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　３ 年　 ６ 月　 ８ 日現在

機関番号：１７１０２
研究種目：基盤研究(C)（一般）
研究期間：2017～2020
課題番号：１７Ｋ００４０７

研究課題名（和文）混合正則化モデリングを軸としたヘテロ生物データ群からの機械学習の研究

研究課題名（英文）Study on machine learning approaches for heterogeneous biological data based on mixing regularization models

研究代表者
丸山　修（Maruyama, Osamu）

九州大学・芸術工学研究院・准教授

研究者番号：２０２８２５１９

交付決定額（研究期間全体）：（直接経費）　　3,400,000円

研究成果の概要（和文）：「タンパク質複合体予測問題」と「E3ユビキチン・ライゲース結合部位予測問題」に対して，評価関数のモデル化とギブス・サンプリング・アルゴリズムに基づく最適化技法を開発し，一定の成果を得ることができた．とくに，「E3ユビキチン・ライゲース結合部位予測問題」に対しては，生物学的知見に基づき，複雑な尤度関数と複数の事前分布を設計し，これらから得られる事後確率を最適化する崩壊型のギブス・サンプリング・アルゴリズムを提案することができた．また，計算機実験において既存手法より予測精度が優れていることを示した．

研究成果の学術的意義や社会的意義
本研究では，データ・ドリブンな研究手法として，データに基づくモデリングを行い評価関数を設計するスキームを確立することが出来た．とくに，事後確率によるモデリングの柔軟性の高さとギブス・サンプリングをベースとした最適化の汎用性の組み合わせの利点を示すことが出来た．また，取り組んだ生物学的個別課題「E3ユビキチン・ライゲース結合部位予測問題」に関しては，タンパク質配列上の結合部位の予測情報は新規薬剤の設計などに応用できる可能性がある．

研究成果の概要（英文）：For the problem of protein complex prediction and that of E3 ubiquitin ligase binding site prediction, I got successful results to some extent, by modeling target data, designing evaluation functions, and constructing optimization algorithms based on the Gibbs sampling algorithm. Particularly, for the problem of E3 ubiquitin ligase binding site prediction, I designed complicated likelihood functions, the multiple prior distributions based on biological knowledge, and collapsed Gibbs sampling algorithm for the posterior probability distribution derived from the likelihood functions and the prior probability distributions. I showed that the proposed method is superior to existing methods in prediction accuracy on our target datasets.

研究分野：バイオインフォマティクスとComputational biology

キーワード：モチーフ　ギブス・サンプリング　結合部位　E3ユビキチン　タンパク質disorder　混合正則化　バイオインフォマティクス　デグロン

様　式　C－１９、F－１９－１、Z－１９（共通）

１．研究開始当初の背景

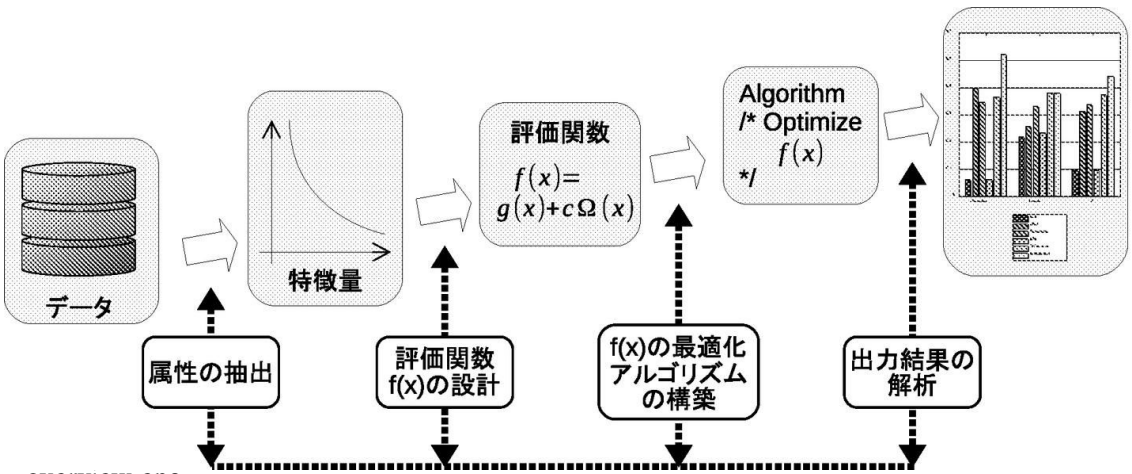
バイオインフォマティクスや Computational biology 分野の最適化問題における評価関数の設計は重要である．評価関数の設計は既存の生物学的データの数理的解釈や表現に他ならず，評価関数は出力された解の解析結果を大きく左右するからである．そこで，評価関数の（最適）設計を優先し，定式化された評価関数の最適化はマルコフ連鎖モンテカルロ法に基づく汎用最適化サンプリング・アルゴリズムに任せるという枠組みの研究を遂行することとした．

２．研究の目的

（１）混合正則化モデリングを軸に，階層ベイズモデリングや近似モデリングなどの様々なモデリングを最優先で追求し評価関数 f(x)を設計する．
（２）最適化関数に比較的制約を設けない．
（３）マルコフ連鎖モンテカルロ法という汎用的最適化の枠組みを用いて f(x)を最適化するアルゴリズムを構築する．
この作戦のもと，次の２つの「生物データ群からの予測問題」に対して予測手法の向上を目的とした：
（A）タンパク質複合体予測問題
（B）E3 ユビキチン・ライゲース結合部位予測問題

３．研究の方法

研究目的で挙げた個別問題に対して，次の基本ステップを試行錯誤的に繰り返し研究を着実に遂行することとした（図参照）：



overview.eps

ステップ１：予測対象の関連データベースや入力データに対し，特徴的な属性の抽出や学習を行い，混合正則化モデリングを軸に様々なモデリングの検証を行い，評価関数 f(x)を定式化する．
ステップ２：まずは f(x)のための汎用最適化アルゴリズムを構築し実働化するが，f(x)の形が固まれば最適化アルゴリズムの高速化を検討する．
ステップ３：計算機実験を行い，結果を解析し，既存手法のものと比較し，改善の糸口を探る．

４．研究成果

「タンパク質相互作用ネットワークからのタンパク質複合体予測問題」に関して，次の内容の論文を発表することができた：
（１）予測されたタンパク質複合体の相互の重なりをモデル化した．
（２）（１）に基づく評価関数を最適化するサンプリング・アルゴリズムを構築した．
（３）（２）のアルゴリズムを実装し比較実験を行い，その有効性を実証した．
相互に重なりあるタンパク質複合体の数理的モデルを提案できたことは成果であり，また，そのような角度も考慮してタンパク質複合体予測問題を考えるべきという結果を得たことは意義深い．しかし，最初の個別課題として挙げていた「タンパク質複合体の内部構造モデリングに基づくタンパク質複合体予測手法の開発」に関しては，いろいろと試行錯誤したが否定的な結果しか得られなかった．理由としては，次が挙げられる：
1．相互排他的な関係にあるタンパク質間相互作用の情報のタンパク質複合体予測問題に対する貢献度が比較的高くないかもしれない．

2．相互排他的な関係にあるタンパク質間相互作用の情報のモデル化が不適切だったかもしれない．
3．相互排他的な関係にあるタンパク質間相互作用の情報の他に，本問題に有効と思えるゲノムワイドなデータが存在しなかった．

次に「E3 ユビキチン・ライゲース結合部位予測のための崩壊型ギブス・サンプリング・アルゴリズム DegSampler」を発表した．この手法は次の３つの特徴を有する．
（１）配列モチーフの出現位置の事前情報は通常一様分布であるが，本研究では，タンパク質の各サイトの disorder 値を基にした事前情報を定式化している．これにより，E3 ユビキチン・ライゲースの結合部位の予測精度が格段に向上することを確認している．
（２）タンパク質配列を構成するアミノ酸残基はそれぞれ特有の化学的特性を有する．とくに，結合部位の既知のコンセンサス・パターンを見ると，各サイトは極性，無極性，正電荷，負電荷で分類できることが分かる．そこで，この特性を数理モデルで捉えた尤度関数を定式化し提案手法 DegSampler の事後確率に組み込んでいる．
（３）タンパク質の機能領域に関するデータベースである ELM(Eukaryotic Linear Motif) を用いて，提案手法 DegSampler の性能を網羅的に評価し，既存手法より優れていることを確認している．
この研究では，E3 ユビキチン・リガーゼが結合する基質の部位を予測する配列モチーフを推定するためにの崩壊型ギブス・サンプリング・アルゴリズムを構築した．このアルゴリズムにより最適化される事後確率は次の特徴を有している：
(1) アミノ酸の化学特性に基づく尤度関数を有している．
(2) モチーフ出現位置の事前分布としてタンパク質配列の各位置の disorderness を利用する．
(3) 崩壊型のギブス・サンプリングであるため計算の効率が高く結果のブレが相対的に少ない．
以上の特徴をもつ DegSampler を用いた 36 個の E3 に対する計算機実験を行ったところ，既存手法よりも格段に良い予測精度を得た．

引き続き，「E3 ユビキチン・ライゲース結合部位予測問題」に取り組み，崩壊型ギブス・サンプリング・アルゴリズムの改良を行った．確率配列モチーフのでデファクトスタンダードは位置依存スコア行列(PSSM; position-specific scoring atrix)である．これは各位置に独立なカテゴリカル分布が対応することにより構成されるモデルである．このモデルの欠点は，モチーフの位置間の独立性を仮定していることである．予測精度向上のアプローチとして，位置依存スコア行列(PSSM; position-specific scoring atrix)の相異なる位置に出現する文字間の依存関係を許す配列モチーフモデルを定式化し，これを最適化するアルゴリズムを開発した．そして次の結果を得ている：
（１）調査対象として選んだ 36 種類の E3 特異的基質タンパク質配列集合に対する E3 ユビキチン・ライゲースの結合部位の予測精度は，DegSampler version 3 とその前の version とほぼ同程度であった．
（２）しかしながら，36 種類の結果を個別に見ていくと，予測精度が大きく違っている場合が 36 個中 9 個存在した．これはモチーフ・モデルごとに得意不得意があることと示唆するという興味深い結果である．

５．主な発表論文等

〔雑誌論文〕　計5件（うち査読付論文　5件／うち国際共著　1件／うちオープンアクセス　2件）

１．著者名 Maruyama Osamu、Matsuzaki Fumiko	４．巻 NA
２．論文標題 DegSampler3: Pairwise Dependency Model in Degradation Motif Site Prediction of Substrate Protein Sequences	５．発行年 2019年
３．雑誌名 Proceedings - 2019 IEEE 19th International Conference on Bioinformatics and Bioengineering, BIBE 2019	６．最初と最後の頁 11-17
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/BIBE.2019.00012	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 Nakajima Natsu、Hayashida Morihiro、Jansson Jesper、Maruyama Osamu、Akutsu Tatsuya	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 Maruyama Osamu、Matsuzaki Fumiko	４．巻 1
２．論文標題 [Regular Paper] DegSampler: Collapsed Gibbs Sampler for Detecting E3 Binding Sites	５．発行年 2018年
３．雑誌名 Proc. of 2018 IEEE 18th International Conference on Bioinformatics and Bioengineering (BIBE)	６．最初と最後の頁 1-9
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/BIBE.2018.00009	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 Maruyama Osamu、Kuwahara Yuki	４．巻 18
２．論文標題 RocSampler: regularizing overlapping protein complexes in protein-protein interaction networks	５．発行年 2017年
３．雑誌名 BMC Bioinformatics	６．最初と最後の頁 491
掲載論文のDOI（デジタルオブジェクト識別子） 10.1186/s12859-017-1920-5	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 Maruyama Osamu	4．巻 28
2．論文標題 Two Challenging Difficulties of Protein Complex Prediction	5．発行年 2018年
3．雑誌名 Agriculture as a Metaphor for Creativity in All Human Endeavors. FMfI 2016. Mathematics for Industry	6．最初と最後の頁 139-145
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/978-981-10-7811-8_14	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 －

〔学会発表〕　計5件（うち招待講演　0件／うち国際学会　1件）

1．発表者名 Osamu Maruyama
2．発表標題 DegSampler3: Pairwise dependency model in degradation motif site prediction of substrate protein sequences
3．学会等名 2019 IEEE 19th International Conference on Bioinformatics and Bioengineering, BIBE 2019（国際学会）
4．発表年 2019年

1．発表者名 丸山　修
2．発表標題 E3結合部位予測のための崩壊型ギブス・サンプラーDegSampler
3．学会等名 分子生物情報研究会（SIG-MBI）
4．発表年 2019年

1．発表者名 丸山　修
2．発表標題 正負例配列集合のためのコンセンサス・モチーフによるクラスタリング・アルゴリズム
3．学会等名 日本バイオインフォマティクス学会（JSBi）九州地域部会セミナー宮崎開催
4．発表年 2017年

１．発表者名 丸山　修
２．発表標題 Regularizing protein complexes by mutually exclusive protein-protein interactions
３．学会等名 第6回生命医薬情報学連合大会(IIBMP 2017)
４．発表年 2017年

１．発表者名 Ryo Shimizu, Wan Kin Au Yeung, Hidehiro Toh, Hiroyuki Sasaki and Osamu Maruyama
２．発表標題 Predicting Discriminative Motifs for DNA Methylation in Mammalian Development
３．学会等名 バイオインフォマティクス学会年会　第9回生命医薬情報学連合大会 IIBMP2020
４．発表年 2020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

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関