若手研究者インターナショナル・トレーニング・プログラム(ITP)

バイオインフォマティクスとシステムズバイオロジーの国際連携教育研究プログラム 応募書類

Name:　烏山 昌幸

Title: 機械学習によるグラフ上でのデータ解析手法に関する研究

Institute: 京都大学化学研究所バイオインフォマティクスセンター

Partner institute of your choice：Centre for Computational Biology, Mines ParisTech

Duration of your choice: 2013 年 10 月 2 日 – 2013 年 12 月 26 日

Plan：

要旨

タンパク質相互作用ネットワーク等，生物学におけるグラフ型データの解析に対する需要は高い．しかし，古典的な統計学の枠組みではこのような構造を持つデータや，さらにそこに数値データも付随するような事例を扱うのは難しい．本応募はこのような問題に対する機械学習アプローチに関する研究のために, 生物学データの機械学習による解析における第一人者であるフランス Mines ParisTech の Jean-Philippe Vert 教授の研究室に滞在することを希望するものである．はじめに研究の背景となる事項について紹介し，その後，具体的な滞在の目的と計画について述べる．

背景

解析技術の発達に伴って，DNA や遺伝子など生体上の重要な情報が高速・大量に取得できるようになり，膨大な情報から有用な知見を如何に引き出すかが関連学術分野のみならず社会全体にとっても大きな関心事となっている．ただし，問題となるのはデータの絶対量だけではない. 生物学分野においては古典的な統計学で扱われるような数値データのみならず, より複雑な構造を持つデータを扱わなければならない. ここでは**構造を表現するモデリング手法として「グラフ」を考える**.

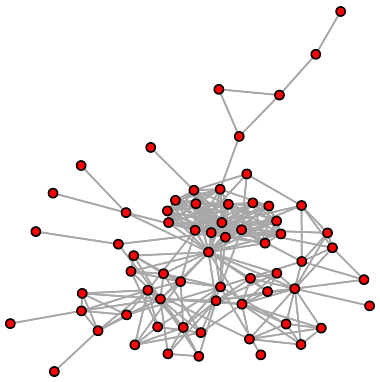

図 1: タンパク質相互作用ネットワーク. 各頂点がタンパク質であり，相互作用は頂点間の辺で表現される.

グラフは頂点集合とそれらを結ぶ辺から成り, 様々なデータを表現することができる. 例えばタンパク質の相互作用や代謝の経路, あるいは遺伝子制御等のネットワークデータはグラフとして表現できる構造データの例である. タンパク質相互作用の場合なら，各タンパク質が頂点，相互作用の有無がグラフの辺で表現される(図 1)．このようなグラフに対して従来型の**数値データも同時に付随する**ことでさらに複雑な状況が生まれている(図 2). 再びタンパク質ネットワークの場合を考えると，各タンパク質に対応する遺伝子発現量の数値を併せて解析することで生物機能の発現とネットワーク構造の関係性を解析する．例えばある特定の状況(癌細胞など) において相互作用ネットワーク上のどの部分が活性しているかを見つけるのは，生体内のメカニズムの理解，バイオマーカーの発見，ひいては薬剤の開発などにとって非常に重要となる．単純な数値を扱う古典的な統計学や，数学的グラフを扱うグラフ理論で

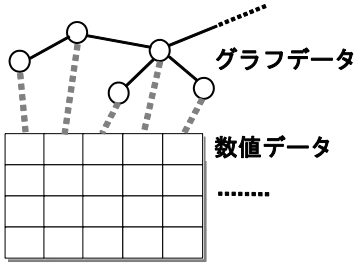

図 2: データの模式図. グラフの各頂点に数値データが対応する.

Plan (Continued)

はこのような問題に対応することは難しく，**多様化するデータタイプに対応する新たな枠組みが必要とされている．**

<u>滞在目的・計画</u>

今回希望する滞在ではこのような**グラフ上での数値データの隠れた構造を発見する「機械学習」アルゴリズムの研究**を計画している．機械学習は情報科学の一分野であり，大量データからの知識獲得のための方法論・アルゴリズムの理論や実装を考える分野である．この分野はコンピュータ技術の進歩とともに急速に発展し，現在では証券市場の解析からゲームソフトまで幅広く応用されており，生物学データの解析においてもその有用性はすでに広く認識されている（例えば[1]）．

本応募ではフランス Mines ParisTech の Jean-Philippe Vert 教授のグループ(http://cbio.ensmp.fr/~jvert/)への滞在を希望する．Vert 教授は機械学習アプローチによる生物学データの解析について引用数の多い論文を多数発表しており，理論と実践の両側面に関する貢献がよく知られている．機械学習の方法論に関しては LASSO と呼ばれる予測モデルから重要な変数を発見する手法に関する研究[3]などは応募者の今回の研究テーマあるいは過去の研究とも関連が強く，教授との議論は行いやすい関係にある．

応募者は現在，グラフと数値の複合型データに対してグラフィカルガウシアンモデルと呼ばれるモデルに基づくアプローチについて検討している．このモデルではグラフをガウス分布における依存関係として捉えることでグラフ情報と数値情報を同時に扱うことができる．推定されたガウス分布から行列分解などの手法を用いて情報を取り出すことで，グラフのどの部分構造が活発な相互作用を持つかを見出すアルゴリズムの構築を行う．そこで，Vert 教授らのグループの研究と関連の深い LASSO 型の変数選択手法と応募者のアプローチを組み合わせることで抽出した情報がさらにどういった状況で重要か，例えば特定の癌と関連している等，を発見するアルゴリズムが設計できる．

このアイデアに基づいた方法論についてすでに予備的な結果を得ているため[3]，到着後なるべく早く現地でプレゼンテーションを行い，内容について議論を始める．推定手法に関して洗練化を行いつつ，実際のタンパク質データ等を使った計算機実験を進めていく．観察された結果について順次議論を行い，必要に応じて修正していくプロセスを繰り返す．期間中にある程度実験を収束させ，論文を作成し投稿まで進めることを最終目標とする．すでにプロトタイプがあるため，このスケジュールの達成可能性は十分に高いと考えている．この論文をベースに帰国後もアルゴリズムの拡張を行い，背景で述べた問題に対する一つの枠組みの完成を目指す．これは機械学習によるグラフデータ解析に関する新たな枠組みであり，この**滞在が果たし得る役割は大きい**．また，今回を機に **Vert 教授との今後の研究協力体制を確立することも重要な目的の一つ**である．

[1] E. Barillot, L. Calzone, P. Hupe, J-P. Vert, and A. Zinovyev, Computational Systems Biology of Cancer, CRC Press, 2012.

[2] L. Jacob, G. Obozinski and J.-P. Vert, Group Lasso with Overlaps and Graph Lasso, *Proceedings of the 26th Annual International Conference on Machine Learning*, 433-440, 2009.

[3] M. Karasuyama and H. Mamitsuka, Latent Factor based Subnetwork Feature Selection, *Machine Learning and Applications to Biology*, 2013.