様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　２９　年　　５　月　２８　日現在

機関番号：　１４５０１
研究種目：　基盤研究(B)（一般）
研究期間：　2012～2016
課題番号：　２４３０００５６

研究課題名（和文）分子表面の三次元データマイニングによるタンパク質機能知識の創出

研究課題名（英文）Three-dimensional data mining from protein molecular surfaces for discovery of knowledge about protein functions

研究代表者

大川　剛直（OHKAWA, Takenao）

神戸大学・大学院システム情報学研究科・教授

研究者番号：３０２２３７３８

交付決定額（研究期間全体）：（直接経費）　　13,500,000円

研究成果の概要（和文）：本研究では、タンパク質の機能に関連する知識の創出を目的として、分子表面データに対する三次元データマイニング技術を開発した。分子表面を三次元画像と見なし、多数の特徴点から構成される点群あるいはグラフとして表現することにより、機能に関わる部位を、点群のマッチング結果に対するバイクラスタリングや最適グラフ発見により、抽出する方式を提案した。また、タンパク質の機能に関する観点からタンパク質構造解析文献を検索する手法や文献から機能に関する知識を抽出する手法についても提案した。

研究成果の概要（英文）：In this research, we aimed to develop several methods of three-dimensional data mining from protein molecular surfaces for discovery of knowledge about protein functions. Molecular surfaces are regarded as three-dimensional images that are expressed by point clouds or graphs. The functional sites of proteins are extracted by biclustering of matched points in the point clouds or by an optimal graph discovery. In addition, we proposed an effective method to search for related articles on protein structure analysis by considering a user's intention and developed a new method for the automatic extraction of protein-protein interaction information from scientific articles by predicting dominant keywords.

研究分野：　情報工学

キーワード：　知識発見　データマイニング　バイオインフォマティクス　三次元データ　バイオデータ処理　バイクラスタリング

１．研究開始当初の背景

近年、構造解析技術の進展に伴い、様々なタンパク質の立体構造が明らかにされ、多数の三次元構造データが蓄積されつつある。タンパク質の機能に深く関わる重要部位は、生物の進化の過程で保存される傾向にあるため、多数のタンパク質において頻出する類似構造、すなわち保存性の高い局所的な構造を発見することにより、機能部位に特異的な構造上の特徴に関する知見を獲得し、さらにこれまで知られていなかった新規の機能部位に関する情報を提供することが期待できる。これまでに、タンパク質の相互作用部位の網羅的比較をもとにタンパク質を分類する試みは見られるが、分子表面全体の網羅的なマイニングにより、頻出する局所部位を発見する取り組みは、ほとんど行われていない状況にある。

２．研究の目的

タンパク質分子表面データに対する網羅的な三次元データマイニング技術を開発する。

分子表面を三次元画像と見なすことにより、特徴点抽出と特徴量計算を行い、空間上の特徴点集合である三次元点群や各特徴点を頂点とするグラフとして、タンパク質分子表面を記述することで、タンパク質局所部位の物理的・化学的特性を考慮した頻出パターン発見、最適パターン発見を実現する。このような多数の点から構成される点群や巨大なグラフを対象とする網羅的なマイニングを実用時間で実行するための高速化技法として、パターン照合用メモリ型プロセッサの導入とそのためのアルゴリズム開発や抽象グラフの導入についても検討する。開発する三次元データマイニング技術の活用により、タンパク質の結合部位や機能に関与する重要部位の発見など、タンパク質の機能に関する新たな知識の創出を目指す。

また、発見された機能部位などの評価を円滑に行うため、類似機能を持つタンパク質について記述された文献の検索や、文献からタンパク質相互作用情報を自動抽出する手法についても開発する。

３．研究の方法

(1) タンパク質分子表面を三次元画像として捉え、これを特徴点集合として記述するとともに、高速なパターンマッチングとバイクラスタリング処理により、機能に関与する重要部位を特定・抽出する方式について検討する。

(2) タンパク質分子表面を構成するポリゴンデータをもとに、グラフ表現し、結合に寄与する部位の予測を、最適グラフ発見問題として定式化するとともに、抽象グラフの導入により、その高速化を実現する方式について検討する。

(3) タンパク質構造解析について記述された文献を対象として、既存のデータベースの活用によりタンパク質の機能の観点から、関連文献を検索する方式について検討する。

(4) 機械学習により、相互作用するタンパク質ペアに関する知識を文献から自動抽出する方式について検討する。

４．研究成果

(1) 多くのタンパク質は、その分子表面上で他の物質と結合あるいは相互作用することにより、機能する。結合現象には、2 つの分子がお互いに離れた位置関係にある状態から相互認識を行う過程が存在し、そのような認識の目標となる局所的な重要部位が分子表面上に存在すると考えられる。また、同様な物質と結合するタンパク質は形状や物性が類似する部分構造を有することが多く、そのような重要部位は類似するタンパク質群において共通に観察される部位である。そこで、タンパク質分子表面間の網羅的比較を通して、分子表面上から重要部位を抽出する手法を提案した。

提案手法では、特徴点集合として表現したタンパク質分子表面に対して、三次元空間上での高速なマッチング処理により、各タンパク質間の共通部分構造を発見する。具体的には、重要部位を発見する対象となるクエリタンパク質と複数の既知の参照タンパク質とのマッチング結果から、クエリタンパク質の特徴点と参照タンパク質の特徴点間の対応関係を網羅的に求め、これをバイナリ行列形式で表現する。このようにして生成されるバイナリ行列に対してバイクラスタリング処理を行うことで、クエリタンパク質と複数の参照タンパク質間の共通部分構造をバイクラスタとして抽出する。このとき、列が参照タンパク質に対応することから、タンパク質間の類似性をバイクラスタリング処理に反映するための列間類似度、行が位置関係を持つ特徴点に対応することから、空間上の近接関係を反映するための行間類似度を導入する。さらに、列には、同一参照タンパク質に対する複数のマッチング結果を含むため、同一タンパク質列群からの排他的列選択を考慮するとともに、これを効率的に実行するためのランダムサンプリングを導入した新たなバイクラスタリングアルゴリズムを考案した。

60 種類のタンパク質に対して、各タンパク質をクエリ、残りの 59 種類を参照タンパク質と想定し、網羅的な分子表面マッチングにより生成されるバイナリ行列を対象にバイクラスタリングを行うことで、重要特徴点を抽出した。得られたバイクラスタに含まれる特徴点が、実際に結合に関与するとされているアミノ酸残基に対応する割合（適合率）により、抽出結果を評価した。

まず、ランダムサンプリングを行った場合と、網羅的に探索した場合における、バイクラスタリング処理に要した時間を図 1 に、抽出結果の適合率を表 1 に示す。これより、ランダムサンプリングによって、精度を落とす

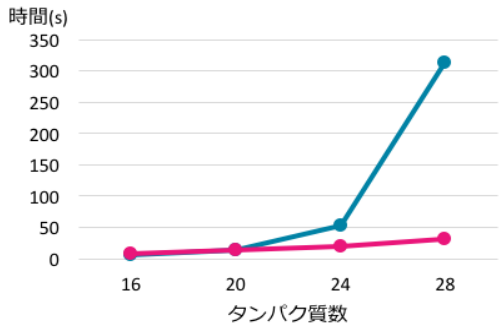

図 1　計算時間の比較（●網羅的探索、●ランダムサンプリング）

表 1　適合率の比較

タンパク質数	16	20	24	28
網羅的探索	0.618	0.656	0.672	0.650
ランダム	0.616	0.662	0.671	0.656

表 2 既存手法との比較

	TOP1	TOP1～3
Bimax	0.548	
BiBit	0.572	
PDNS	0.583	
提案手法	0.667	0.727

ことなく、処理時間を大幅に短縮できていることを確認した。

次に、既存のバイクラスタリング手法と提案手法による適合率の比較結果を表 2 に示す。表において、TOP1 は 1 つの参照タンパク質に対して、対応する特徴点の個数が最大となるマッチングのみを用いた場合、TOP1～3 は、特徴点の個数が上位 3 位までのものから、排他的に 1 つを選択した場合を意味する。これより、既存手法に比べて高い精度で結合に関与する重要部位を抽出することができ、提案手法の有効性を確認した。さらに、個別の結果に対する考察により、分子が結合する際に目印とする重要部位が、結合位置から離れた位置において発見されていることが確認されており、提案手法の独創性を示すことができた。

(2)　タンパク質の分子表面を記述したポリゴンデータをもとに、結合部位を予測する手法とその高速化方式について提案した。

提案手法では、同一のリガンドが結合するタンパク質の間で頻出し、それ以外のタンパク質ではあまり観察されないような類似局所部位を結合部位として予測する。タンパク質分子表面をグラフ表現することで、局所部位の類似性評価を類似グラフ探索として実現する。このとき、結合部位になる可能性が高い部分グラフに高い評価値を与える評価関数を導入することで、結合部位予測問題を最適グラフ発見問題として定式化する。一般に、タンパク質表面を記述したグラフは、そのサイズが大きいため、類似グラフ探索に大きな計算コストを要する。そこで、提案手法では類似する頂点集合を 1 つの頂点に集約して再構成した抽象グラフを利用することで、処理の効率化を図る。すなわち、最適グラフ発見においては、多数の参照タンパク質に対応するグラフの中から、クエリとなるグラフに類似する部分グラフがいくつ存在するかを求めることが基本的な処理となるが、抽象グラフに対して事前にこの処理を適用することによって、全てのグラフを探索することなく、存在する個数の上限を求めることが可能となり、これにより、計算量を削減する。

一方で、抽象グラフの利用により、結合部位になる可能性の低い部分グラフを探索する問題が生じる。この問題を解決するために、提案手法では、複数の抽象グラフを平均化した平均抽象グラフを新たに導入し、結合部位になる可能性が高い部分グラフに対してのみ類似グラフ探索を行うことで、さらなる計算コストの削減を試みる。

5 種類のリガンドのいずれかに結合する計 37 種類のタンパク質における分子表面データを対象として、結合部位の予測に要した時間を比較した結果を図 2 に示す。これより、平均抽象グラフを導入せずに、抽象グラフのみを導入した手法（■）は、抽象グラフを用いなかった手法（■）に比べて、予測時間を 70%程度に抑えることができていることがわかる。さらに、平均抽象グラフを導入することにより（■）、その半分の計算時間で予測を実現している。また、結合部位の予測正解率に関しては、図 3 に示すように、抽象グラフを導入しなかった手法と比べると、結合するリガンドの種類によっては、やや劣る場合が見られるが、平均抽象グラフの導入については、その利用の有無にかかわらず、同程度の精度が得られており、平均抽象グラフを導入した提案手法の有効性を確認した。

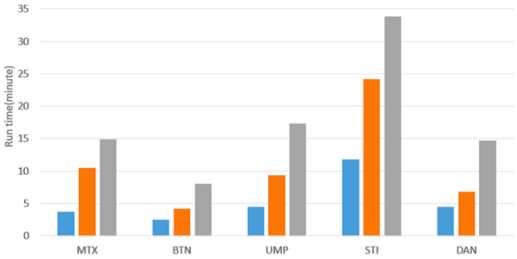

図 2　計算時間（■平均抽象グラフ、■抽象グラフのみ、■抽象グラフなし）

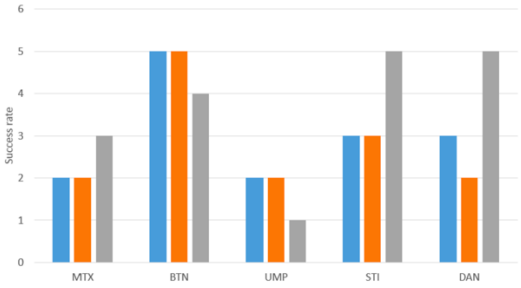

図 3　予測正解率（■平均抽象グラフ、■抽象グラフのみ、■抽象グラフなし）

(3) タンパク質の機能や構造の解明を目的として、様々な構造解析手法が開発され、解析結果はタンパク質構造解析関連論文として蓄積・公開されている。そこで、タンパク質構造解析に関する論文を対象に、注目している論文をクエリとして、その論文に類似する論文を検索するための新しい手法を提案した。

提案手法では、2 つの論文を入力クエリとして用い、両者の関連性を概念階層グラフから評価することにより、検索に際して、どのような観点から論文を類似すると捉えているか、すなわち検索における意図を定量化する。さらに、論文を概念集合に変換する際に用いるデータベースとして、主にタンパク質の生物学的機能の観点からの概念階層である GO (Gene Ontology)と、より幅広い生物学医学関連の用語をまとめた MeSH (Medical Subject Headings) の 2 つを利用し、両者の特性を相補的に利用した類似度算出手法を定式化した。

入力クエリを構成する 2 つの論文の組合せ 21 ペアを 1 つのテストセットとし、合計 3 つのテストセットを対象とした検索実験を実施した。文献引用情報に基づいて用意した擬似正解データを用いて提案手法による検索結果の良否を平均適合率 (MAP) により評価した。評価結果を表 3 に示す。表 3 より、3 つのテストセットのいずれにおいても、GO に基づく概念集合のみを用いた場合、MeSH に基づく概念集合のみを用いた場合に比較して、高い値が得られ、提案手法の有効性を確認できる。また、検索の意図を注目度として定量化することの有効性が示される。さらに、擬似正解データを用いる代わりに、検索結果の上位 3 位までに選ばれた文献に対して、手作業により、論文内容に基づいて、望ましい検索結果であるかどうかを判断した結果、提案手法により、正解率が顕著に上昇していることを確認した。

(4) 多くのタンパク質は、他のタンパク質と相互作用することにより、機能を果たすことが知られている。このようなタンパク質間相互作用に関する情報については、様々な実験結果をもとにデータベース化が進められている一方で、論文中に自然言語を用いて記述されているものも多数あり、文献テキストから抽出・整理することが望まれている。そこで、文献テキストの記述上の特徴をもとに、機械学習により相互作用情報の有無を判定することで、相互作用するタンパク質ペアを自動抽出する新しい手法を提案した。

表 3 検索評価結果

テストセット	提案手法	GO のみ	MeSH のみ	提案手法 (注目度無)
1	0.610	0.580	0.518	0.560
2	0.580	0.560	0.475	0.550
3	0.518	0.450	0.329	0.463

提案手法では、文献内のテキストを表現するための多数の特徴の中で、特定のキーワード (bind や interact などの英単語) が、相互作用の有無の判定に際して大きな影響を与えることに着目し、このようなキーワードを優性キーワード (dominant keyword, DK) と名付け、DK の有無により、訓練データを複数に分割し、分割された訓練データごとに学習することで、抽出精度の向上を図る。このとき、あるキーワードが DK であるか否かは、そのキーワードそのものに対して絶対的に定まるのではなく、事例 (文中における相互作用有無の判定対象となるタンパク質ペア) 毎に決まることに留意し、各事例に対して、その事例に出現するキーワードが DK であるかどうかを予測する方式を導入している。すなわち、図 4 に示すように、訓練データ内の事例を、DK を持つと仮定できるもの ($DK = 1$) と、DK を持たないと仮定できるもの ($DK = 0$) に分割し、それぞれで学習した分類器を用いた判定結果の成否をもとに、各事例における DK の有無の仮定を検証することで、仮定を更新する。このプロセスを分割が収束するまで繰り返すことにより、より正確な DK の有無の予測を実現する。

5 種類のコーパス (相互作用の有無についてラベル付けされた事例集合) に対する相互作用情報抽出実験により、得られた F 値 (適合率と再現率の総合評価値) を表 4 に示す。いずれのコーパスに対しても、事例集合のサブセットへの分割 (MC) の有効性、ならびに特徴選択 (FS) と DK 予測 (DK) を併用することの有効性が示されている (SC はいずれも導入無し)。さらに、4 つのコーパスを対象に、提案手法を既存手法と比較した結果を表 5 に示す。これより、2 つのコーパスに対して、提案手法により、最大の F 値が得られることを確認した。

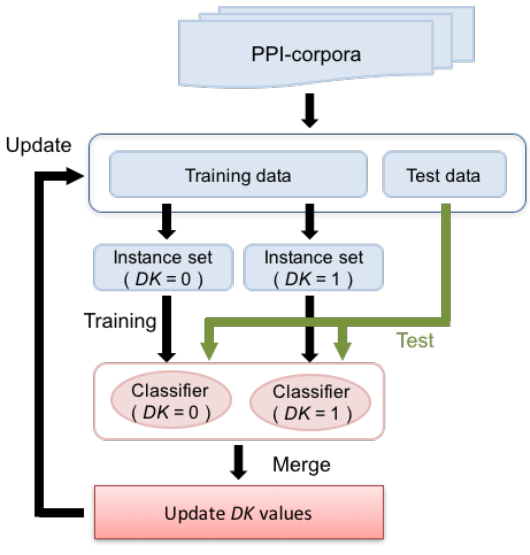

図 4 DK 更新アルゴリズム

表 4 5 種類のコーパスにおける *F* 値

	LLL	HPRD50	IEPA	AImed	BioInfer
SC	0.821	0.723	0.661	0.573	0.711
MC	0.836	0.720	0.671	**0.603**	0.711
DK-MC	0.848	0.750	0.690	0.600	0.717
FS-MC	0.847	0.754	0.690	0.583	0.719
DK-FS-MC	**0.850**	**0.770**	**0.692**	0.600	**0.727**

表 5 関連研究との比較

コーパス	手法	*F* 値
LLL	Fundel et al.	0.820
	Fayruzov et al.	0.780
	Van Landeghem et al.	0.820
	提案手法	**0.850**
HPRD50	Van Landeghem et al.	0.710
	提案手法	**0.770**
IEPA	Van Landeghem et al.	**0.710**
	提案手法	0.692
	Giuliano et al.	**0.639**
	Mitsumori et al.	0.543
	Fayruzov et al.	0.450
AImed	Van Landeghem et al.	0.620
	Edit of Erkan et al.	0.556
	Cosine of Erkan et al.	0.581
	提案手法	0.600

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計 14 件）

① Phan Thi Thanh Thuy and Takenao Ohkawa, “Protein-protein Interaction Extraction with Feature Selection by Evaluating Contribution Levels of Groups Consisting of Related Features”, BMC Bioinformatics, Vol. 17, Suppl. 7, pp. 517-531, DOI: 10.1186/s12859-016-1100-z, 査読有 (2016).

② 伊藤あずさ, 大川剛直, “概念階層グラフを利用した検索意図の反映が可能な蛋白質構造解析文献検索手法”, 電気学会論文誌C, Vol. 135, No. 3, pp. 340-348, 査読有 (2015).

③ Azusa Ito and Takenao Ohkawa, “A Method of Searching for Related Literature on Protein Structure Analysis by Considering a User’s Intention”, BMC Bioinformatics, Vol. 16, Suppl. 7, DOI: 10.1186/1471-2105-16-S7-S4, 査読有 (2015).

④ Shun Koyabu, Thi Thanh Thuy Phan, and Takenao Ohkawa, “Extraction of Protein-Protein Interaction from Scientific Articles by Predicting Dominant Keywords”, BioMed Research International, Vol. 2015, DOI: 10.1155/2015/928531, 査読有 (2015).

⑤ Natsumi Kurumatani, Hiroyuki Monji, and Takenao Ohkawa, “Binding Site Extraction by Similar Subgraphs Mining from Protein Molecular Surfaces and Its Application to Protein Classification”, International Journal on Artificial Intelligence Tools, Vol. 23, No. 3, DOI: 10.1142/S0218213014600070, 査読有 (2014).

⑥ Takuma Mitsui and Takenao Ohkawa, “Binding Site Extraction by Detecting Optimal Graphs from Protein Molecular Surfaces”, International Journal of Bioscience, Biochemistry and Bioinformatics, Vol. 4, No. 1, pp. 28-32, DOI:10.7763/IJBBB.2014.V4.305, 査読有 (2014).

⑦ Tomoki Aso and Takenao Ohkawa, “Method of Retrieving Articles on Protein Structure Analysis from User Intention”, International Journal of Bioscience, Biochemistry and Bioinformatics, Vol. 3, No. 3, pp. 182-186, DOI: 10.7763/IJBBB.2013.V3.192, 査読有 (2013).

⑧ Kazunori Miyanishi and Takenao Ohkawa, “A Method of Extracting Sentences Containing Protein Function Information from Articles by Iterative Learning with Feature Update”, Computational Intelligence Methods for Bioinformatics and Biostatistics, Lecture Notes in Computer Science, Vol. 7845, pp. 81-94, DOI: 10.1007/978-3-642-38342-7_8, 査読有 (2013).

〔学会発表〕（計 16 件）

① Masaya Yatori, Takuma Mitsui, and Takenao Ohkawa, “Optimal Graph Detection with Summary Graph for Identification of Ligand-Binding Site from Protein Molecular Surface”, Biotechnology and Bioinformatics Symposium 2016, 2016 年 12 月 8-9 日, Provo, Utah (USA).

② Hiroto Nishimura, Kento Sakaue, and Takenao Ohkawa, “Extraction of Protein Recognition Spots by Biclustering Considering Exclusive Selection of Column”, Biotechnology and Bioinformatics Symposium 2016, 2016 年 12 月 8-9 日, Provo, Utah (USA).

③ Thi Thanh Thuy Phan, Takenao Ohkawa, and Akihito Yamamoto, “Protein-protein Interaction Extraction from Literature with Evaluation of Cross-corpus Learning”, 人工知能学会第 101 回人工知能基本問題研究会, 2016 年 8 月 7-8 日, 北海道大学（北海道・札幌市）.

④ 八鳥真弥, 三井拓真, 大川剛直, “最適グラフ発見に基づく蛋白質表面からの結合部位抽出におけるグラフの抽象化”, 第 108 回 MPS・第 46 回 BIO 合同研究発表会, 2016 年 7 月 4-6 日, 沖縄科学技術大学院

大学（沖縄県・国頭郡恩納村）.

⑤ 西村宏人, 阪上絢人, 大川剛直,“蛋白質分子表面マッチングと項目集合からの排他的選択を考慮したバイクラスタリングを用いた重要特徴点抽出”,, 第 108 回 MPS・第 46 回 BIO 合同研究発表会, 2016 年 7 月 4-6 日, 沖縄科学技術大学院大学（沖縄県・国頭郡恩納村）.

⑥ Phan Thi Thanh Thuy and Takenao Ohkawa, “Protein-protein Interaction Extraction with Feature Selection by Evaluating Contribution Levels of Groups Consisting of Related Features”, Biotechnology and Bioinformatics Symposium 2015, 2015 年 12 月 10-11 日, Provo, Utah (USA).

⑦ Azusa Ito and Takenao Ohkawa, “A Method of Searching for Related Literature on Protein Structure Analysis by Considering a User’s Intention”, Biotechnology and Bioinformatics Symposium 2014, 2014 年 12 月 11-12 日, Provo, Utah (USA).

⑧ 伊藤あずさ, 大川剛直, “概念階層グラフを利用した検索意図の反映が可能な蛋白質構造解析関連文献検索手法”, 電気学会第 58 回情報システム研究会, 2014 年 5 月 16 日, 電気学会会議室（東京都）.

⑨ 車谷奈都実, 大川剛直, “3 次元画像特徴量を用いた蛋白質分子表面比較”, 情報処理学会 第 29 回バイオ情報学研究会, 2012 年 6 月 28-29 日, 沖縄科学技術大学院大学（沖縄県・国頭郡恩納村）.

⑩ 小薮 駿, 大川剛直, “複数の分類器に基づく半教師あり学習を用いた文献からの蛋白質間相互作用抽出”, 情報処理学会 第 29 回バイオ情報学研究会, 2012 年 6 月 28-29 日, 沖縄科学技術大学院大学（沖縄県・国頭郡恩納村）.

⑪ Kazunori Miyanishi and Takenao Ohkawa, “A Method of Extracting Sentences Containing Protein Function Information from Articles by Iterative Learning with Feature Update”, the Ninth International Meeting on Computational Intelligence Methods for Bioinformatics and Biostatistics, 2012 年 6 月 12-14 日, Houston, Texas (USA).

６．研究組織

(1)研究代表者

大川 剛直（OHKAWA, Takenao）

神戸大学・大学院システム情報学研究科・教授

研究者番号：30223738

(2)研究分担者

（　　）

研究者番号：

(3)連携研究者

（　　）

研究者番号：

(4)研究協力者

（　　）