Vol.2012-BIO-29 No.19
2012/6/29

# タンパク質間ドッキング予測における目的関数の機械学習を用いた動的調整

藤原　隆之[1]　松崎　由理[1]　石田　貴士[1]　秋山　泰[1]

**概要**：
タンパク質間ドッキング予測ソフトウェア “MEGADOCK” では，目的関数に形状相補性と静電相互作用の 2 つの項を用いているが，その最適なバランスは対象毎に一定ではなく，それを決定することは困難である．そのため，先行研究として予測精度改善のため目的関数のうち静電相互作用項の重みをタンパク質の表面電荷等の特徴から動的に調整する手法が提案されたが，いくつかの問題を含んでいた．そこで，本研究では従来手法の再検証を行い，サポートベクター回帰を用いた改良を提案する．改良された手法では従来使用されたデータセットにおいて予測性能の向上が確認され，その上で新たなデータセットへの適用も行った。

**キーワード**：タンパク質間ドッキング，機械学習，サポートベクター回帰

# Dynamic adjustment of the objective function for protein-protein docking prediction by means of machine learning

Fujiwara Takayuki[1]　Matsuzaki Yuri[1]　Ishida Takashi[1]　Akiyama Yutaka[1]

***Abstract:*** The protein-protein docking software “MEGADOCK” uses the two terms in its target function; shape complementarity and electrostatic. However, the optimal balance between those two terms is defferent for each protein. Thus, dynamic adjustment of the weight of the electrostatic term based on the surface charge of a protein was proposed in a previous work. In this work, we improved the method by using support vector regression and additional characteristics of a protein. By using our new method, we achieved the better prediction performance for the data used in the previous study. We also applied the method to new data set.

***Keywords:*** protein-protein docking, machine learning, support vector regression

## 1. 序論

タンパク質間相互作用 (Protein-Protein Interaction, PPI) は生命現象において中心的な役割を担っており，その解明は生命現象の解明につながると期待され，盛んな研究が行われている．計算機上で PPI 予測を行う様々な手法が提案されてきた中で，近年タンパク質間ドッキング (Protein-Protein Docking, PPD) を利用し，PPI 予測を行うという手法が提案されており，PPD 技術はさらに重要なものとなっている [1], [2]. タンパク質間ドッキングはタンパク質の物理化学的性質などを目的関数に用いてドッキングスコアを導出し，そのスコアから相互作用する場合の予測複合体構造を生成するというものである. ここで，多くの場合ドッキングスコアは複数の項からなる目的関数それぞれの項ごとに重みを付けて足し合わせる方法が一般的であるが，タンパク質ごとに最適な重みは異なっており，一つの同定された重みではなく，タンパク質の性質によって重みを動的に調整することで予測精度の改善が見込める．

[1] 東京工業大学　大学院情報理工学研究科　計算工学専攻
Graduated School of Information Science and Engineering, Tokyo Institute of Technology

そこで我々のグループ [3] は，網羅的 PPI 予測システム “MEGADOCK”[2] の目的関数の重みを，タンパク質の表面電荷などの情報から動的に決定する手法を提案した．以前の研究では一定の改善結果を出したが，その手法にはいくつかの問題点が含まれていた．そこで本研究ではサポートベクター回帰を用いた既存手法の改良の提案を行う．

## 2. 既存研究とその問題点

以前の研究では MEGADOCK の目的関数で用いられている形状相補性と静電相互作用の 2 つの項のうち，静電相互作用項の重みをデフォルトの値から $\alpha$ 倍するという定義のもと，その $\alpha$ を決定するという方法を用いた．以前の提案手法では，タンパク質表面の溶媒露出面積 (Accessible Surface Area, ASA) における負の電荷の偏りを考慮した決定式が最良の改善結果を出した．その式を以下に示す．

$$\sigma = S^-_{chg}/S_{pol}(ligand) - S^-_{chg}/S_{pol}(receptor)$$
$$\alpha^* = 3^Z$$

$S^-_{chg}$ : 負の電荷を有する部分の ASA
$S^+_{chg}$ : 正の電荷を有する部分の ASA
$S_{pol}$ : 極性を有する部分の ASA
$Z$ : $\sigma$ の Z 値

しかし以前の研究で提案した $\alpha$ 決定式は恣意的な式の形とパラメータが用いられており，それについて物理化学的根拠がなく，真に最適な式であるかの議論がなされていなかった．そこでサポートベクター回帰 (Support Vector Regression，SVR) [4] を用いてこの問題を解決することを考える．サポートベクター回帰とは，サポートベクターマシン [4] の原理を回帰問題に応用した，非線形回帰にも適用可能なカーネルベースの手法である．

## 3. サポートベクター回帰による既存手法の改良

### 3.1 特徴量選択

学習に用いる特徴量は以下に示す 3 つを用いる．それぞれの特徴量について，リガンドとレセプターの両方を用いるので，実際の特徴量の数は 6 つとなる．

$S^+_{chg}/S_{pol}$: 極性を持つ部分の ASA に対する正の電荷を持つ部分の ASA の割合
$S^-_{chg}/S_{pol}$: 極性を持つ部分の ASA に対する負の電荷を持つ部分の ASA の割合
$S_{pol}/S_{tot}$: 全体の ASA に対する極性を持つ部分の ASA の割合

上記の特徴量は重みの決定に有効であると思われる．まず，$S^-_{chg}/S_{pol}$ については既存研究で最も良い結果を出した $\alpha$ 決定方法に用いられた特徴量である．さらに，物理化学的には負の電荷だけでなく正の電荷も用いるべきとの考えから $S^+_{chg}/S_{pol}$ も用いる．しかし，これら 2 つの特徴量のみでは，電荷を有する部分の ASA が全体の ASA の中で少ない割合である場合にも，偏りによっては $\alpha$ を大きくすべきと判断される恐れがある．そこで，$S_{pol}/S_{tot}$ を導入することで，全体の ASA の中での極性を持つ部分の ASA の割合が学習の際に考慮され，この問題についても対応できると考えられる．

### 3.2 最適な $\alpha$ の決定方法

サポートベクター回帰において学習させる最適な $\alpha$ については，まず事前に離散的に $\alpha$ を変化させて実際にドッキングを行ったのち，$\alpha$ の変化と予測結果の関係を調べた．それぞれのタンパク質について最良の予測結果を導く $\alpha$ を用いることとする．

## 4. 結果と考察

ドッキング予測精度の評価方法は，MEGADOCK においてデフォルトで出力される 2000 位までの候補構造の RMSD (Root Mean Square Deviation) をすべて計算し，各順位までで最小の RMSD をプロットしたグラフにおける AUC(Area Under the Curve) がどれだけ小さくなるかという基準を用い，既存研究で最も良い結果を出した負の電荷の偏りを考慮する手法との比較を行う．用いるデータセットは Protein-Protein Docking Benchmark 2.0[5] 中の 44 例である．$\alpha$ を変化させたときの，$\alpha = 1$ のときの AUC に対する割合を $R_{AUC}$ とする．それぞれの手法について $R_{AUC}$ の平均値，$R_{AUC} < 0.9$ を満たす例の個数，$R_{AUC} > 1.1$ を満たす例の個数を**表 1** に示す．また両手法について，データセット中のタンパク質についての $1 - R_{AUC}$ の累積値をまとめたグラフを**図 1** に示す．表 1 について，提案手法は既存手法の結果に比べ，悪化した個数はほぼ同数ながら改善した個数を伸ばしている．図 1 についても，提案手法は悪化の具合を抑えながら，改善の具合を大きく伸ばしている．以上の結果から， 提案手法は既存研究の結果を上回っていると言える．このことから，既存手法の恣意的な式やパラメータよりも最適な式が，サポートベクター回帰によって導出されたと考えられる．

### 4.1 新しいデータセットへの適用

前述の実験で用いた Benchmark 2.0 はやや古いデータセットであるため，本稿では新しいデータセット Protein-Protein Docking Benchmark 4.0[6] 中の 132 例についても実験を行った．Benchmark 4.0 に適用する際，提案手法については Benchmark 2.0 で学習を行い，既存手法については Benchmark 2.0 における $\alpha$ 決定式の Z 値を用いていた部分を Benchmark 2.0 で得られた平均と標準偏差を使用した Z 値に準ずる値に変更する．Benchmark 2.0 の場合と

**表 1** Protein-Protein Docking Benchmark 2.0 への実験結果
**Table 1** The result of the experiment for Protein-Protein Docking Benchmark 2.0.

| | 既存手法 | 提案手法 |
|---|---|---|
| $R_{AUC}$ の平均 | 0.97 | 0.92 |
| $R_{AUC} < 0.9$ を満たすタンパク質の個数 | 7 | 13 |
| $R_{AUC} > 1.1$ を満たすタンパク質の個数 | 2 | 1 |

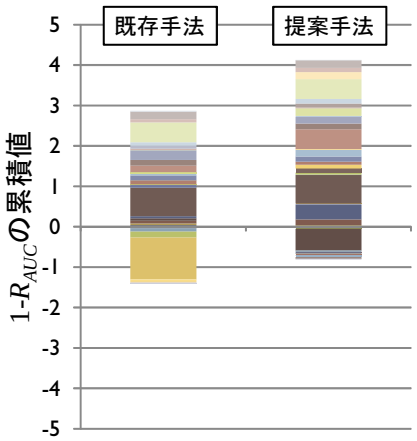


**図 1** $1 - R_{AUC}$ の累積値 (Benchmark 2.0)
色分けされた部分はそれぞれ個々のタンパク質を表す.
**Fig. 1** The accumulated value of $(1 - R_{AUC})$ for Benchmark 2.0. Each color represents the contribution of each protein.

**表 2** Protein-Protein Docking Benchmark 4.0 への実験結果
**Table 2** The result of experiment for Protein-Protein Docking Benchmark 4.0.

| | 既存手法 | 提案手法 |
|---|---|---|
| $R_{AUC}$ の平均 | 1.05 | 1.05 |
| $R_{AUC} < 0.9$ を満たすタンパク質の個数 | 13 | 12 |
| $R_{AUC} > 1.1$ を満たすタンパク質の個数 | 28 | 27 |

同様に，全てのケースに対して計算した $R_{AUC}$ の平均値，$R_{AUC} < 0.9$ を満たす例の個数，$R_{AUC} > 1.1$ を満たす例の個数を**表 2** に示す．両手法について，結果が大きく悪化しており，予測結果の改善は全くできていない．原因としては，Benchmark 2.0 のデータが偶然偏っていた可能性，考慮すべき特徴量が他にも数多く存在することが考えられる．例えば，本研究ではタンパク質表面における電荷の偏りを用いたが，電荷を有する部分が一箇所に集中しているか，分散して均等に分布しているか，といった点や電荷の強さでその影響度合いは大きく異なるであろう．従って，今後より効果的な特徴量を用いることによって結果が改善する可能性がある．

## 5. 結論

本稿では，タンパク質間ドッキング予測精度の向上を目的として，予測に用いる目的関数をタンパク質の性質に基づきサポートベクター回帰を用いて動的に調整するという手法を提案した． Benchmark 2.0 を用いて既存研究との比較を行い，ドッキング予測精度が向上したことを確認した．しかし，Benchmark 4.0 に適用した際に，既存研究，提案手法共にその性能は大きく低下した．原因としては，最初に用いたデータセットのタンパク質が偏っていた可能性，予測に使用した特徴量の数が少なかったことが考えられる．今後の課題として，電荷の分布の詳細等の特徴量を増やすことで予測精度向上を図ることが考えられる．

## 参考文献

[1] D.Juan, F.Pazos, A.Valencia: “High-confidence prediction of global interactomes based on genome-wide coevolutionary networks”, *PNAS*, 105(3): 1–6, 2008.
[2] M.Ohue, Y.Matsuzaki, Y.Akiyama: “Docking-calculation-based method for predicting protein-RNA interactions”, *Genome Inform*, 25(1): 25–39, 2011.
[3] 松崎 裕介, 大上 雅史, 松崎 由理, 佐藤 智之, 関嶋 政和, 秋山 泰: “タンパク質の特性に基づく unbound ドッキングのための剛体予測手法の改良” , 情報処理学会研究報告バイオ情報学 (BIO), 2010-BIO-20(4): 1–8, 2010.
[4] V.N.Vapnik: *The Nature of Statistical Learnig Theory*, Spring-Verlog, New York, 1995.
[5] J.Mintseris, K.Wiehe, B.Pierce, R.Anderson, R.Chen, J.Janin, Z.Weng: “Protein-Protein Docking Benchmark 2.0: an update”, *Proteins*, 60(2): 214–216, 2005.
[6] H.Hwang, T.Vreven, J.Janin, Z.Weng: “Protein-protein docking benchmark version 4.0”, *Proteins*, 78(15): 3111–3114, 2010.