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

令和　２ 年　 ６ 月　 ６ 日現在

機関番号：５５２０１
研究種目：基盤研究(C)（一般）
研究期間：2016～2019
課題番号：１６Ｋ００３９２

研究課題名（和文）大域的構造抽出と相互作用確率モデルによるタンパク質機能予測法

研究課題名（英文）Development of protein function prediction methods with global substructures and interaction stochastic models

研究代表者
林田　守広（HAYASHIDA, Morihiro）

松江工業高等専門学校・電気情報工学科・准教授

研究者番号：４０４０２９２９

交付決定額（研究期間全体）：（直接経費）　　3,500,000円

研究成果の概要（和文）：タンパク質の機能を理解するために，タンパク質と相互作用する生体分子や複合体を形成するタンパク質を知ることは有用である．タンパク質とRNAの間の残基と塩基の相互作用予測手法およびヘテロ二量体タンパク質複合体の予測手法を改良した．さらに同じ機能を示すタンパク質やRNAの配列をそれぞれ集約する文字列の集合上の確率分布を推定する手法を開発し，また中央文字列を見つける手法を高速化するとともに，タンパク質のドメイン構成や生物学的ネットワークに対する文法圧縮手法を開発した．

研究成果の学術的意義や社会的意義
タンパク質相互作用およびタンパク質複合体の予測精度を向上させることによってタンパク質の機能を推定する手がかりとし，病原菌の感染経路の解明など生物学，医学へ貢献する．また文字列の集合上の確率分布や中央文字列の厳密解法，一般化Series-Parallelグラフの最小文法を与えることはタンパク質やRNAの配列解析や生物学的ネットワーク構造の解析に有用であるだけでなく数理学的にも意義がある．

研究成果の概要（英文）：It is useful for understanding protein functions to identify biomolecules interacting with proteins and proteins consisted in a complex. We improved prediction methods for protein-RNA residue-base contacts and heterodimeric protein complexes. We developed an estimation method of a Laplace-like mixture on a set of strings, and enhanced the speed of finding median strings. Furthermore, we developed grammar-based compression methods for domain sequences of proteins and generalized series-parallel graphs.

研究分野：生物情報学

キーワード：タンパク質相互作用予測　タンパク質二量体予測　中央文字列　文法圧縮　一般化Series-Parallelグラフ

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

個々のタンパク質と他の生体分子との相互作用を理解することは，タンパク質の機能および関連する病気の原因を明らかにする手がかりとなる．タンパク質は DNA の遺伝子配列をもとに転写と翻訳の過程を経てアミノ酸のポリペプチド鎖として生成され，特定の立体構造を形成することによって機能を示す．タンパク質の中にはドメインと呼ばれる，固有の機能あるいは構造をもつものが存在し，分類方法の違いによっていくつかのデータベースが構築されている．プロファイル隠れマルコフモデルに基づく Pfam データベースや二次構造に基づいた SCOP データベースなどがある．

２．研究の目的

現在までに開発してきたタンパク質相互作用予測やタンパク質複合体予測の高精度化と，タンパク質内部の構造を詳細に理解することによって，タンパク質の機能推定に役立てることである．

３．研究の方法

(1) タンパク質の機能を予測する手法の開発に関して，タンパク質と RNA との間のアミノ酸残基と塩基との相互作用を予測する手法を高精度化する．タンパク質と RNA との相互作用は，RNA のスプライシングや転写後調節，タンパク質の翻訳などに係わっており重要である．またタンパク質と RNA の複合体のリボヌクレオタンパク質について X 線回析法による立体構造の決定がいくつか行われているが，機能が完全に解明されたものは少ない．本研究ではこれまでに提案した条件付き確率場モデルを用いた擬似対数尤度の最大化によるアミノ酸残基と塩基との相互作用予測手法を改良する．条件付き確率場の入力としてはタンパク質アミノ酸配列または RNA 塩基配列の多重配列アラインメントから得られる相互情報量を，アミノ酸残基と塩基の間の進化的な関係の強さを表す量として用いる．

(2) 細胞内の生体分子間の相互作用ネットワークの構造やタンパク質立体構造を解析するためのグラフに対する文法圧縮手法を開発する．これまでの研究では木構造を形成する糖鎖などの生体分子がどのような規則で形成されるのかを，木構造に対する最小の文法を見つけることで解析してきた．しかしながら遺伝子転写制御ネットワークなどの閉路をもったグラフには適用できない．根付きの木構造に対して最小の文法を見つけたときには生成規則として，根における兄弟ノードの分割と，内部ノードにおける親子ノードの分割を考えた．本研究ではこの文法規則を閉路をもつグラフへの自然な拡張となるように文法圧縮を行う対象のグラフを検討し，最小の文法を見つけることでグラフを圧縮する手法を開発する．

(3) タンパク質の部分構造であるドメインに注目し一つの生物種におけるタンパク質全体についてドメイン構成の進化的側面から解析を行う．これまでの研究ではタンパク質に含まれる機能に着目し，ドメインを要素とする集合の族に対する文法圧縮の手法を開発した．集合の生成規則としては，突然変異によるドメイン集合の生成と遺伝子重複によるドメイン集合の複製を考慮した．一方でタンパク質はアミノ酸残基が直鎖状に並んだ構造であるため，ドメインもアミノ酸残基に従って並べることができる．本研究ではドメインに順序付けを行ったときの文法圧縮の手法を開発するとともに多数の生物種に適用し検証する．

(4) タンパク質アミノ酸配列や核酸塩基配列などは文字列によって表現される．同じ機能を示す複数のアミノ酸配列または塩基配列があったときに，代表となる文字列，または複数の文字列を生起する，文字列の集合上の確率分布を見つける手法を開発する．

(5) タンパク質間相互作用の解明によって，病原体と宿主の間の相互作用であれば病原体感染のメカニズム解明および感染症の新たな治療法の発見につながる可能性がある．宿主内におけるタンパク質間相互作用ネットワークのあるノード周辺の局所的なグラフ構造に基づいた特徴量およびアミノ酸配列に基づいた特徴量を用いて，病原体内のタンパク質との相互作用を予測する手法を開発する．

４．研究成果

(1) タンパク質アミノ酸残基と RNA 塩基との間の相互作用を予測する手法の開発について，初年度には多重配列アラインメントから得られる相互情報量を入力として SMM(Support Matrix Machine)，OS-ELM 等の機械学習手法を検討したが，計算機実験による検証の結果は従来手法と同程度の予測精度に留まった．

次年度においては条件付き確率場モデルの改良を目標として，アミノ酸残基と塩基との間の相互作用に対応する確率変数の値を，相互作用するときは 1，しないときは 0 としていたものを，相互作用しないときは -1 としたイジングモデルに似たポテンシャル関数を導入した．従来モデルでは相互作用の確率変数が特徴を選択する形となっていたが，提案モデルでは隣り合うアミノ酸残基および塩基の間の相互作用の有無が同じか否かによってポテンシャル関数が増減する．計算機実験による提案モデルの評価を行うために，先行研究で用いたのと同じタンパク質構造データバンク(PDB)から立体構造が決定された 13 のタンパク質 RNA 複合体を使用した．またアミノ酸の極性や親水性などの特性を考慮し，似た特性をもつアミノ酸は同じとみなして 8,10,15 のグループにそれぞれ分類するとともに，条件付き確率場モデルへの入力として，多重配列アラインメントから得られる相互情報量からバイアスを除いた Dunn ら[①]が提案した MIp あ

るいは Ekeberg ら[②]が提案した plmDCA を用いた．交差検証を用いた評価の結果，MIp，plmDCA のどちらを入力としたときも AUC(Area Under the ROC Curve)の平均値が従来モデルよりも大きくなり予測精度が改善することを確認した．特に提案モデルにおいて，10 または 15 のアミノ酸分類で plmDCA を用いたときに平均 AUC は 0.699 となった．

(2) 生物学的ネットワークのような閉路をもつグラフに対する文法圧縮手法の開発について，木文法に対する分割型文法を一般化 Series-Parallel グラフに拡張した文法を定義し，この文法の下での最小文法を見つけることで閉路をもつグラフを圧縮する手法を開発した．根付き木での分割点は根または内部ノードであったが，提案する文法では対象とするグラフに閉路を含むため 1 ノードでの分割に加えて 2 ノードでの分割に対応する生成規則をもつ．これらの生成規則は 2 端点グラフに対する直列合成，一般化直列合成，並列合成に対応し，木グラフの他，外平面グラフを含む一般化 Series-Parallel グラフを構成する．このグラフの一つが入力として与えられたとき，入力グラフを再構成するために必要な生成規則の数が最小となるように分割するノードを決定する．本研究ではこの問題を整数線形計画問題へ定式化することで解く方法を開発した．

(3) ある生物種のタンパク質全体におけるドメイン構成が進化的にどのように形成されてきたかを解析するために，一つのタンパク質がもつドメインの列のタンパク質全体での集合を対象として，これまでの研究と同様に，進化の過程に基づく生成規則の下での最小の文法を見つける手法を開発した．生成規則としては，突然変異によるドメイン列の生成と遺伝子重複によるドメイン列の複製を考えるが，複製においては複製元のタンパク質のドメイン列から複製先のタンパク質のドメイン列への，ドメインの挿入によるレーベンシュタイン距離を生成規則のコストとして導入した．一つの生物種に含まれるタンパク質全体を構成するすべての生成規則の中から最小コストの文法を見つける問題は最小全域木問題に帰着され大域最適解を多項式時間で見つけることができる．UniProt データベースに含まれる生物種のうち，真核生物 73 種，細菌 328 種，古細菌 14 種にそれぞれ提案手法を適用した結果，細菌，古細菌では大きく圧縮された種はほとんどなく，真核生物ではばらつきがあり近縁種であっても圧縮率に違いが出た．

(4) 同じ機能を示すような複数の配列に対する代表となる配列として中央文字列または中心文字列が知られている．文字列の集合上の確率分布と文字列の間の距離が与えられたとき，中央文字列は与えられた文字列との距離の和を最小とする文字列と定義され，中心文字列は与えられた文字列との距離の最大を最小とする文字列と定義される．これまでの研究では文字列間の距離にレーベンシュタイン距離を用いた場合に整数線形計画問題への定式化を通して中央文字列，中心文字列を見つける手法を提案していた．本研究ではレーベンシュタイン距離が三角不等式制約を満たすことから，整数線形計画問題へ新たな制約式を追加することで計算時間を大幅に削減できることを示した．

一方で同じ機能を示す複数の配列は文字列の集合上の一つの確率分布から生起したと考えることができる．本研究では文字列の集合上にラプラス様混合確率分布を導入し，尤度に EM(期待値最大化)アルゴリズムを用いることで確率分布を推定する手法を開発した．Rfam データベースから取得した 6 つの配列ファミリー，計 289 本の RNA に混合数を 6 として提案手法を適用し 6 つの中心となる文字列を推定するとともに有効性を確認した．

(5) 病原体と宿主の間でのタンパク質間相互作用を予測する手法を開発した．宿主内でのタンパク質相互作用ネットワークにおける対象となるタンパク質を含む 5 ノード以下の部分グラフの頻度から得られる特徴量およびアミノ酸配列に基づく特徴量を用い，分類器として確率的勾配降下法および SCW(Soft Confidence Weighted learning)法を組み合わせた．4 種の病原体，B. anthracis，F. tularensis，S. typhi，Y. pestis，それぞれとヒトとの間のタンパク質に適用し交差検証による計算機実験の結果，提案手法の F 値が既存手法を上回った．さらにネットワークの局所構造が予測の精度向上に寄与することを確認した．

(6) 細胞内の複合体はいくつかの分子が結合することによって機能を示す．本研究では二量体を形成する異なる二つのタンパク質を予測する手法の精度を改善した．二つのタンパク質の相互作用の信頼度，系統学的なプロファイルを特徴量として，MLPK (Metric Learning Pairwise Kernel)，TPPK (Tensor Product Pairwise Kernel) の二つのペアワイズカーネルを適用した．酵母菌の 152 のヘテロ二量体を含むデータセットを用いた交差検証の結果，正規化した最小カーネルと MLPK の組み合わせによる F 値が 0.686 となり予測精度を改善させた．

(7) 遺伝子発現プロファイルから肺がんに罹患しているかどうかを判別する手法を開発した．ヒトのタンパク質相互作用ネットワークにスペクトラルクラスタリングを適用し遺伝子発現プロファイルと組み合わせた結果を畳み込みニューラルネットワークの入力とした．GEO データベースの GSE66499 から取得した 380 のサンプルを用いて交差検証を行った結果，提案手法は RBF カーネルを用いたサポートベクトルマシン，あるいはランダムフォレストよりも高い予測精度を達成した．

＜引用文献＞

① Dunn, S.D. *et al.*, Bioinformatics, 24, 333-340, 2008.

② Ekeberg, M. *et al.*, Phys. Rev. E, 87, 012707, 2013.

５．主な発表論文等

〔雑誌論文〕　計9件（うち査読付論文　9件／うち国際共著　4件／うちオープンアクセス　6件）

１．著者名 Koyano Hitoshi、Hayashida Morihiro、Akutsu Tatsuya	４．巻 106
２．論文標題 Optimal string clustering based on a Laplace-like mixture and EM algorithm on a set of strings	５．発行年 2019年
３．雑誌名 Journal of Computer and System Sciences	６．最初と最後の頁 94～128
掲載論文のDOI（デジタルオブジェクト識別子） 10.1016/j.jcss.2019.07.003	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 Matsubara Teppei、Ochiai Tomoshiro、Hayashida Morihiro、Akutsu Tatsuya、Nacher Jose C.	４．巻 17
２．論文標題 Convolutional neural network approach to lung cancer classification integrating protein interaction network and gene expression profiles	５．発行年 2019年
３．雑誌名 Journal of Bioinformatics and Computational Biology	６．最初と最後の頁 1940007～1940007
掲載論文のDOI（デジタルオブジェクト識別子） 10.1142/S0219720019400079	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Takuya Mori, Hayliang Ngouv, Morihiro Hayashida, Tatsuya Akutsu, Jose Nacher	４．巻 12
２．論文標題 ncRNA-disease association prediction based on sequence information and tripartite network	５．発行年 2018年
３．雑誌名 BMC Systems Biology	６．最初と最後の頁 37
掲載論文のDOI（デジタルオブジェクト識別子） 10.1186/s12918-018-0527-4	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 Natsu Nakajima, Morihiro Hayashida, Jesper Jansson, Osamu Maruyama, Tatsuya Akutsu	４．巻 13
２．論文標題 Determining the minimum number of protein-protein interactions required to support known protein complexes	５．発行年 2018年
３．雑誌名 PLOS ONE	６．最初と最後の頁 e0195545
掲載論文のDOI（デジタルオブジェクト識別子） 10.1371/journal.pone.0195545	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 該当する

1．著者名	4．巻
Morihiro Hayashida, Noriyuki Okada, Mayumi Kamada, Hitoshi Koyano	6
2．論文標題	5．発行年
Improving conditional random field model for prediction of protein-RNA residue-base contacts	2018年
3．雑誌名	6．最初と最後の頁
Quantitative Biology	155～162
掲載論文のDOI（デジタルオブジェクト識別子）	査読の有無
10.1007/s40484-018-0136-7	有
オープンアクセス	国際共著
オープンアクセスとしている（また、その予定である）	－

1．著者名	4．巻
Yu Bao, Morihiro Hayashida, Pengyu Liu, Masayuki Ishitsuka, Jose Nacher, Tatsuya Akutsu	25
2．論文標題	5．発行年
Analysis of critical and redundant vertices in controlling directed complex networks using feedback vertex sets	2018年
3．雑誌名	6．最初と最後の頁
Journal of Computational Biology	1071～1090
掲載論文のDOI（デジタルオブジェクト識別子）	査読の有無
10.1089/cmb.2018.0019	有
オープンアクセス	国際共著
オープンアクセスではない、又はオープンアクセスが困難	－

1．著者名	4．巻
Ruan Peiying、Hayashida Morihiro、Akutsu Tatsuya、Vert Jean-Philippe	19
2．論文標題	5．発行年
Improving prediction of heterodimeric protein complexes using combination with pairwise kernel	2018年
3．雑誌名	6．最初と最後の頁
BMC Bioinformatics	39
掲載論文のDOI（デジタルオブジェクト識別子）	査読の有無
10.1186/s12859-018-2017-5	有
オープンアクセス	国際共著
オープンアクセスとしている（また、その予定である）	該当する

1．著者名	4．巻
Morihiro Hayashida, Hitoshi Koyano	690
2．論文標題	5．発行年
Finding median and center strings for a probability distribution on a set of strings under Levenshtein distance based on integer linear programming	2017年
3．雑誌名	6．最初と最後の頁
Communications in Computer and Information Science	108-121
掲載論文のDOI（デジタルオブジェクト識別子）	査読の有無
10.1007/978-3-319-54717-6_7	有
オープンアクセス	国際共著
オープンアクセスではない、又はオープンアクセスが困難	該当する

1．著者名 Yu Bao, Morihiro Hayashida, Tatsuya Akutsu	4．巻 17
2．論文標題 LBSizeCleav: improved support vector machine (SVM)-based prediction of Dicer cleavage sites using loop/bulge length	5．発行年 2016年
3．雑誌名 BMC Bioinformatics	6．最初と最後の頁 487
掲載論文のDOI（デジタルオブジェクト識別子） 10.1186/s12859-016-1353-6	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 該当する

〔学会発表〕　計7件（うち招待講演　0件／うち国際学会　7件）

1．発表者名 Teppei Matsubara, Tomoshiro Ochiai, Morihiro Hayashida, Tatsuya Akutsu, Jose Nacher
2．発表標題 Convolutional neural network approach to lung cancer classification integrating protein interaction network and gene expression profiles
3．学会等名 The 18th IEEE International Conference on Bioinformatics and BioEngineering（国際学会）
4．発表年 2018年

1．発表者名 Morihiro Hayashida, Kousei Ishibashi, Hitoshi Koyano
2．発表標題 Analyzing order of domains in grammar-based compression of proteomes
3．学会等名 The 24th International Conference on Parallel and Distributed Processing Techniques and Applications（国際学会）
4．発表年 2018年

1．発表者名 Morihiro Hayashida, Jose Nacher, Hitoshi Koyano
2．発表標題 Artificial neural network approach to prediction of protein-RNA residue-base contacts
3．学会等名 The 12th International Conference on Bioinformatics Models, Methods and Algorithms（国際学会）
4．発表年 2019年

１．発表者名
Morihiro Hayashida, Noriyuki Okada, Mayumi Kamada, Hitoshi Koyano

２．発表標題
Improving conditional random field model for prediction of protein-RNA residue-base contacts

３．学会等名
The 11th International Conference on Computational Systems Biology（国際学会）

４．発表年
2017年

１．発表者名
Takuya Mori, Hayliang Ngouv, Morihiro Hayashida, Tatsuya Akutsu, Jose Nacher

２．発表標題
ncRNA-disease association prediction based on sequence information and tripartite network

３．学会等名
The 16th Asia Pacific Bioinformatics Conference（国際学会）

４．発表年
2018年

１．発表者名
Morihiro Hayashida, Hitoshi Koyano, Tatsuya Akutsu

２．発表標題
Grammar-based compression for directed and undirected generalized series-parallel graphs using integer linear programming

３．学会等名
The 9th International Conference on Bioinformatics Models, Methods and Algorithms（国際学会）

４．発表年
2018年

１．発表者名
Jira Jindalertudomdee, Morihiro Hayashida, Jianning Song, Tatsuya Akutsu

２．発表標題
Host-pathogen protein interaction prediction based on local topology structures of a protein interaction network

３．学会等名
IEEE 16th International Conference on Bioinformatics and BioEngineering（国際学会）

４．発表年
2016年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考
研究分担者	小谷野　仁 (KOYANO Hitoshi) (10570989)	国立研究開発法人農業・食品産業技術総合研究機構・農業情報研究センター・研究員 (82111)	