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　　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：　2014～2016
課題番号：　２６３３０３３０

研究課題名（和文）大規模バイオデータに対する混合正則化モデリングと最適化サンプリング技法の研究

研究課題名（英文）Mixture modeling of regularization terms with optimization sampling strategies and its application to biological large scale data

研究代表者

丸山　修（Maruyama, Osamu）

九州大学・マス・フォア・インダストリ研究所・准教授

研究者番号：２０２８２５１９

交付決定額（研究期間全体）：（直接経費）　　3,700,000円

研究成果の概要（和文）：正則化モデリングとマルコフ連鎖モンテカルロ法に基づく最適化を軸に，タンパク質複合体予測問題とガウス分布のベイズ推定問題に対して成果を得ることが出来た．とくに，タンパク質複合体予測問題に関しては，相互に排他的なタンパク質間相互作用に基づく正則化項の有効性の検証やサイズ２，３の小さいタンパク質複合体の教師付き学習の手法の開発など様々な角度から研究を展開した．
さらに，タンパク質複合体予測問題の挑戦的課題である「個々のタンパク質複合体同士が共通のタンパク質を共有すること」を制御する正則化項のモデル化に取り組み，実施した計算機実験において，その提案手法は既存手法よりも優れた予測率することを報告している．

研究成果の概要（英文）：Based on regularization modeling and Markov chain Monte Carlo algorithms, we have developed methods for the protein complex prediction problem and the Bayes estimation of Gaussian distributions. Especially, for the protein complex prediction problem, we have empirically shown the effectiveness of a regularization term based on the information of mutually exclusive protein-protein interactions. In addition, we have developed a supervised learning algorithm for protein complexes with 2 or 3 components. Furthermore, we have designed a regularization term for controlling overlaps between predicted complexes, and showed that the new method with that regularization term outperforms others.

研究分野：　バイオインフォマティクス

キーワード：　正則化　マルコフ連鎖モンテカルロ法　タンパク質複合体　ガウス分布　ベイズ推定　タンパク質間相互作用　教師付き学習　べき法則

## １．研究開始当初の背景

バイオインフォマティクス分野では，与えられたデータをうまく説明するモデル・パラメータなどの仮説を推定する問題が盛んに研究されている．そのような問題においては，仮説の良し悪しを定める評価関数が必要となるが，特に「生物学的に」自然な解を高評価とする仕組みを有する評価関数の設計が生物学的問題に関しては重要である．
この問題に対するアプローチの１つは，正則化モデリングである．正則化モデリングとは，べき乗則や L1 ノルムなどの正則化（罰則）項を評価関数に加えるモデリングであり，汎化誤差を小さく抑えた自然な解を見つけ出すための技法である．しかしながら，一般に推定問題の説明対象は，複雑で多様である．そこで，複数の正則化項を同時に用いる「混合正則化」の考え方に基づき個々の推定や予測の問題に対する評価関数を設計し，精度の向上の研究を実施することとした．

## ２．研究の目的

「混合」正則化モデリングとは，「複数」の正則化項を用いた複雑な評価関数の設計を目指すモデリングである．本研究では，具体的な題材としてタンパク質複合体予測問題などのバイオインフォマティクス分野の重要な予測・推論問題に対して，混合正則化モデリングにより評価関数を定式化し，これをマルコフ連鎖モンテカルロ法に基づくサンプリング・アルゴリズムで最適化する手法を追求することとした（図１参照）．また，個別問題の各手法の背後に共通に存在する本枠組みの原理を明確化・体系化し再利用可能な形にまとめることを試みることとした．

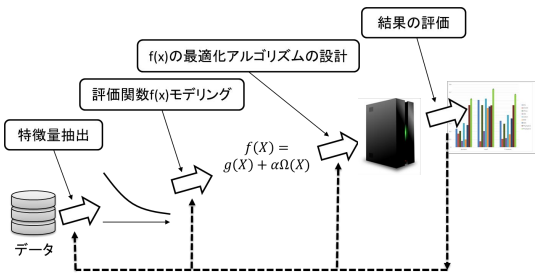


図１　推定プロセスの全体像

## ３．研究の方法

バイオインフォマティクス分野の個別の問題に対して，次の基本ステップの試行錯誤を繰り返しながら実施していくことにより新しい最適化手法を考案することとした．
ステップ１：対象問題の関連データベースの解析や仮説の検証を行い，特徴的な属性変数を同定する．または，最適化の計算の負荷を軽減するなどの妥当な仮定から制約条件を定める．
ステップ２：ステップ１の結果を含めた混合正則化モデリングにより評価関数 f(x)を定式化する．
ステップ３：マルコフ連鎖モンテカルロ法に基づく f(x)を最適化するアルゴリズムを構築し計算プログラムを実装する．
ステップ４：計算機実験を行い，結果を既存手法のものと比較する．

## ４．研究成果

（１）「タンパク質複合体予測問題」に対しては次の研究成果を得ている．

① 排他的タンパク質間相互作用情報に基づく正則化項の設計（発表論文 ②）：

相互に排他的なタンパク質間相互作用(mutually exclusive protein-protein interaction)とは，同時には実現し得ないがそれぞれの存在は確認されている２つのタンパク質間相互作用である（図２参照）．

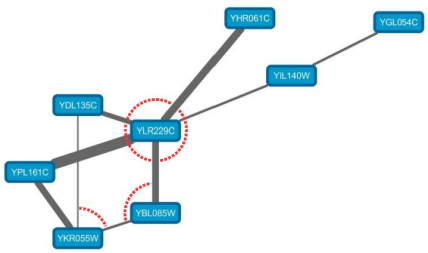


図２ 排他的タンパク質間相互作用の例　赤の破線で繋がっているタンパク質間相互作用は互いに排他的である．

この情報は正確なタンパク質複合体予測に有効と考えられるので，この情報に基づく正則化項を設計し，すでにいくつかの他の正則化項を含む評価関数に組み込むことにより新しい予測手法 PPSampler2-PIME を開発した．計算機実験により，大幅ではないが確実に予測精度が改善されることを確認した．

② 共通のタンパク質を含む相異なるタンパク質複合体を予測する手法の開発（発表論文①，④）：

既知のタンパク質複合体の中には，互いに共通のタンパク質を含む複数のタンパク質複合体が知られている．これらを予測する問題は意外に難しく，この問題点に真っ向から挑戦している研究は僅かである．
この問題に対して，研究期間中に２つの研究を実施した．１つは，研究代表者がここ数年開発してきた混合正則化に基づく評価関数を最適化するシステム PPSampler2 を繰り返し用いることにより互いに重なり合う予測複合体を生成する手法 ReSAPP の研究である．ReSAPP の結果の方が，PPSampler2 より統計的に有意に良いという結果を得ている．
もう一つの研究は，「個々のタンパク質複合体同士が共通のタンパク質を共有する状態」を直接制御する正則化項を設計する研究である．具体的にそのような正則化項を

Jaccard インデックスの考え方に基づき設計し，また既存手法である PPSampler2 の個々の正則化項を改善した．そして，これらの正則化項を統合した評価関数を最適化するアルゴリズムをマルコフ連鎖モンテカルロ法に基づき開発し，これを予測ツール RocSampler として公開した．実施した計算機実験において，RocSampler は既存手法よりも優れた予測率を有することを報告している．図３は，予測した２つのクラスターが互いに重複しかつそれぞれが既知のタンパク質複合体とマッチしている例を示している．

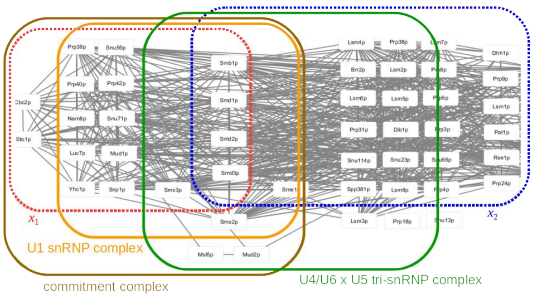


図３ 予測された複合体 $x_1$ と $x_2$ は互いに重複しかつ既知のタンパク質にマッチしている例

③ 最小サイズ（２と３）のタンパク質複合体の教師付き学習の手法の開発（発表論文③，⑥）：
この研究テーマは，研究代表者がこの科研費プロジェクトが始まる前年の 2013 年に発表した引用論文①を起源とする研究課題である．この論文で問題の重要性を明らかにし，問題を教師付き学習問題として定式化している．その後，本科研費プロジェクト期間中に，国内外の研究者と協力して研究を進め，教師付き学習の精度を高めることに成功している．

（２）「遺伝子発現データからの遺伝子ネットワークの推定」（発表論文 ⑤）に関して次の結果を得ている．
Gaussian グラフィカル・モデリングは，遺伝子ネットワークの推定においてよく使われるモデルの１つである．遺伝子ネットワーク推定問題の特徴の１つは，与えられるサンプル数が少ないことである．この仮定に加えて，遺伝子ネットワークはスケール・フリーであると仮定した．つまり，解くべき問題は，少ないサンプルからモデル・パラメータである共分散の構造がスケール・フリーである Gaussian 分布を推定する問題となる．これに対して，共分散の構造のスケール・フリー性を捉える正則化項を定式化し，対応するサンプリング最適化アルゴリズムを開発した．その性能評価では，既存手法よりも高い予測精度を得ることができた（図４参照）．

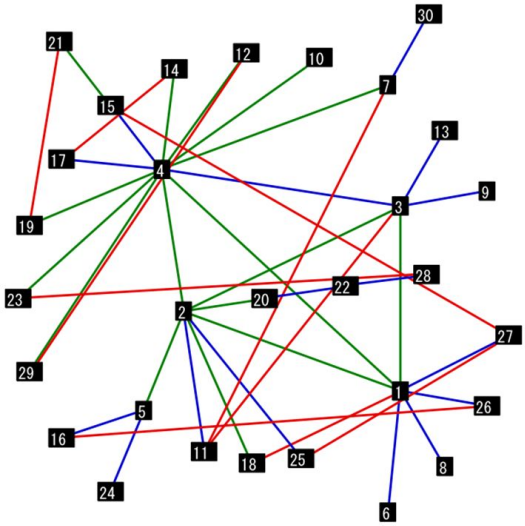


図４ 推定された遺伝子ネットワークの例

（３）個別問題の各手法の背後に共通に存在する本枠組みの原理を明確化・体系化し再利用可能な形にまとめることに関しては，実際に定式化した正則化項の数が一桁台にとどまっているので抽象化して考えるレベルに達しなかった．この点は今後の課題である．

＜引用文献＞

① Osamu Maruyama, Heterodimeric protein complex identification by naïve Bayes classifiers, BMC Bioinformatics, 2013, 14:347, 10.1186/1471-2105-14-347


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）

① Osamu Maruyama, Yuki Kuwahara, RocSampler: Regularizing overlapping protein complexes in protein-protein interaction networks, in the Proceedings of 2016 IEEE 6th International Conference on Computational Advances in Bio and Medical Sciences (ICCABS), 査読有り, 2016, 10.1109/ICCABS.2016.7802774

② Osamu Maruyama, Limsoon Wong, Regularizing predicted complexes by mutually exclusive protein-protein interactions, in the Proceedings of 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), 査読有り, 2015, pp. 1068 – 1075, 10.1145/2808797.2808817

③ Chern Han Yong, Osamu Maruyama, Limsoon Wong, Discovery of small protein complexes from PPI networks with size-specific supervised weighting, BMC systems biology, Vol. 8, Suppl. 5, 2014, S3, 10.1186/1752-0509-8-S5-S3

④ So Kobiki, Osamu Maruyama, ReSAPP: Predicting overlapping protein complexes by merging multiple-sampled partitions of proteins, Journal of bioinformatics and computational biology, Vol. 12, Issue 6, 2014, 1442004, https://doi.org/10.1142/S0219720014420049

⑤ Osamu Maruyama, Shota Shikita, A scale-free structure prior for Bayesian inference of Gaussian graphical models, in the Proceedings of 2014 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2014, 131 – 138, 10.1109/BIBM.2014.6999141

⑥ Peiying Ruan, Morihiro Hayashida, Osamu Maruyama, Tatsuya Akutsu, Prediction of heterotrimeric protein complexes by two-phase learning using neighboring kernels, BMC bioinformatics, Vol. 15, Suppl. 2, 2014, S6, 10.1186/1471-2105-15-S2-S6

⑦ Yasuhiro Okamoto, Kensuke Koyanagi, Takayoshi Shoudai, Osamu Maruyama, Discovery of Tree Structured Patterns Using Markov Chain Monte Carlo Method, Proc. 7th IADIS International Conference on Information Systems 2014, 2014

〔学会発表〕（計９件）
① Osamu Maruyama, Protein complex prediction, Forum "Math-for-Industry" 2016 - Agriculture as a metaphor for creativity in all human endeavors, 2016/11/23, **ブリスベン（オーストラリア）**

② Osamu Maruyama, RocSampler: Regularizing overlapping protein complexes in protein-protein interaction networks, 2016 IEEE 6th International Conference on Computational Advances in Bio and medical Sciences (ICCABS), 2016/10/13, **アトランタ（米国）**

③ Osamu Maruyama, Regularizing predicted complexes by mutually exclusive protein-protein interactions, International Symposium on Network Enabled Health Informatics, Biomedicine and Bioinformatics (HI-BI-BI 2015), 2015/8/26, **パリ（フランス）**

④ Osamu Maruyama, Protein Complex Prediction, Kick-off Meeting of IMI Australia Branch in La Trobe--Mathematics Bridge over the Pacific for Competitive Edge in Industry, 2015/3/13, **メルボルン（オーストラリア）**

⑤ So Kobiki, ReSAPP: Predicting overlapping protein complexes by merging multiple-sampled partitions of proteins, GIW ISCB-ASIA 2014, 2014/12/16, **東京国際交流館（東京都・）江東区青海）**

⑥ Osamu Maruyama, A scale-free structure prior for Bayesian inference of Gaussian graphical models, IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2014/11/4, **ベルファスト（イギリス）**

〔図書〕（計１件）
① R. Nishii, S.-i. Ei, M. Koiso, H. Ochiai, K. Okada, S. Saito, T. Shirai (Eds.), Springer Japan, A Mathematical Approach to Research Problems of Science and Technology, 2014, 507, **分担：** Osamu Maruyama, Markov Chain Monte Carlo Algorithms, 349 - 363, https://link.springer.com/chapter/10.1007/978-4-431-55060-0_26#page-1

〔産業財産権〕

○出願状況（計０件）

○取得状況（計０件）

〔その他〕
ホームページ等
http://imi.kyushu-u.ac.jp/~om/

**６．研究組織**
**(1)研究代表者**
**丸山　修（Maruyama，Osamu）**
**九州大学・マス・フォア・インダストリ研究所・准教授**
**研究者番号：20282519**

**(2)研究分担者**

**(3)連携研究者**

**(4)研究協力者**