はより保存されているということを念頭に，配列類似性検索で用いられている PAM50 置換行列を用いた．この関数を用いて得られた，出現しうる全ての三角形パターンの間の類似度行列$\boldsymbol{S}$に対し，多次元尺度構成法を適用し，二重中心化を行い固有値分解することにより各三角形の高次元空間中での座標を得た（Fig. 3 中[before])[3,5]. 一方，二重中心化後の行列$\boldsymbol{D}$において負の要素をすべて 0 とし，スパースな形にした行列$\boldsymbol{D'}$を構成し，その後の演算を行う方法を[after]とした．そして，各三角形パターンの座標と，実際の基質結合部位より観測される三角形の，頻度のベクトルの線形和によって基質結合部位のベクトル表現を定義した．このベクトル表現間でのコサイン距離を計算することで基質結合部位間の類似度を計算するものと定義した.

三角形パターン間の類似度行列に対し多次元尺度構成法を行う中で，大規模な固有値分解を要する点に，計算量的困難が存在した．乱択アルゴリズムを使用した固有値分解法を用いることにより，多次元尺度構成法の計算を現実的な時間で可能になるよう解決した.

2 つのデータセットを用いて新規手法の性能を評価した．1 つは *Ito138* という，難易度の高いデータセットであり，これを用いて評価した結果を Fig. 4 (左)に ROC 曲線を用いて示した．新規手法は既存の 1,540 種類の三角形を用いる手法と比べ，高精度な基質結合部位間の類似度を得られた．また，*APocS3* という難易度の低いデータセットを用いて速度はあまり速くないものの，最適化問題を解き，高い精度で類似度を得る(即ち，構造アラインメントを用いる)既存手法，APoc と比較した．Fig. 4(右)に同様に示したように，新規手法は高精度であった.

既存手法に共通して存在する問題点を，新たなアルゴリズムを導入することで解決した．開発した新規手法が，仮説のように，既存の高速な手法に比べてより高性能な基質結合部位のベクトル表現法となっていることが確認された.

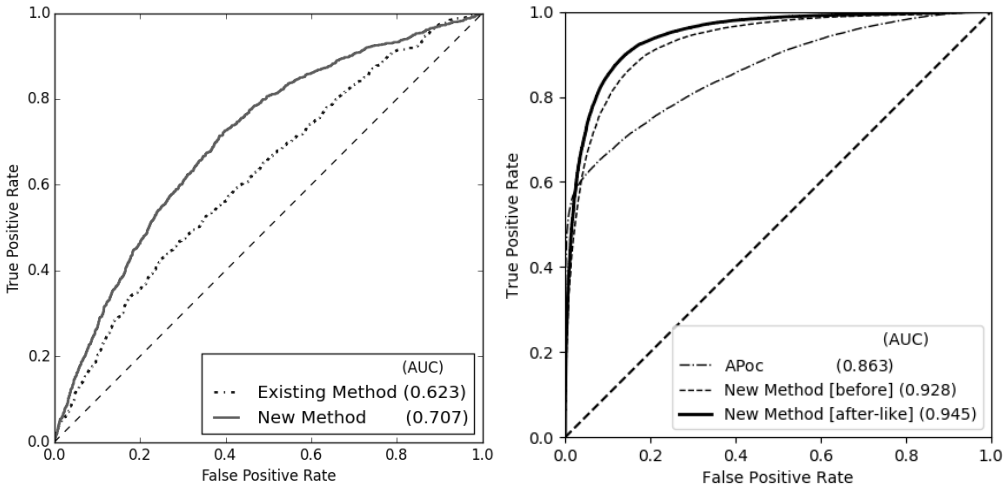

Fig.4 (左) *Ito138* を用いた，1,540 種類の三角形を用いる手法と，新規手法の性能比較.
(右) *APocS3* を用いた，APoc との性能比較.

・発表論文

1) Nakamura T, Yamada K D, Tomii K, Katoh K, *Bioinformatics*, 2018, 2) Nakamura T†, Oda T†, Fukasawa Y, Tomii K, *Proteins*, 2017 (†co-first author),
3) Nakamura T, Tomii K, *Biophysics and Physicobiology*, 2016, 4) Lensink M F, Velankar S, Nakamura T(73 番目), Tomii K, Wodak S J(97 名省略), *Proteins*, 2016,
5) Nakamura T, Tomii K, *Methods*, 2016