Vol.2019-MPS-123 No.53
Vol.2019-BIO-58 No.53
2019/6/19

マルチドメインタンパク質の相互作用残基ペア予測を用いた立体構造予測手法の改良

松野 駿平[1,2]　大上 雅史[1]　秋山 泰[1,a)]

概要：タンパク質はその立体構造を知ることで機能を解明することができる．しかし，マルチドメインタンパク質は構造が不安定なものが多いため，全長での構造決定が困難であることが多いまた，マルチドメインタンパク質はドメインごとに分解され構造を同定されることが多く，各ドメインの構造は公共データベースに多く収録されている．そこで，近年の計算機の性能の向上に伴い，計算機を用いたマルチドメインタンパク質の構造予測手法が注目されている．既存手法では，立体構造予測を行う際にテンプレートとなるタンパク質を用いているが，テンプレートとなるタンパク質が存在しない場合，予測の精度が落ちてしまう．本研究はマルチドメインタンパク質の立体構造を，テンプレートを必要とせずに予測することを目的とする．そこで，ドメイン毎の構造から剛体ドッキングを行うことで予測構造を生成し，より正解構造に近い構造を上位にリランキングする手法を改良した．提案手法では，テンプレートとなるマルチドメインタンパク質の構造を必要とせずに得られる相互作用に関するスコアと，ドッキング計算時に得られる立体構造に関するスコアを新たにスコア関数に組み込むことで，55 個のマルチドメインタンパク質のうち，50 個のタンパク質の構造を予測するこに成功した（「成功」の定義は 4.3.2 節で説明）．

キーワード：タンパク質立体構造予測，相互作用残基，リランキング，ドメインドッキング

Improvement of three-dimensional structure prediction method using interaction residue pair for multidomain proteins

Shumpei Matsuno[1,2]　Masahito Ohue[1]　Yutaka Akiyama[1,a)]

Abstract: Proteins can be elucidated by knowing their three-dimensional structures. However, structural determination often difficult due to structural instability of multidomain proteins. Therefore, with the improvement of computer performance in recent years, computational structure prediction methods have attracted attention. In this study, we aim to predict the three-dimensional structure of multidomain proteins without the need for a multidomain protein as a template. We generate a predicted structure by rigid-body docking of individual domains and rerank with the score function we improved. In the proposed method, with using the score for the interaction obtained without requiring the structure of the template multidomain protein and the score for the three-dimensional structure obtained during docking calculation, it is possible to predict the structure of 50 multidomain proteins out of 55.

Keywords: protein structure prediction, interaction residue, reranking, domain docking

[1] 東京工業大学 情報理工学院 情報工学系
Department of Computer Science, School of Computing, Tokyo Institute of Technology
[2] 産業技術総合研究所・東京工業大学 実社会ビッグデータ活用オープンイノベーションラボラトリ
AIST-Tokyo Tech Real World Big-Data Computation Open Innovation Laboratory (RWBC-OIL), AIST
[a)] akiyama@c.titech.ac.jp

Vol.2019-MPS-123 No.53
Vol.2019-BIO-58 No.53
2019/6/19

1. 序論

原核生物や真核生物の遺伝子の半分以上はタンパク質ドメインと呼ばれる部分構造を複数有するマルチドメインタンパク質を生成する [1]．それぞれのタンパク質ドメインは密で安定な構造に折りたたまれており，異なるマルチドメインタンパク質内でも構造が保存されている．

タンパク質はその立体構造によって機能を持つため，その立体構造を理解することは重要である [2]．通常はタンパク質の立体構造は X 線結晶構造解析法や NMR 解析法，クライオ電子顕微鏡などによって決定される．しかし，マルチドメインタンパク質は一般的に結晶化が困難であるため，全ての構造を実験によって同定することが難しい．一方で，全体の立体構造が解明されていない場合でも，タンパク質ドメインの立体構造情報が公共データベース（PDB）に多く収録されている．そこで，個々のドメインの構造から計算機を用いて全体構造を予測することで，コストを削減することが期待されている．立体構造予測の手法には *ab initio* の手法 [3][4] と，テンプレートベースの手法 [5]，テンプレートフリーの手法 [6] があり，研究が進められている．

ab initio の手法では，タンパク質全体の構造を計算機によって組み立てていくために計算資源が大量に必要であり，すべてのマルチドメインタンパク質の構造をこの手法を用いて予測することは困難である [4]．またテンプレートベースの手法では，立体構造情報が既知である相同なタンパク質が必要であり，そのようなタンパク質がない場合，予測精度が低くなる場合がある [5]．

一方で，テンプレートフリーの手法では，タンパク質ドメインがマルチドメインタンパク質内でも構造が保存される性質を利用し，タンパク質ドメインの立体構造を用いて剛体ドッキングによって構造を予測する．この手法はマルチドメインタンパク質全体の構造を，PDB に格納されているタンパク質ドメインの立体構造から予測することができるという点で優れている．Hirako ら [7] は，2 個のドメインで構成されるマルチドメインタンパク質である 2 ドメインタンパク質の立体構造予測をするために，剛体ドッキングツールで構造モデルを生成し，スコア関数を用いてドメイン同士の適切なドッキングモデルを選択する手法を開発している．しかしながら，この手法に用いられるスコア関数には，相同性のあるテンプレートタンパク質が必要であり，そのタンパク質がない場合に適用することができない．したがって本研究では，Hirako らの手法 [7] を改良し，相同なテンプレートタンパク質を必要としない手法を提案する．

2. 先行研究

2.1 DINE

Hirako ら [7] は 2 つのドメインをもつ 2 ドメインタンパク質の構造予測手法を開発した．手順を以下と図 1 に示す．

既存手法 [7] の手順

手順 1. ZDOCK 3.0.1[8] を用いて予測構造を 2,000 個生成する
手順 2. 各予測構造に対してスコア関数（DINE）を用いてスコアを算出する
手順 3. 手順 2 のスコアを用いて予測構造をリランキングする

用いられるスコア関数（DINE）は結合エネルギースコア（S_{zrank}），ドメインの相互作用面スコア（S_{int}），ドメイン間距離スコア（S_{ete}）の重み付き線形和で，式 (1) のように定義される．各項に付随する $w_{zrank}, w_{int}, w_{ete}$ は各スコアの重みである．

$$S_{DINE} = w_{zrank}S_{zrank} + w_{int}S_{int} + w_{ete}S_{ete} \quad (1)$$

結合エネルギースコア（S_{zrank}）

ZRANK[9] はファンデルワールスエネルギー，静電相互作用エネルギー，脱溶媒和エネルギーから計算される [10][11]，複数のポテンシャルを用いてスコアを付与するリランキングシステムである．天然に存在するタンパク質の立体構造はそのアミノ酸配列情報によってエネルギーが最も小さくなるようにフォールディングされることが知られており，ZRANK によるスコアがより小さいものが天然に存在する立体構造に近いと言える [12]．生成された予測構造に対して ZRANK[9] を用いて結合エネルギーに関するスコアを計算し，式 (2) によって S_{zrank} を算出する．ただし，zr は各構造モデルの ZRANK スコアであり，max_{zr}, min_{zr} は

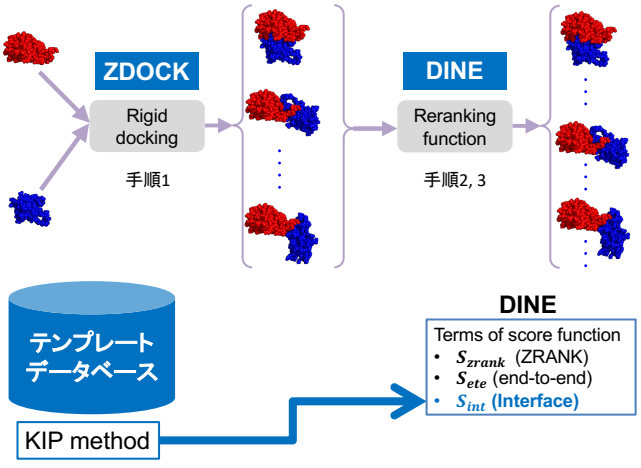

図 1 既存手法 [7] における予測手順
Fig. 1 The way of existing method[7]

それぞれ最大値と最小値である．

$$S_{zrank} = \frac{-zr + max_{zr}}{max_{zr} - min_{zr}} \tag{2}$$

相互作用面スコア（S_{int}）

30%以上の配列相同性があるドメインは，マルチドメイン内での空間的配置と相互作用面が等しくなる場合がある [13][14]．したがって，相互作用面を予測することが可能な場合，精度向上が見込まれる．その手順は，まず既知のマルチドメインタンパク質の構造をクラスタリングし相互作用している残基を記録する．そして，クエリのドメインと配列相同性があるタンパク質がデータベースに存在するとき，相互作用している残基をアラインメントし，予測相互作用残基を出力する．この予測相互作用残基と，構造モデルの相互作用残基の比を S_{int} とする．

ドメイン間距離スコア（S_{ete}）

マルチドメインタンパク質はドメイン間の距離がリンカー領域によって制限されている．したがって，既知のマルチドメインタンパク質のリンカー領域の残基数とドメイン間の距離を用いて，ドメイン間距離スコアを式 (3) で定義する．ただし，$M(L)$ は $|d_e - m_e(L)|$，SD_e は構造モデルのドメイン間の距離，$m_e(L), SD_e(L)$ はリンカーの残基数が L であるマルチドメイン間の距離の平均と標準偏差を表している．平均と標準偏差は DINE[7] で用いられている 1,657 個のマルチドメインタンパク質のデータベースから算出されている．

$$S_{ete} = \begin{cases} 1 & \text{if } M(L) \leq SD_e(L) \\ 2 - \frac{|d_e - m_e(L)|}{SD_e(L)} & \text{if } SD_e(L) < M(L) \leq 2SD_e(L) \\ 0 & \text{if } M(L) > 2SD_e(L) \end{cases} \tag{3}$$

3. 提案手法

3.1 提案手法の概要

既存手法である Hirako らの手法 [7] ではドッキング計算を用いてマルチドメインタンパク質の構造を予測している．しかし，相互作用残基を予測する際に配列相同性のある，全体構造が既知のタンパク質の情報を用いて行っている．この手法では，テンプレートタンパク質が存在しない場合に，適用することができない．そこで本研究では，よりマルチドメインタンパク質立体構造予測の適用範囲を拡大するために，相同なテンプレートとなるタンパク質を必要としない接触アミノ酸残基スコアと，ドッキング計算を行う際に算出されるドッキングスコアを既存手法の相互作用スコアの代わりに用いることを試みた．構造予測手法のフローを図 2 に示し，以下に手順を示す．

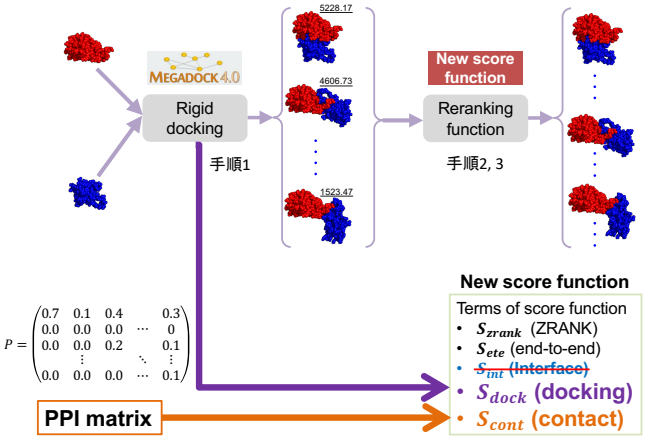

図 2 提案手法のフロー
Fig. 2 The flow of prediction

提案手法の手順

手順 1. MEGADOCK 4.0.2[15] を用いて予測構造を 10,800 個生成する
手順 2. 各予測構造に対してスコア関数を用いてスコアを算出する
手順 3. 手順 2 のスコアを用いて予測構造をリランキングする

3.2 ドメインの定義

本研究では，2 ドメインタンパク質を SCOP[16] の定義にしたがって分割した．しかし，ドメインとドメインをつなぐリンカー領域は常に明確に定義されているわけではない．そこで，N 末端側のドメインの二次構造をとる最後の残基と，C 末端側のドメインの二次構造をとる最初の残基との間をリンカー領域と定義した．つまりドメイン間の二次構造をとらない残基をリンカー領域と定義した．残基ごとに二次構造を形成しているかは DSSP[17] を用いて決定した．

3.3 モデルの生成

構造モデルの生成は MEGADOCK 4.0.2[15] を用いて行った．各回転角ごとにドッキングスコアの上位 3 個の構造モデルを出力し，回転角の数は 3,600 とした．したがって，構造モデルは $3 \times 3{,}600 = 10{,}800$ 個となる．各々のクエリとしたマルチドメインタンパク質に対して，10,800 個の構造モデルそれぞれに算出されるドッキングスコアを最小値が 0，最大値が 1 になるように正規化し，S_{dock} とした．また本研究では，ドメインドッキングをする際にリンカー領域を除いて行った．

3.4 結合エネルギースコア（S_{zrank}）

ZRANK[9] によって自由結合エネルギーを算出するために，reduce[18] を用いて構造モデルに水素を付加した．水

素を付加した構造モデルを ZRANK[9] の入力とし，結合エネルギーを計算し，最小値が 0，最大値が 1 になるように正規化することで S_{zrank} を得た．

3.5 ドメイン間距離スコア（S_{ete}）

生成された 10,800 個の構造モデルに対して，ドメイン間距離をそれぞれ計算し，ドメイン間距離スコア（S_{ete}）を得る．ドメイン間距離は，N 末端ドメインの C 末端残基の C_α 原子と，C 末端ドメインの N 末端残基の C_α 原子間の距離と定義した．

3.6 改良したスコア関数

既存手法のスコア関数で用いられる相互作用面予測に基づくスコア（S_{int}）の代わりに，Protein-Protein Interaction（PPI）から導かれる接触アミノ酸残基スコア（S_{cont}）と MEGADOCK[15] によるドッキング計算から導かれるドッキングスコア（S_{dock}）を取り入れたものが提案手法で用いるスコア関数であり，以下の式 (4) で定義される．S_{zrank} と S_{ete} は先行研究の DINE で用いられている項と同様であり，以下では式 (4) の下線で示す新規に提案した S_{cont} と S_{dock} について説明する．

$$Score = w_{zrank}S_{zrank} + w_{ete}S_{ete} + \underline{w_{cont}S_{cont} + w_{dock}S_{dock}} \tag{4}$$

接触アミノ酸残基スコア（S_{cont}）

相互作用する残基の組み合わせに注目し，設けた項である．本研究では，2 つのドメイン内にある C_α 原子間の距離が 8 Å 以内であれば，その残基ペアが相互作用していると定義する．タンパク質は多量体であることが多く，PDB には多量体の立体構造が格納されている．そこで，PDB に格納されている二量体のタンパク質のうち，UniProtID ベースで冗長性を排除した複合体構造 12,532 個から相互作用する位置に存在するアミノ酸残基ペアの数をカウントし，アミノ酸残基ペアの数を示す 20×20 上三角行列 (PPI Matrix) $\mathbf{P} = (p_{ij})$ を求めた．ただし $\mathbf{P}$ の各要素はカウントした相互作用残基ペアの数を式 (5) により最小値を 0，最大値を 1 に正規化した値とした．

$$p'_{ij} = \frac{p_{ij} - min(p_{ij})}{max(p_{ij}) - min(p_{ij})} \tag{5}$$

また，生成したドメインドッキング構造モデルはドメイン間で相互作用しているアミノ酸残基ペアの数をカウントし，20×20 上三角行列 (Contact Matrix) $\mathbf{C} = (c_{ij})$ を求めた．この 2 つの行列に対し，

$$\mathbf{S} = \mathbf{P} \odot \mathbf{C} \tag{6}$$

によって行列 $\mathbf{P} = (p_{ij})$ を求め，

$$S_{cont} = \frac{\sum_{i,j=1}^{20} s_{ij}}{\sum_{i,j=1}^{20} c_{ij}} \tag{7}$$

によって各構造モデルごとにスコアを計算し，p_{ij} と同様に最小値が 0，最大値が 1 になるように正規化することで，接触アミノ酸残基スコア S_{cont} を求めた．ただし，式 (6) の $\odot$ は行列のアダマール積（要素積）を表す．すなわち $s_{ij} = p_{ij}c_{ij}$ である．

MEGADOCK によるドッキングスコア（S_{dock}）

MEGADOCK 4.0.1[15] を用いてドメインドッキングによる構造モデルを生成する際に，ドッキングスコアが算出される．このドッキングスコアを 0 から 1 に正規化した値をドッキングスコア（S_{dock}）とした．

4. 実験

4.1 データセット

スコア関数の重みを最適化するデータセットと，テスト用のデータセットは Hirako らの研究 [7] で用いられているデータセットを用いた．スコア関数の重み（$w_{zrank}, w_{ete}, w_{cont}, w_{dock}$）を最適化する際には Wollacott ら [3] が用いていた 2 ドメインタンパク質のベンチマークセット 76 個のうち，SCOP[16] ではシングルドメインタンパク質と定義されていた 14 個のタンパク質を除いた，62 個のタンパク質を使用する．また．テスト用のデータセットは pyDockTET[6] で用いられていた 77 個の非冗長なデータセットを使用する．

4.2 スコア関数の重みの最適化

スコア関数の重み $\{w_{zrank}, w_{ete}, w_{cont}, w_{dock}\}$ を訓練用データセットを用いて最適化する．手順は以下に示す．

スコア関数の重みの最適化の手順

手順 1. 重みそれぞれを 1 から 10 まで 1 刻みで変えたスコア関数でモデルをリランキングする

手順 2. 予測が成功したタンパク質の数が最も多くなるときの重みの組み合わせを探索する

手順 3. 最適な重みの組み合わせが複数ある場合，リランキングした結果の上位 500 位に存在する許容構造が最も多いものを選択する

重みを最適化したスコア関数を用いて，テスト用のデータセットをリランキングし，スコア関数の評価をする．

4.3 評価指標

4.3.1 構造モデルの評価

生成された構造モデルの評価には，Hirako ら [7] が採用している平均二乗偏差（RMSD）を用いた評価指標を本研究でも採用した．定義を式 (8) に示す．ただし，N は C_α

の原子数，d_i は構造モデルとネイティブ構造の残基番号が同じである C_α 原子間の距離である．

$$RMSD = \sqrt{\sum_{i=1}^{N} \frac{d_i^2}{N}} \quad (8)$$

また，構造モデルの RMSD を計算する際は，構造モデルの 2 つのドメインのうち，残基数の多いドメインをネイティブ構造と重ね，残基数が少ないドメインのネイティブ構造と構造モデル間の RMSD を計算した．

4.3.2 スコア関数の評価

スコア関数の評価は，1 つのマルチドメインタンパク質に対し，10,800 個の構造モデルをスコア関数を用いてリランキングした際に，ランクが上位から N 位までに許容構造（≤ 10 Å）が現れる場合，そのタンパク質は N 位（$N = 10, 20, 30, 40, 50, 100, 200, 300, 400, 500$）で予測が成功したとする．特に，$N = 10$ 位で予測が成功した場合，そのマルチドメインタンパク質の予測が成功したとし，スコア関数の評価はデータセットに含まれるマルチドメインタンパク質の予測が成功した数とする．

5. 結果

5.1 スコア関数の重みの最適化

訓練用のデータセット（62 個）を用いてスコア関数の 4 つの重み $\{w_{zrank}, w_{ete}, w_{cont}, w_{dock}\}$ を最適化した．全ての重みが 1 である初期状態の場合，62 個のうち 51 個のタンパク質で予測が成功した．予測の「成功」の定義は、4.3.2 節で述べたとおりである。前章で述べた方法で重みを最適化した結果，最適化された重みは $\{w_{zrank}, w_{ete}, w_{cont}, w_{dock}\} = \{9, 2, 1, 4\}$ となった．この重みの組み合わせを用いたスコア関数で訓練用のデータセットを予測した場合，52 個のタンパク質で予測が成功した．重みを最適化することで新たに予測に成功したタンパク質（PDB ID: 1BKB）は，初期状態のスコア関数を用いた場合より，RMSD が小さい構造モデルが上位にリランキングされた．

5.2 スコア関数の評価

上記で求めた重みを用いてテスト用のデータセット（55 個）をリランキングした結果，提案手法ではテスト用のデータセット 55 個のうち 50 個のマルチドメインタンパク質の予測に成功した．テンプレートがなく，相互作用面を予測できない場合の Hirako らの手法（$S_{zrank} + S_{ete}$）[7] より，提案手法では 3 個多く予測に成功した．

5.3 新たに用いた項

接触アミノ酸残基スコア（S_{cont}）

接触アミノ酸残基のみを用いてリランキングした場合，10 位以内に許容構造が存在するのは 55 個のタンパク質のうち 1 個のみであった．この予測に成功したタンパク質（PDB ID: 1BAG）は，RMSD が 4.3 Å の構造モデルを 9 位にリランキングした．

MEGADOCK によるドッキングスコア（S_{dock}）

S_{dock} のみを用いた予測は，単一項を用いた予測のなかで最も多く，55 個のうち 48 個のタンパク質で予測が成功した．

6. 考察

6.1 スコア項の寄与

スコア関数の各項（$S_{zrank}, S_{ete}, S_{cont}, S_{dock}$）の寄与を調べるために，用いる項を変え，重みを最適化した関数を用いた予測結果を表 1 に示す．また，各関数を用いてテスト用データセットをリランキングし，許容構造がどのように分布しているかを図 3 に示す．

結合エネルギースコア（S_{zrank}）

提案手法の各スコア項の重みは訓練用データセットを用いて最適化されている．その中で，最も大きい値となったのが結合エネルギーに基づいたスコア項，S_{zrank} の重み w_{zrank}（$= 9$）であり，他の重みの 2 倍以上の値となった．また，S_{zrank} が含まれているスコア関数と含まれていないスコア関数を比較した場合，含まれているスコア関数がより成功率が高い結果となった，したがって，結合エネルギースコアはスコア関数に大きく寄与していると考えられる．一方，S_{zrank} のみを用いた予測で失敗している 10 個のタンパク質のうち，5 個は提案手法でも予測に失敗している．

表 1　スコアの組み合わせによる成功率の違い
Table 1　Difference of success rate by combination of scores

スコア関数の項				10 位	20 位	500 位
S_{zrank}				0.82	0.85	0.95
	S_{ete}			0.75	0.87	0.93
		S_{cont}		0.02	0.05	0.51
			S_{dock}	0.87	0.91	0.96
S_{zrank}	$+S_{ete}$			0.85	0.85	0.95
S_{zrank}		S_{cont}		0.84	0.84	0.93
S_{zrank}			$+S_{dock}$	0.85	0.89	0.95
	S_{ete}	$+S_{cont}$		0.09	0.13	0.67
	S_{ete}		$+S_{dock}$	0.87	0.91	0.96
		S_{cont}	$+S_{dock}$	0.84	0.85	0.95
S_{zrank}	$+S_{ete}$	$+S_{cont}$		0.85	0.87	0.95
S_{zrank}	$+S_{ete}$		$+S_{dock}$	0.91	0.91	0.98
S_{zrank}		$+S_{cont}$	$+S_{dock}$	0.85	0.89	0.95
	$+S_{ete}$	$+S_{cont}$	$+S_{dock}$	0.87	0.91	0.96
S_{zrank}	$+S_{ete}$	$+S_{cont}$	$+S_{dock}$	0.91	0.93	0.98

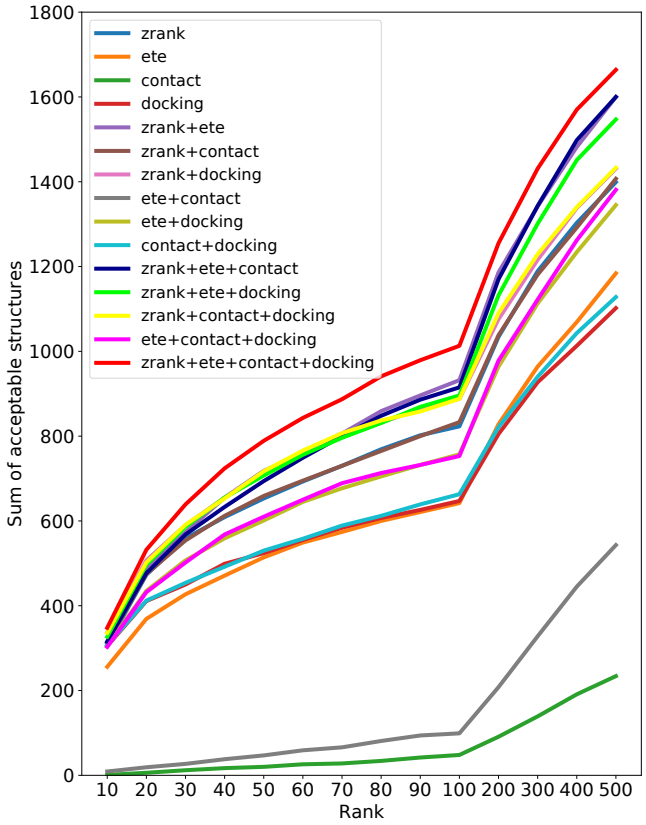

図 3　許容構造の総数

Fig. 3　Total number of acceptable structure of each rank

ドッキングスコア（S_{dock}）

ドッキングスコアのみを用いる予測は，単一のスコア項を用いる予測の中で最も高い予測精度となっている．加えて，S_{dock} の重み，w_{dock} は w_{zrank} についで高い値（$= 4$）となっており，提案手法の精度に寄与していると考えられる．また，結合エネルギーに基づいたスコアのみでは予測できていないが，提案手法で予測に成功している 5 個のタンパク質が存在する．これらはドッキングスコアを用いた予測で上位にリランキングされており，そのため提案手法で予測できていると考えられる．実際に，S_{dock} 以外の 3 個の項を用いたスコア関数より提案手法の方が成功率が高い結果となっている．

ドメイン間距離スコア（S_{ete}）

ドメイン間距離に基づく項，S_{ete} のみを用いての予測精度は高くなく（成功率 $= 75\%$），単一での予測は困難であった．しかし，$\{S_{zrank}, S_{dock}\}$ を用いるスコア関数より，$\{S_{zrank}, S_{ete}, S_{dock}\}$ を用いたスコア関数の方が予測精度が高い．これは，ドメイン間の距離を制限するという目的を達成しているためであると考えられる．

接触アミノ酸残基スコア（S_{cont}）

w_{cont} の値は 4 つのスコア項のうち最も小さい 1 となった．また，単一での予測に成功したのは 55 個のうち 1 個と最も低い値となっており，改善の必要がある．S_{cont} のみを用いた予測で唯一成功したタンパク質（PDB ID: 1BAG）では，図 4 のように予測構造にばらつきがある．RMSD が 10 Å を唯一下回っているのはマゼンダ（9 位）であり，1 位（黄），2 位（赤），3 位（灰）は天然構造からは大きく離れている．しかし，接触アミノ酸残基スコアを用いた提案手法は，より多くの許容構造を上位に予測することが明らかとなった（図 3）．これは，複合体のタンパク質で相互作用している残基ペアの種類が，ドメイン間での相互作用にも影響することを示している．

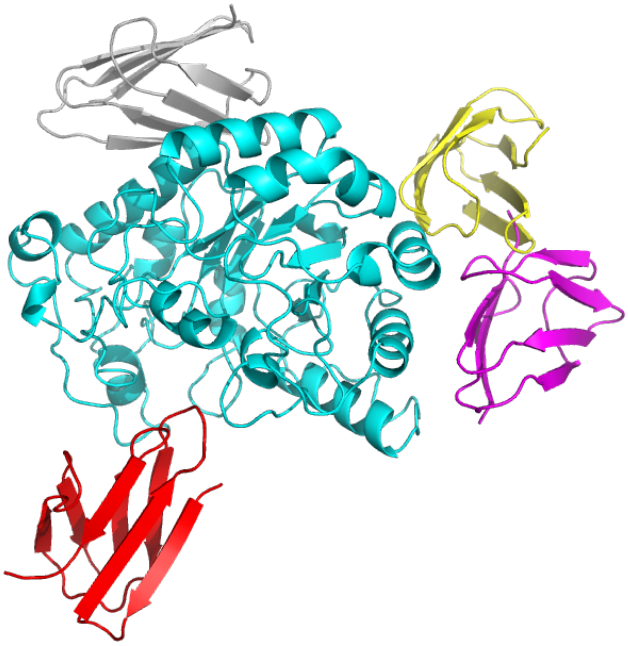

図 4　PDB ID: 1BAG の構造モデル（黄（1 位），赤（2 位），灰（3 位），マゼンダ（9 位，許容構造））

Fig. 4　Poses of predicted structure of PDB ID: 1BAG (yellow (rank 1), red (rank 2), gray (rank 3), magenta (rank 9, acceptable))

表 2　予測に失敗したタンパク質における許容構造の最も高いランク（括弧内は RMSD(Å)）

Table 2　The best rank of acceptable pose that failed to predict (value in parentheses is RMSD(Å))

PDB	S_{zrank}	S_{dock}	提案手法	相互作用面積
1etpB	1,243 (1.2)	36 (1.1)	224 (1.1)	784 Å^2
1i8dB	1,296 (2.4)	891 (0.9)	2,189 (0.9)	821 Å^2
1ik6A	106 (3.1)	14 (1.0)	16 (1.0)	1,090 Å^2
1kzlA	1,728 (2.9)	12 (1.2)	213 (1.1)	979 Å^2
1p5uA	472 (2.1)	3,255 (2.1)	211 (2.1)	585 Å^2

6.2　スコア関数の評価

予測に失敗したタンパク質を表 2 に示す．予測に失敗した 5 個のうち，結合エネルギーによる予測に失敗しているタンパク質は提案手法でも失敗していた．これは w_{zrank}

の値が大きく，予測精度が悪い場合に，提案手法での予測に失敗してしまうことを示している．また，予測に失敗したタンパク質のネイティブ構造の相互作用面が 1,400 Å^2 より小さい．この場合，相互作用の弱さゆえにタンパク質ドッキングの結果が悪くなってしまう [19] ため，結合エネルギーによる予測と，ドッキングによる予測の精度が悪くなってしまい，予測に失敗していると考えられる．

7. 結論

本研究では，2 つのドメインをもつマルチドメインタンパク質の構造を計算機を用いて予測する手法の改良を行った．部分構造を用いて剛体ドッキングを行い立体構造を予測する既存手法に Hirako らの手法 [7] がある．しかし，相互作用面を予測する際にクエリとなるマルチドメインタンパク質と相同性のあるテンプレートとなるタンパク質が必要となる．本研究では，テンプレートとなるタンパク質を必要とせずに相互作用に関するスコアを用いるために接触アミノ酸残基スコア（S_{cont}）を提案し，またドッキング計算時に算出されるドッキングスコアをスコア関数に用いた．これらにより，テンプレートとなるタンパク質がない場合でも予測が可能となり，接触アミノ酸残基スコアを用いることで，より多くの許容構造を上位にリランキングすることが可能となった．

しかし，接触アミノ酸残基スコアのみを用いた予測は困難であり，相互作用面が小さいマルチドメインタンパク質に対しての予測には課題がある．接触アミノ酸残基スコアを用いることで相互作用面をより精度良く予測することが可能であれば，予測の精度向上が見込まれる．また，ドメイン間距離に基づくスコアは，条件に当てはまる構造モデルには一様なスコアが付与されている．このスコアを段階的に変化させる，もしくはよりネイティブ構造に近い構造モデルに対してスコアを高く付与するなどのような改善の余地があり，これらは今後の課題である．

謝辞 本研究の一部は，JSPS 科研費 (17H01814, 18K18149)，JST 世界に誇る地域発研究開発・実証拠点（リサーチコンプレックス）推進プログラム，文部科学省地域イノベーション・エコシステム形成プログラム，AMED 創薬等先端技術支援基盤プラットフォーム (BINDS) (JP19am0101112) の支援を受けて行われた．

参考文献

[1] Apic, G., Gough, J. and Teichmann, S.A. Domain Combinations in Archaeal, Eubacterial and Eukaryotic Proteomes. *J. Mol. Biol.*, Vol. 310(2), 311-325 (2001).

[2] Vogel, C., Bashton, M., Kerrison, N.D., Chothia, C. and Teichmam, S.A. Structure, function and evolution of multidomain proteins. *Curr. Opin. Struct. Biol.*, Vol. 2(14), 208-216 (2004).

[3] Wollacott, A.M, Zanghellini, A., Murphy, P. and Baker, D. Prediction of structures of multidomain proteins from structures of the individual domains. *Protein. Sci.*, Vol. 16(2), 165-175 (2007).

[4] Hardin, C., Pogorelov, T.V. and Luthey-Schulten, Z. Ab initio protein structure prediction. *Curr. Opin. Struct. Biol.*, Vol. 12(2), 176-182 (2002).

[5] Xu, D., Jaroszewski, L., Li. Z. and Godzik, A. AIDA: ab initio domain assembly for automated multi-domain protein structure prediction and domain-domain interaction prediction. *Bioinformatics.*, Vol. 31(13), 2098-2105 (2015).

[6] Cheng, T.M.K., Blundell, T.L. and Fernandez-Recio, J. Structural assembly of two-domain proteins by rigid-body docking.*BMC Bioinformatics*,Vol. 9, 441 (2008).

[7] Hirako, S. and Shionyu, M. DINE: A Novel Score Function for Modeling Multidomain Protein Structures with Domain Linker and Interface Restraints. *IPSJ Trans. on Bioinformatics.*, Vol. 5,18-26 (2012).

[8] Chen, R., Li, L. and Weng, Z. ZDOCK: An initial-stage protein-docking algorithm. *Proteins.*, Vol. 52(1), 80-87 (2003).

[9] Prierce, B. and Weng, Z. ZRANK: reranking protein docking predictions with an optimized energy function. *Proteins*, Vol. 67(4), 1078-1086 (2007).

[10] Brooks, B.R., Bruccoleri, R.E., Olafson, B.D., States, D.J., Swaminathan, S. and Karplus, M. CHARMM: a program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.*, Vol. 4(2) (1983).

[11] Zhang, C., Vasmatzis, G., Cornette, J.L. and DeLisi, C. Determination of atomic desolvation energies from the structures of crystallized proteins. *J. Mol. Biol.*, Vol. 267(3), 707-726 (1997).

[12] Anfinsen, C.B. Principles that govern the folding of protein chains. *Science.*, Vol. 181(4096), 223-230 (1973).

[13] Aloy, P., Ceulemans, H., Stark, A. and Russell, R.B. The relationship between sequence and interaction divergence in proteins. *J. Mol. Biol.*, Vol. 332(5), 989-998 (2003).

[14] Korkin, D., Davis, F.P. and Sali, A. Localization of protein-binding sites within families of proteins. *Protein Sci.*, Vol. 14(9), 2350-2360 (2005).

[15] Ohue, M., Shimoda, T., Suzuki, S., Matsuzaki, Y., Ishida, T. and Akiyama, Y. MEGADOCK 4.0: An ultra-high-performance protein-protein docking software for heterogeneous supercomputers. *Bioinformatics.*, Vol. 30(22), 3281-3283 (2014).

[16] Murzin, A.G., Brenner, S.E., Hubbard, T. and Chothia, C. SCOP: A structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* Vol. 247(4), 536-540 (1995).

[17] Kabsch, W. and Sander, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features, *Biopolymers*, Vol. 22(12), 2577-2637 (1983).

[18] Word, J.M., Lovell, S.C., Richardson, J.S. and Richardson, D.C. Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation. *J. Mol. Biol.*, Vol. 285(4), 1735-1747 (1999).

[19] Vajda, S. Classification of protein complexes based on docking difficulty. *Proteins.*, Vol. 60(2), 176-180 (2005).