3Z-03

# 多次元尺度構成法を用いた 蛋白質ポケット部位の縮約ベクトル表現によるポケット構造比較

中村　司†　　富井　健太郎‡,†

東京大学大学院　新領域創成科学研究科　メディカル情報生命専攻†

産業技術総合研究所　生命工学領域　創薬基盤研究部門‡

## 1. はじめに

蛋白質立体構造中における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 PDB (Protein Data Bank) には，2014 年 9 月までに，結合する低分子化合物が既知である，基質結合部位が約 30 万個登録されている．また，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 520 万のポケット形状が推定されている [1]．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，三角形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[2]．

Structural comparison of protein pockets by a reduced vector representation derived from multidimensional scaling of generalized description of pockets

Tsukasa Nakamura†, Kentaro Tomii‡, †

†Department of Computational Biology and Medical Sciences, Graduate School of Frontier Sciences, The University of Tokyo

‡Biotechnology Research Institute for Drug Discovery, Department of Life Science and Biotechnology, National Institute of Advanced Industrial Science and Technology (AIST)

## 2. 方法

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，多次元尺度構成法を用いて空間中に配置した．これにより，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でも，基質結合部位の特徴ベクトルの次元数は，多次元尺度構成法の結果として得られる次元数(139 次元)に抑えることが出来，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 [1]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に対し，多次元尺度構成法を適用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．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得た (図 1)．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．

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[3]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．

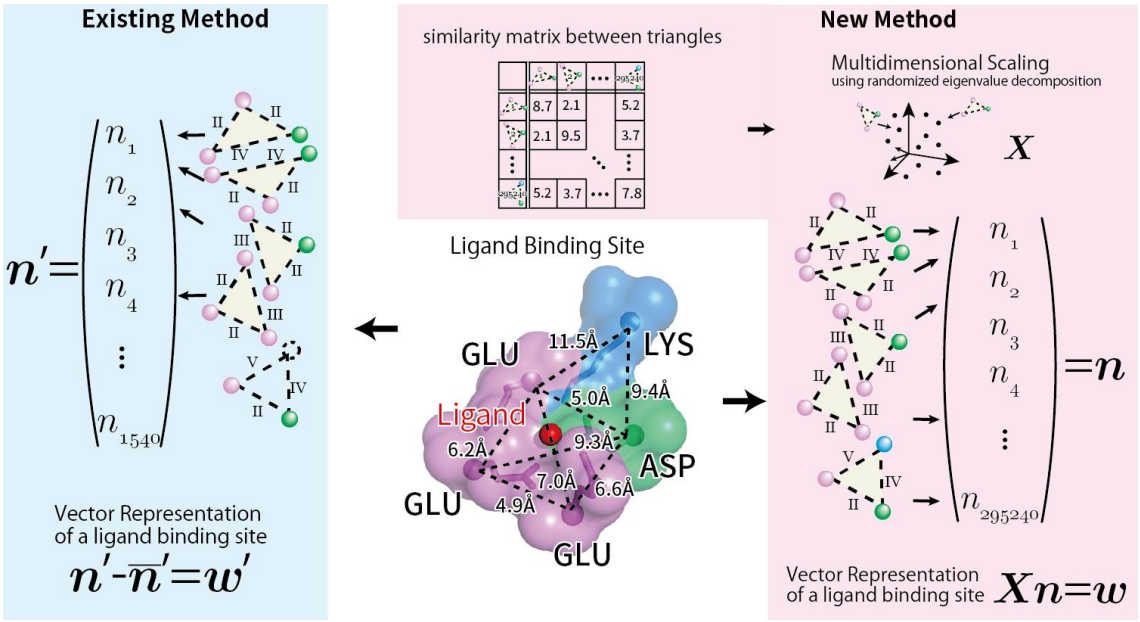


図 1 新規手法の概要

## 3. 結果と考察

2 つのデータセット[2]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図 2（上）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．図 2（下）に同様に示したように，新規手法は高精度であった．

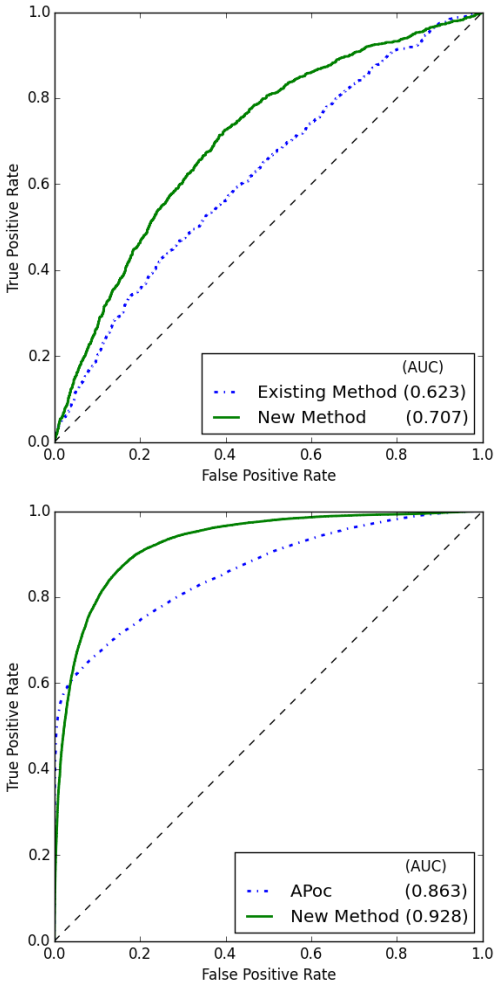


図 2（上）*Ito138* を用いた，1,540 種類の三角形を用いる手法と，新規手法の性能比較．（下）*APocS3* を用いた，APoc との性能比較．

## 4. まとめ

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．

参考文献
1) Ito, J.-I. et al.: PDB-scale analysis of known and putative ligand-binding sites with structural sketches, Proteins, Vol.80, No.3, pp.747–763 (2012).
2) Nakamura, T. and Tomii, K.: Protein ligand-binding site comparison by a reduced vector representation derived from multidimensional scaling of generalized description of binding sites, Methods, doi:10.1016/j.ymeth.2015.08.007.
3) Halko et al.: Finding Structure with Randomness: Probabilistic Algorithms for Constructing Approximate Matrix Decompositions, SIAM Review, Vol.53, No.2, pp.217–288 (2011).