様式Ｃ－１９

科学研究費助成事業（科学研究費補助金）研究成果報告書

平成 24 年　5 月 23 日現在

機関番号：１４５０１
研究種目：基盤研究(C)
研究期間：2009～2011
課題番号：２１５００１３９
研究課題名（和文）　タンパク質機能知識の発見のための異種データハイブリッドマイニング

研究課題名（英文）　Hybrid data mining 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

研究成果の概要（和文）：タンパク質の三次元構造データ、タンパク質間相互作用や代謝関係を表現したネットワークデータ、タンパク質構造解析について記述した文献などのテキストデータを相互に活用するデータマイニング手法を提案した。これにより、タンパク質機能部位の予測、文献からの機能情報の抽出、利用者の意図を考慮した類似文献検索、生体機能の観点からの生物系統知識の発見など、タンパク質機能に関する各種知識の発見を可能とした。

研究成果の概要（英文）：Several methods for hybrid data mining from the various types of data such as the protein 3D structure data, the network data (e.g. protein-protein interaction data and metabolic pathways), and the text data describing the results of the protein structure analysis have been proposed. Applications to the knowledge discovery of protein functions, which include prediction of the protein functional sites, information extraction from the biomedical articles, retrieval of similar articles considering user's aspect, and classification of organisms from the viewpoint of the biological function, have been developed based on the proposed data mining methods.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
２００９年度	1,000,000	300,000	1,300,000
２０１０年度	1,500,000	450,000	1,950,000
２０１１年度	900,000	270,000	1,170,000
年度			
年度			
総　計	3,400,000	1,020,000	4,420,000

研究分野：情報科学
科研費の分科・細目：情報学・知能情報学
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

１．研究開始当初の背景

タンパク質は、生体における各種機能の主体となる高分子であり、生物が生命現象を営む上で極めて重要な役割を持つ。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。また、大規模実験により、あるタンパク質と他のタンパク質との間の相互作用性の有無に関する網羅的なデータが揃いつつある。さらに、研究者が発表した論文の中には、対象としているタンパク質の構造や、化学的性質、そこから推測される機能などの情報が記載されている。

このように、タンパク質に関わる代表的なデータとして、三次元構造データ、関係データ（相互作用データ）、文献データが挙げら

れるが、これらは、データ構造的観点からも、意味内容の観点からも、大きく異なる種類のデータと言える。その一方で、互いに深い関連性を有しており、異種データを相補的に利用し、効果的に組み合わせることにより、それぞれ単独では得ることが困難なタンパク質の機能に関わる重要な知識を発見することが期待されている。

２．研究の目的

構造データマイニングアルゴリズム、タンパク質構造データからの表面モチーフ発見、文献からの情報抽出などの研究を融合的に発展させ、三次元構造データ・関係データ・文献データを有機的に連携させた新しいハイブリッドマイニングの方式を提案するとともに、これをタンパク質機能知識の自動発見に応用することを目的とする。具体的な研究項目を以下に列挙する。

(1) タンパク質三次元構造データから導出可能な局所的空間類似性とタンパク質間相互作用データの特徴的対応関係を相補的に利用することにより、タンパク質の機能に関わる部位の特徴を精度良く自動抽出する方式を確立する。

(2) 三次元構造データから得られる距離情報・位相情報、相互作用データから得られる関係情報、さらに文献テキストデータを融合的に処理する手法を考案し、タンパク質の機能に関連する各種知識を発見するための方式を確立する。

３．研究の方法

(1) タンパク質三次元構造の局所的類似性とタンパク質間相互作用との関連性に着目し、タンパク質の機能に関わる部位（機能部位）を自動抽出する方式について検討する。

(2) 三次元構造データとテキストデータの利用により、文献のテキスト中に存在しているタンパク質名の特定やタンパク質機能情報の抽出が、少数の訓練データから可能な方式について検討する。

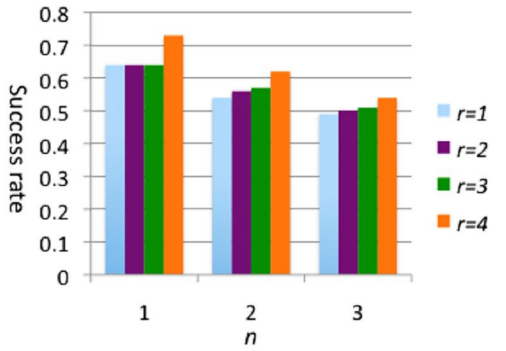
(3) 三次元構造データと機能データを利用することで、文献間の関連性を評価する仕組みを導入し、これをもとに、類似文献を検索する手法について検討する。

(4) タンパク質代謝ネットワークをもとに、生物種の類縁関係に関する知識の獲得手法について検討する。

４．研究成果

(1) 三次元構造データと相互作用データのハイブリッド利用によるタンパク質機能部位抽出

①タンパク質立体構造データとタンパク質ータンパク質相互作用ネットワークを併用することにより、機能未知タンパク質の相互作用部位を抽出する手法を提案した。提案手法では、あるタンパク質に対してネットワーク上で近傍に位置するタンパク質(周辺タンパク質)が類似機能を有することに着目し、周辺タンパク質をその立体構造類似度をもとに、特徴的部位を共有するクラスタに分類する。そして、各クラスタに機能未知タンパク質を追加して構成されるタンパク質グループから構造や物性が類似する部分構造をマイニングすることで相互作用部位を特定する。さらに、予測結果を既知機能部位と仮定し、クラスタ再構築を繰り返すことにより、精度向上を実現した。相互作用部位が既知である蛋白質の実データをもとに抽出実験を行い、三次元構造データと相互作用データの相補的利用が機能部位抽出に有効であることを示した（下図参照)。



n: # of proteins with unknown interaction site

r: # of cycles in repetitive prediction process

②タンパク質分子表面のグラフ表現データをもとに，頻出する類似部分グラフを発見することにより、与えられたタンパク質の結合部位を予測する新しい手法を提案した。提案手法では、タンパク質のリガンド結合特異性に注目し、結合リガンドに基づく分類データを活用することで、特定のリガンド結合グループのみに高頻度に出現する局所構造をグラフマイニングの枠組で抽出することにより、結合部位の予測を試みていることに特徴がある。このとき、類似リガンドに結合する複数のグループ間において、結合部位の構造も類似することを考慮し、グループの自動統合の枠組を導入することで、予測精度の向上を図った。5 つのグループに所属する 37 の蛋白質分子表面データに対して結合部位予測手法を適用した結果、3 つのグループに対して 40%以上の精度で、第 1 位に結合部位を予測することに成功した。また、統合処理を利用することで、その精度を 50%以上に向上で

きることを確認した。

(2) 三次元構造データとテキストデータのハイブリッド利用による文献からのデータマイニング
①文献テキストデータからのマイニングにおいて、文献中のタンパク質名を的確に特定することが重要である。そこで、特に訓練例が十分に得られない場合に、訓練例拡張により、高い判定精度でタンパク質名が特定可能な手法を提案した。提案手法では、訓練例の一部を妥当性チェック用集合とし、この文集合に対してタンパク質名のタグ付けを行う。タグ付け結果を評価し、タグ付け精度が向上するように外部コーパスからの文選択を行い、訓練例を拡張する。以上の処理を繰り返すことで効果的な文の選択を実現し、タンパク質名を高精度に特定することに成功した。
②タンパク質構造解析に関する文献を対象として、該当するタンパク質の構造データと文献テキストデータの両者から構成される特徴空間を構築し、機械学習によりタンパク質機能情報を抽出する方式を提案した。特に、能動学習と半教師付き学習の考え方を導入することで、小規模な訓練データを用いて初期学習された分類器の出力結果の信頼度を、正解が判明しているデータ集合との特徴空間上での距離分布に基づき算出し、信頼度が高いデータ集合を次の学習サイクルにおける訓練データセットに追加するとともに、信頼度が低いデータ集合に関してユーザからのフィードバックを得ることで、少数の訓練データに基づく効果的な学習を実現した。
③テキストデータのマイニングにおいて、構造データの相補的利用が重要な役割を果たす一方で、異分野の訓練データの活用を図る転移学習を導入する際には、容易に構造データが利用できるとは限らないことが問題となる。そこで、新しい転移学習の枠組である選択的転移学習手法を提案した。提案手法では、構造データから得られる特徴が、情報抽出結果に対してどのように寄与するかに着目することにより、抽出対象テキストを、転移学習の適用が有効と期待される文集合とそうでない文集合に分割する。これにより、構造データの利用と転移学習の併用を可能とし、相互作用情報の抽出精度向上を達成した。

(3) 構造データと機能データの利用による文献間関連性の抽出とその関連文献検索支援への応用を図った。具体的には、タンパク質構造解析に関する文献に内包されているタンパク質機能情報や構造情報を、PDB、SCOP、PubMed、GO、PROSITE などの関連データベースの統合利用により抽出し、これをもとに、文献間の関連性を評価可能な方式を提案した。特に、利用者によって適切に選択された少数の入力文献をもとに、概念間の関連性を調整することにより、利用者の意図の反映を可能とし、関連文献検索支援への応用を図った。 2 つの文献に対する共引用文献を正解文献とする検索実験を行い、PubMed による検索結果と比較した結果、提案手法と PubMed における MAP(Mean Average Precision)値はそれぞれ 0.725, 0.660 となり、提案手法が関連文献をより的確に検索できることを確認した。

(4) 定常状態を満たす代謝構造を利用することで，生物種間の系統分類に新たな知見を与えることが可能との観点から、代謝反応の重要度に基づくペナルティ付き pseudo alignment による代謝パスウェイ比較手法を提案した。提案手法では、定常状態を満たす最小の部分構造である elementary flux mode を抽出し、この部分代謝構造間での類似性をもとに生物種間の比較を行う。このとき、重要な酵素の欠損には、より大きなペナルティを与えるべきであるとの考えに基づき、酵素の重要性をペナルティとして用いた pseudo alignment を導入している。抽出した elementary flux mode を対象に、酵素反応の重要度をペナルティとする pseudo alignment を行うことにより、機能の観点に基づく生物種間距離の算出を実現する。38 種の生物に対し、提案手法を用いて比較解析実験を行い、系統樹を構築した結果、3 ドメイン説を生体内の機能的観点から再現することに成功した。また，好気性、嫌気性など、生存環境に関わる機能の観点からの分類が、高い精度で可能となり、提案手法の有効性を確認した。

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 7 件）（総計 16 件）

① Kazunori Miyanishi and Takenao Ohkawa: A method of extracting sentences containing protein function information from articles by iterative learning with feature update, Proceedings of the Ninth International Meeting on Computational Intelligence Methods for Bioinformatics and Biostatistics, 10pages, 査読有 (To appear).
② Hiroyuki Monji, Satoshi Koizumi, Tomonobu Ozaki, and Takenao Ohkawa: Interaction site prediction by structural similarity to neighboring clusters in protein-protein interaction networks, BMC

Bioinformatics, Vol.12 (Suppl 1), S39, 10pages, 査読有 (2011).

③ Riku Kyogoku, Ryo Fujimoto, Tomonobu Ozaki, and Takenao Ohkawa: A method for supporting retrieval of articles on protein structure analysis considering users' intention, BMC Bioinformatics, Vol.12 (Suppl 1), S42, 9pages, 査読有 (2011).

④Kazunori Miyanishi, Tomonobu Ozaki, and Takenao Ohkawa: A method to extract sentences containing protein function information with training data extension based on user's feedback, IPSJ Transactions on Bioinformatics, Vol. 3, pp. 82-90, 査読有 (2010).

⑤ Kazunori Miyanishi, Tomonobu Ozaki, and Takenao Ohkawa: A Method to identify protein names with iterative extension of training data set, Proceedings of the 2nd International Conference on Bioinformatics and Computational Biology (BICoB-2010), pp.178-183, 査読有 (2010).

⑥ Yuta Ashida, Tomonobu Ozaki, and Takenao Ohkawa: A comparative analysis of metabolic pathways based on metabolic steady states, IPSJ Transactions on Bioinformatics, Vol.2, No.4, pp.83-92, 査読有 (2009).

⑦ Kazunori Miyanishi, Tomonobu Ozaki, and Takenao Ohkawa: Selection of effective sentences from a corpus to improve the accuracy of identification of protein names, IPSJ Transactions on Bioinformatics, Vol.2, No.4, pp.93-100, 査読有 (2009).

〔学会発表〕(計 7 件) (総計 11 件)

① 京極 陸, 大川 剛直: 文の接続関係を考慮した蛋白質構造解析文献からの相互作用記述文抽出方法, 情報処理学会 第 25 回バイオ情報学研究会, 2011 年 6 月 23-24 日, 琉球大学

② 文字 宏之, 大川 剛直: タンパク質の立体構造情報と類似部分グラフマイニングを利用した結合部位の自動抽出に関する研究, 情報処理学会 第 25 回バイオ情報学研究会、2011 年 6 月 23-24 日, 琉球大学

③ 麻生 知希, 大川 剛直: 利用者の意図を考慮した概念的観点に基づく蛋白質構造解析文献検索手法, 情報処理学会 第 25 回バイオ情報学研究会, 2011 年 6 月 23-24 日, 琉球大学

④ 宮西 一徳, 尾崎 知伸, 大川 剛直: ユーザフィードバックに基づく訓練データ拡張を伴う蛋白質機能情報文抽出に関する研究, 情報処理学会 第 23 回バイオ情報学研究会, 2010 年 12 月 16-17 日, 九州大学

⑤ 小泉 敏史, 尾崎 知伸, 大川 剛直: 蛋白質-蛋白質相互作用ネットワークにおける周辺クラスタとの立体構造類似度を考慮した相互作用部位予測, 人工知能学会 第 77 回 人工知能基本問題研究会, 2010 年 3 月 17-18 日, 北海道大学

⑥ 藤本 亮, 尾崎 知伸, 大川 剛直: 概念的類似度の更新に基づく観点を考慮した蛋白質構造解析文献の検索支援, 人工知能学会 第 77 回 人工知能基本問題研究会, 2010 年 3 月 17-18 日, 北海道大学

⑦ 宮西 一徳, 尾崎 知伸, 大川 剛直: コーパスからの文選択による事例集合拡張に基づく蛋白質名判定, 情報処理学会 第 19 回バイオ情報学研究会, 2009 年 12 月 17-18 日, 電気通信大学

６．研究組織

(1)研究代表者

大川 剛直 (OHKAWA TAKENAO)

神戸大学・システム情報学研究科・教授

研究者番号：30223738

(2)研究分担者

(3)連携研究者