# 遺伝子データベースとテキストマイニングを用いたタンパク質間の関係性の可視化

**1815070** 武藤克弥　　情報基盤工学講座　指導教員 奥原浩之

## 要約

生命科学分野において、テキストマイニングを用いて，データベースに日々蓄積されていく遺伝子データから遺伝子・タンパク質間の関係性や相互作用を見出すことの重要性は依然として強い．本研究では複数の完全ゲノム中に含まれるタンパク質を抽出し，共通のタンパク質の共起関係を抽出し，それらの関係性を**3D**グラフに可視化する．そして，出力された関係から新たな考察をユーザに促すことを目的とする．

キーワード：
遺伝子データベース, スクレイピング, テキストマイニング, 共起ネットワーク, 可視化

## 1 はじめに

計算機の発展に伴い，年々増加しつつある大量かつ整理されていない文書データに対して，自然言語処理や情報検索技術等を用い，有用な情報を見い出すテキストマイニングが近年盛んになってきている．あるキーワードで検索し，検索結果に現れる文章から自然言語処理などを行って必要な単語のみを抽出したり，得られた情報をデータベース化し，新たな情報提供を行えるソフトウェアを開発したりなど，様々な応用がある．生体・遺伝子情報を扱う生命科学分野においても，日々蓄積されていく文献・遺伝子データベースの中から特定の働きをもつ遺伝子名を検索したり，遺伝子間の関係性や相互作用を見い出したりするのにテキストマイニング技術が用いられている．本研究では，遺伝子間の関係性を可視化することに重点を置いている．まず，**GenBank**から収集したタンパク質リストを用いて，完全ゲノム中のアミノ酸配列からタンパク質を抽出する．次にタンパク質**1**つ**1**つの共起数を計算する．そして複数の完全ゲノムにおいて，あるタンパク質どうしの共起数を合計した隣接行列を作成し，関係性をグラフ上に可視化する．

## 2 テキストマイニングと可視化

### 2.1 テキストマイニングとスクレイピング

情報抽出するための大量の文書データを収集するのにスクレイピングが必要となってくる．スクレイピングとは**Web**上のサイトやニュース記事などから文章を取得する手法であり，効率的にデータ得ることができる．これにテキストマイニングと組み合わせることで自動的に情報抽出する仕組みを作ることができる．

### 2.2 3Dグラフによる可視化

テキストマイニングによって可視化を行う際，頂点(ノード)と辺(エッジ)を用いたグラフを用いることが多い．たいていノードには意味を持った単語が入り，特定の関係性のある単語同士がエッジで結ばれる構造となる．ノードからノードへの流れがある場合，エッジに矢印が付く有効グラフとなり，ない場合は無向グラフと呼ばれる．単語どうしの結びつきとその度合いについての情報は，行と列に全く同じ並びで単語を入れた隣接行列で表される．行列の要素は行の単語ノードから列の単語ノードに向かうエッジの重み(関連度)を表しており，ある**Web**ページの文章中の単語が，他の単語にどのように共起されるかという共起性によって計算されることが多い．

**3D**グラフ**[1]**ではある検索ワードに対して，**Twitter**や**Web**サイトよりスクレイピングしてきた文章から重要となる単語を抽出し，共起頻度を計算したのち隣接行列を作成する．得られた隣接行列の表を**Python**のプログラムによって辞書形式の**Json**ファイルに変換し，**3D**グラフが動作するサーバ上に送ることで可視化を実現する．関連性のある単語同士はエッジで結ばれ，関連度の高さに応じてエッジの太さが変化する．ユーザは検索ワードから複数の関連する単語を得ることができ，アイデアの発想支援につなげることができる．

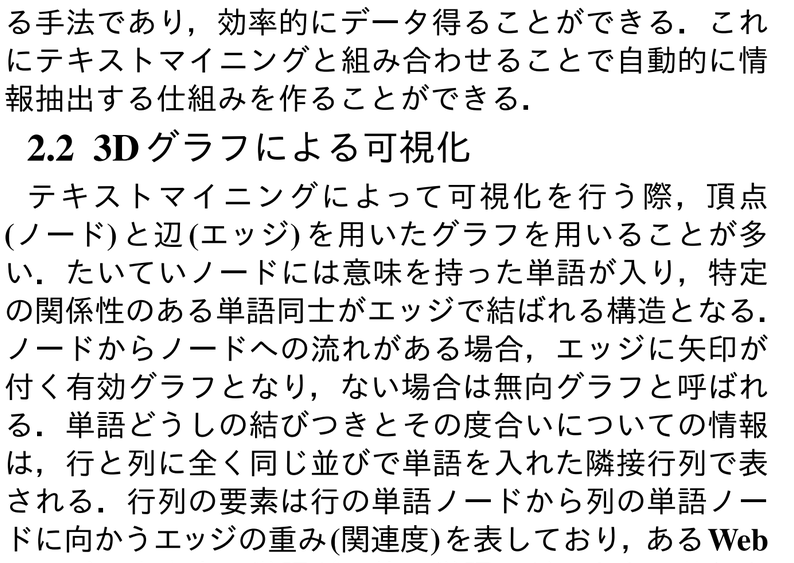


図1　検索ワード「学生」でスクレイピングした場合の3Dグラフ

## 3 生命科学とテキストマイニング

### 3.1 遺伝子モデルの可視化

生命科学分野におけるテキストマイニングとは，文献や遺伝子データベースから表記名・派生語が複数ある遺伝子名等を正確に識別したり，タンパク質間の相互作用の関係性やネットワーク状の遺伝子構造を抽出したりなどが挙げられる．とりわけ，関係性・構造抽出においては，視覚的な見やすさからグラフを用いて可視化されることが多い．遺伝子データベースの１つに**PubMed**がある．**PubMed**とは生物医学雑誌の論文や妙録を電子化したものであり，今までに発表されてきた膨大な数の文献を閲覧することができる．**PubMed**には**MeSH**と呼ばれる医学用語を階層構造に分類したシソーラスがあり，**MeSH**内の用語の共起性を用いて用語間の関係性を調べる研究が多くなされてきた．

**[2]**の研究では**MeSH**から用語を抽出し，ビブリオメトリックの指標として，**Dice**係数の逆数を用いて**2**つの遺伝子間の非類似行列を求め，それをエッジ重みとして可視化した．**[3]**の研究ではサイトカインと疾患の関係性を**MeSH**用語の共起頻度を用い，相関行列を導出してクラスタリングを行った．

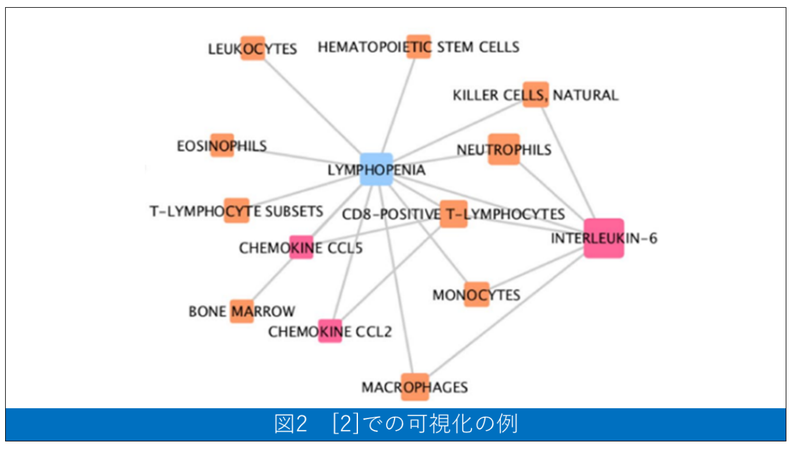


図2　[2]での可視化の例

### 3.2 タンパク質間の共起性

遺伝子やタンパク質の関係性を調べる際に共起分析が良く用いられる．共起分析はテキストマイニングの研究ではよく用いられており，文章中のある単語に付随して出現する単語がある場合，その**2**つの単語には何らかの関連があるものとして分析する手法である．これは，**Web**ブラウザの検索アルゴリズムに等に利用されている．生命科学分野においては**PubMed**などの論文の妙録を集めた**DB**を用い，論文中に出現する複数の遺伝子名やタンパク質名を抽出し，その共起性を分析する研究が数多くなされている．一方で，論文ベースではなく，ゲノム配列内に出現する特定の遺伝子やタンパク質どうしの研究も行われている．

**[4]**では，世界的に**MRG**(金属耐性遺伝子)が抗生物質耐性遺伝子**(ARG)**の増加に影響している傾向がみられることから，可動遺伝因子**(MGEs)**の**DNA**に含まれる**ARG**と**MRG**の共起性を解析している．ここでは**[4]**で用いられている共起性評価指標を説明する．評価指標として遭遇率と平均最小距離の２つがある，遭遇率は**1**つのゲノム配列に対して，**MRG**から距離$5000bp$内にある**ARG**の数を$[200bp, 100000bp]$**(**ステップ数**:**$200bp$**)**の範囲でカウントされる．**MRG,ARG**に属する遺伝子をそれぞれ$M_i$**,**$A_i$とし，この範囲における$M_i$の数を$N_{(M_i)}$，**1**つの**MRG**遺伝子の前後**5000bp**内に含まれる**ARG**の数を$N_{(A_i)}$，とすると，**K**個のゲノムにおける$M_i$と$A_i$の遭遇率$IoE_i$**(1)**が得られる

$$IoE_i = \frac{\sum_{j=1}^{K}(\frac{N_{(A_i)}}{N_{(M_i)}})}{K} \times 100 \qquad (1)$$

平均最小距離$AMD_i$は$M_i$の前後**5000bp**内に複数存在する$A_i$のうち，最も近くにある$A_i$と$M_i$の距離$d_i$における**1**ゲノム中の平均距離であり，**(2)**で与えられる．

$$AMD_i = \frac{\sum_{j=1}^{N_{(M_j)}} min(d_j)}{N_{(M_j)}} \qquad (2)$$

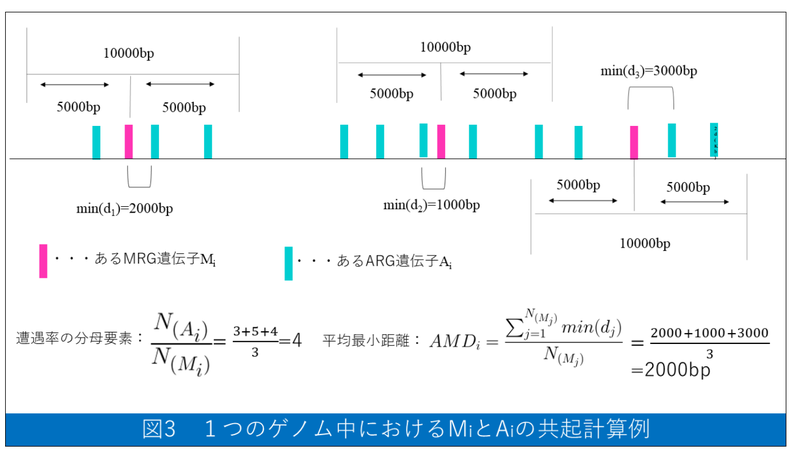


図3　１つのゲノム中におけるMiとAiの共起計算例

## 4 提案手法

本研究では，**NCBI**の完全ゲノムデータベースから完全ゲノムを種別を問わずダウンロードし，各完全ゲノム内の遺伝子**1**つ**1**つのアミノ酸配列を抽出する．同様に**Uniprot**から人に含まれるとされるタンパク質のリストを用意する．得られたアミノ酸配列から該当するタンパク質を割り当てていき，**[4]**の共起評価指標を用いて，完全ゲノム内で抽出されたある２つのタンパク質どうしの共起性を計算する．この計算を全ての組み合わせペアで行い，それを重みとして隣接行列に表現する．そして，共起頻度が高い上位数十件(任意)のみに絞って**3D**グラフに可視化する．最終的には得られたタンパク質の関係性が意味のあるものかどうかを評価し，有効なものであった場合に，その関係についてさらなる調査をユーザに促すような支援を目的とする．

## 5 数値実験並びに考察

## 6 おわりに

## 参考文献

**[1]** 平松 楓也, ”発想支援とジオプロセンシングのシームレスな統合に向けた**QGIS**プラグインの開発”, 富山県立大学学位論文 **2020.**

**[2] Stapley BJ, Benoit G, ”Biobibliometrics: information retrieval and visualization from co-occurrences of gene names in Medline abstracts”, Pacific Symposium on Biocomputing 2000.**

**[3] Nophar Geifman, Anthony D. Whetton, ”A consideration of publication-derived immune-related associations in Coronavirus and related lung damaging diseases”, Journal of Translational Medicine 2020.**

**[4] Li-Guan Li, Yu Xia, Tong Zhang, ”Co-occurrence of antibiotic and metal resistance genes revealed in complete genome collection”, The ISME Journal 2016.**