分子設計情報

教　授：馬見塚 拓　助　教：Canh Hao Nguyen

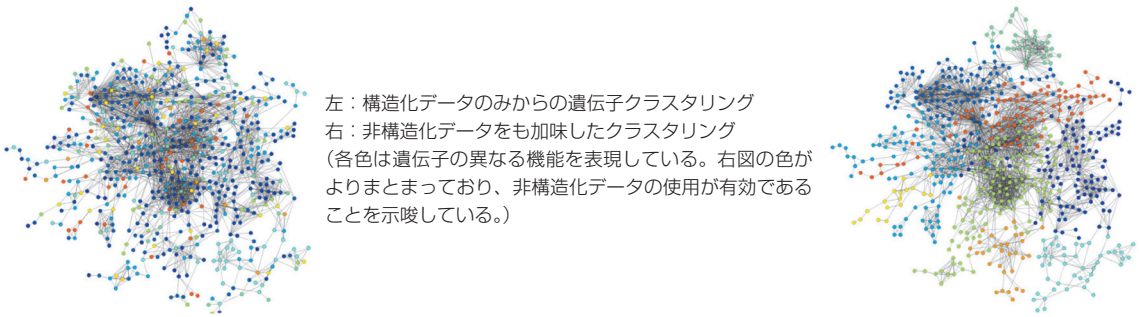
研究概要

生命科学では、近年の実験技術の進歩やビッグサイエンスの潮流により様々な種類のデータが大量に生成され、それらをデータベース化し共有する体制が世界規模で整ってきました。一方、これらのデータが生命現象の解明に十二分に利用されているとは言い難い状況にあります。特に、蓄積したデータから情報処理技術によりデータを解析する「バイオインフォマティクス」が必要です。中でも、データに隠された、内在する有益な情報を計算機により自動的に獲得する技術がひときわ重要でしょう。このような技術の研究分野を計算機科学では「機械学習（machine learning）」あるいは「データマイニング（data mining）」と呼んでいます。機械学習とは計算機がデータの特徴（すなわち、データに内在する規則、パターン、仮説等）を自動的に学習することを指し、データマイニングとは鉱山から貴重な宝石を掘るという mining（採掘する）という言葉になぞらえて、データの山から有益な情報を得ることを指します。いずれも統計科学と密接に関係します。さて、従来、これらの分野で扱うデータは、構造化データと呼ばれるいわゆる表（各事例を行、事例の各属性を列）データで、これに対する解析技術はあまたと提案されてきました。一方、生命科学で近年蓄積されるデータは多様で必ずしも表データではありません。例えば、ゲノム配列、化合物の化学構造式、信号伝達経路等、表で与えられないものが数多くあります。このような非構造化データを表に変換しようとすれば、生命科学にとって重要な情報が欠落する可能性があります。そこで、非構造化データをそのまま扱う機械学習およびデータマイニング技術の構築が非常に重要です。このようなアプローチは生命現象の解明に有益であるだけでなく、計算機科学においても新しい貢献となる研究課題です。当分野では、上記のように、機械学習・データマイニングを中心とした計算機科学（および統計科学）技術の新展開による生命科学および創薬科学の発展への貢献を目指し研究遂行中です。以下、具体的な研究課題の中から 3 つほどを取り上げ簡単に概説します。

1）構造化データと非構造化データの統合データマイニング：近年蓄積された遺伝子をはじめとする生体分子相互の性質はグラフで表現されることがままあります。例として、遺伝子相互作用ネットワーク、タンパク質相互作用ネットワーク、代謝パスウェイなどが挙げられます。一般的な言い方をすれば、これらは事例間の関係性をグラフで表現しています。このような非構造化データ（グラフ）と構造化データを組み合わせ、両データの性質を反映して事例をクラスタリングする（同じ性質毎にまとめる）手法を開発しています。具体的には、遺伝子間の関係性を表現したグラフ（非構造化データ）と発現による遺伝子の類似性を捉えることが可能な cDNA マイクロアレイ（構造化データ）により遺伝子のクラスタリングを行い、遺伝子機能等をより正確に予測する手法です。一例を下図に例示します。現在、グラフのモジュール性の高い場合に有効な手法を開発していますが、今後はグラフの様々な性質を考慮した手法を開発することにより、生体分子の様々な関係性を示す各グラフに適した、生体分子のクラスタリングが可能になるでしょう。

2）木構造データからの学習：非構造化データはグラフばかりではなく、糖鎖の二次元表現など木もあります。木に対する新しい効率的な機械学習手法を考案し、実適用から糖鎖の各クラスのパターン発見と複数糖鎖のアライメントを実現し、今後は自動分類を目指しています。

3）生命科学文献データからの学習：近年大規模に蓄積されている非構造化データの一つには、医学論文等の文献データも挙げられます。これら文献データから有効な知識を効率的に獲得する手法を開発しています。一例は、大規模な文献データの中で、与えられた文章（例えば、「狂牛病の遺伝子の機能は何か？」）に最も関係する文献を探索する情報検索と呼ばれる分野の手法です。他にも、同一文献内に同時に出現する生体分子の共起データから未知の関係性を発見する手法を構築しています。実際、この手法は特定の癌に関係する未知の低分子化合物や遺伝子を高い確度で提示することが示してきました。さらに、大量の文献をその内容により自動的にクラスタリングすることも文献データ処理の上で非常に重要であり、実際頑健で効率的手法を構築してきました。



左：構造化データのみからの遺伝子クラスタリング
右：非構造化データをも加味したクラスタリング
（各色は遺伝子の異なる機能を表現している。右図の色がよりまとまっており、非構造化データの使用が有効であることを示唆している。）

主要論文

- Takigawa *et al.* Mining Significant Substructure Pairs for Interpreting Polypharmacology in Drug-Target Network. *PLoS One*, **6**(2), e16999, 2011.
- Takigawa and Mamitsuka. Graph Mining: Procedure, Application to Drug Discovery and Recent Advance. *Drug Discovery Today*, **18**(1-2), 50-57, 2013.
- Ding *et al.* Similarity-based Machine Learning Methods for Predicting Drug-target Interactions: A Brief Review. *Briefings in Bioinformatics*, **15**(5), 737-747, 2014.