遺伝子発現に関する文献における遺伝子間のネットワーク構築

澁谷 翔吾

1　はじめに

近年，細胞内の遺伝子発現量を測定するために，DNA マイクロアレイ (Deoxyribo Nucleic Acid Microarray)[1] が用いられている．DNA マイクロアレイでは，膨大な数の遺伝子発現変化を網羅的に解析することが可能である．DNA マイクロアレイを用いることで，正常組織と癌組織におけるそれぞれの遺伝子発現の差異が検出でき，癌組織の遺伝子発現による抗癌剤や放射線治療に対する感受性予測が可能になる．このことから，多くの研究者が DNA マイクロアレイを用いて遺伝子発現の研究を行っている．しかし，全ての遺伝子を網羅的見ることでデータが膨大になり，現在のところ，得られたデータの解析が非常に難しく，全ての遺伝子に対するその発現の規則性，法則は導き出されていない [2]．また，情報量が多くなればなるほど，その中から重要な法則性を導き出すことが困難である．これらの理由から，研究者は遺伝子発現について記述された論文，および既知の遺伝子発現に関する知識ベース (Gene Set Enrichment Analysis : GSEA)[*1] を参考にしながら研究を行っている．

本研究では，DNA マイクロアレイの実験結果に関して記述された文献から自然言語処理を行い，遺伝子発現の情報を体系化することを目標としている．遺伝子発現の情報を体系化し，遺伝子同士の関連性を視覚化する．遺伝子発現情報を体系化する上で，本研究では医療の専門用語を関連ある用語で定義した知識ベース（以下，医療概念ベース）を構築し用いる．医療概念ベースを用いることで，文章の表面的な単語だけで文章解析するのではなく，その単語はどういった意味なのかといったことを考慮して解析することが可能になる．本発表では，医療概念ベース構築の前段階として，医療の専門用語ではなく，一般的な用語を関連ある用語で定義した知識ベース（以下，概念ベース）構築について解説し，今後の研究の方向性について解説する．

2　概念ベース

概念ベースとは，ある単語の意味（概念）をその単語に関連のある単語群（属性）で定義した知識ベースである [3]．例えば，人間は，[学校] から「教師」や「生徒」などの単語を連想できる．この場合，概念ベースでは，[学校] の概念を {教師，生徒，…} のように概念の属性群として保持している．概念ベースを構築することで，Web 上の電子文書などを解析する際，表面的な単語だけに留まらず，その単語の意味を考慮した解析が可能になる．

2.1　概念ベースの構築

本節では，医療概念ベース構築の前段階として，一般的な単語を定義した概念ベースを構築する．その流れを以下に解説する．

- 概念となる語の準備
  一般に，概念には国語辞書の見出し語を用いる．今回は，類語玉手箱[*2] という辞書を用いた．または，どのような概念ベースを構築するかによって，その専門記事から名詞を抽出し，それらを概念とすることも可能である．今後，医療概念ベースを構築するにあたっては，医療に関する論文などから名詞を抽出し，概念として定義することを検討している．

- 属性候補の決定
  概念を定義する語，すなわち属性を決定する．本研究では，概念の決定には Wikipedia[*3] を利用している．つまり，概念を Wikipedia で検索したとき，得られる説明文から属性となりうる語を抽出する．本研究では，説明文内に存在する名詞を概念の属性とする (Fig. 1)．

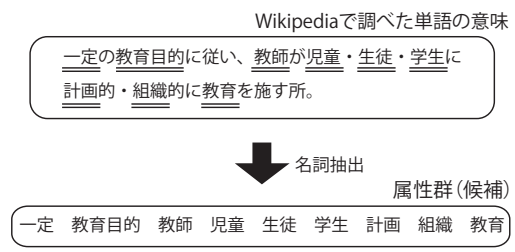

Fig.1　属性の決定 (出典:自作)

例えば，[学校] という単語を Wikipedia で調べたとき，その説明文が「一定の教育目的に従い、教師が児童・生徒・学生に計画的・組織的に教育を施す所。」であるとする．この場合，この説明文から名詞である「一定」，「教育目的」，「教師」，「児童」，「生徒」，「学生」，「計画」，「組織」，および「教育」を抽出し，これを概念の属性候補とする．後に解説する属性への重み付けにおいて，ある一定の数値以上の単語を属性とする．

- 属性の重み付け
  属性がどの程度概念と関連があるかを定義するために，属性には重みを付加する．本研究では，概念と

[*1] http://www.broad.mit.edu/gsea/

[*2] http://www.dictjuggler.net/tamatebako/

[*3] http://www.wikipedia.org/

属性の関連度を算出することで，属性の重み付けを行う．関連度計算には，2 単語の関連度を算出する階層距離計算 [4] を用いる．階層距離計算はシソーラスを用いて単語間の距離を求めることで 2 単語の関連度を求める．シソーラスは単語が階層的に分類されているため，階層の違いを距離に見立てて関連度を計算することができる． 2 単語の関連度（階層距離）は以下の式で計算できる．

$$sim(q,d) = \frac{2c_{qd}}{(d_q+1)+(d_d+1)} \qquad (1)$$

ここで，d_q と d_d は見出し語 q，d が属する意味属性の深さである．また，c_{qd} は見出し語 q が属する意味属性と見出し語 d が属する意味属性の共通の上位属性の中で最も下の階層に位置するカテゴリの深さである． 例えば，概念 [学校] と属性「教師」との階層距離を計算する場合，該当する単語が属する段数と 2 単語に共通する属性体系の段数を求める必要がある．Fig. 2 を用いて解説する．

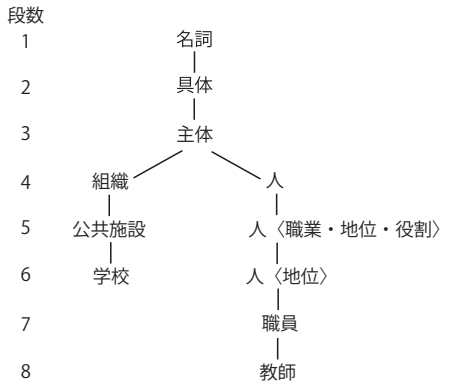

Fig.2　階層距離計算の例 (出典:自作)

概念 [学校] は，上位ノードから辿ると，[名詞，具体，主体，組織，公共機関，学校] となり，第 6 段に属していることが分かる．同様に，「教師」は [名詞，具体，主体，人，人〈職業・地位・役割〉，人〈地位〉，職員，教師] となり，第 8 段に属している．また，2 単語に共通する属性体系の段数は [名詞，具体，主体] となり，第 3 段になる．これらの数値を式 (1) に代入すると，学校と教師の階層距離が算出できる．本例では，階層距離は 0.375 である．0 に近づくほど関連は弱く，1 に近づくほど関連は強い．本研究では，0.5 以上のものを属性としている．以上の操作を概念を定義する語全てに対して行うことで，一つの概念をその属性と重みで定義している．

2.2　概念の例

上述した概念ベース構築により得られた概念の例を Table 1 に示す．属性はその重みの順に上位 3 つまでとする．

概念 [本] は属性として「証券」,「文化」, および「作品」を持っており，それぞれの属性の重みは 0.63，0.57，0.53 である．概念ベースを利用することで，例えば，文章中で [消しゴム] が出てきたとき，その単語に加えて，「鉛筆」,「文房具」, および「文具」といった属性を考慮して解析を行うことができる．

Table1　概念例

概念	属性
本	証券 (0.63), 文化 (0.57), 作品 (0.53)
冷蔵庫	家電 (0.8), 器具 (0.56), 食材 (0.53)
ぬいぐるみ	玩具 (0.80), 包み (0.53)
家	民間 (0.57), 事務所 ((0.57), 世界 ((0.57)
消しゴム	鉛筆 (0.70), 文房具 (0.63), 文具 (0.63)

3　DNA マイクロアレイの遺伝子発現情報の体系化

3.1　DNA マイクロアレイ

DNA マイクロアレイとは，スライドガラスなどの基板上に DNA 断片を固定化した上で，相補的な DNA 鎖同士で塩基対を形成する原理を利用し，遺伝子を検出するセンサデバイスである (Fig. 3).

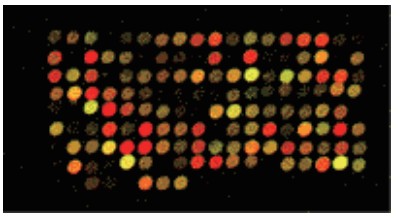

Fig.3　DNA マイクロアレイ (参考文献 [5] より引用)

DNA マイクロアレイを用いることで，膨大な数の遺伝子発現変化を網羅的に解析することが可能である．

DNA マイクロアレイを用いた実験の原理は，DNA の塩基であるアデニン (A)，グアニン (G)，シトシン (C)，およびチミン (T) がアデニンとチミン，グアニンとシトシンという組み合わせで結合する特性を利用する．この原理を利用して，DNA マイクロアレイに固定した DNA 断片と特定の細胞の組織から調整した DNA とを結合させ，目的の細胞，組織でどの遺伝子が作用していたのかを調べる．実験では，DNA マイクロアレイのどのスポットに DNA 断片が結合しているのか試薬を作用させたものとそうでないものの 2 枚を比較する．例えば，3 つの遺伝子に違いが見られるとき，試薬が遺伝子 3 つの働きに何らかの影響を及ぼしているのではないかと予測する．さらに，もしその中で遺伝子 1 つがある特定の病気に関する遺伝子だとすると，その試薬はその病気に対して何らかの効果が期待できる．

DNA マイクロアレイを用いた遺伝子発現の実験では，現在のところ，得られた遺伝子発現データを標準化するには至っていない．また，異なるガラス板間の値をどう揃えるかなど得られた解析データを必ずしも有効に使う

ことができるとは限らない．そのため，研究者は遺伝子発現について記述された論文，および既知の遺伝子発現に関する知識ベースを参考に実験を行う．

3.2 遺伝子発現データ解析の現状

上述したように，DNA マイクロアレイの実験では，現在のところ，得られた遺伝子発現データを標準化するには至っていない．そこで，本研究では，得られた遺伝子発現データを体系化することを目標とする．体系化に用いるデータは，DNA マイクロアレイの実験について投稿された研究論文の序論を利用することを検討している．遺伝子発現データを体系化することで，遺伝子同士の関連性を視覚化する．

3.3 医療概念ベースの利用

得られた遺伝子発現データの体系化するために，研究論文の序論を解析することを検討している．解析対象はバイオインフォマティクスなど専門分野に関する論文であるため，上述した概念ベースではなく，専門用語を定義した概念ベースを利用する必要がある．専門用語を他の用語で定義することで，表面上の単語だけで序論を解析するのではなく，一つの専門用語が何を表しているのかまで考慮して序論を解析することが可能となる．

3.4 遺伝子発現データの体系化

DNA マイクロアレイを用いた遺伝子発現データを体系化する．体系化に用いるデータは研究論文の序論を考えており，その文章を解析することで，遺伝子発現における遺伝子同士の関連をネットワーク化することを考えている．体系化のイメージは Fig. 4 の通りである．

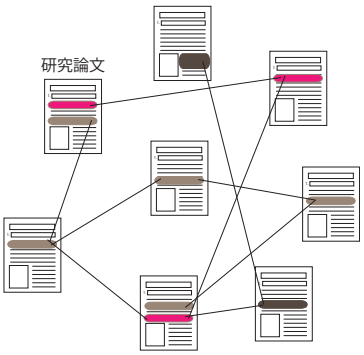

Fig.4　体系化のイメージ (出典:自作)

Fig. 4 内の線は関連を示しており，ある論文の序章に書かれてある遺伝子発現に関する内容は他の論文の序章に書かれてあるそれに関連があるということを示している．序論の内容と他の序論を内容から関連性を考慮し，体系化する．体系化した情報を視覚的に示すことで，遺伝子同士の関連を分かりやすく表示することを考えている．

4 まとめ

本研究の目標は，DNA マイクロアレイを用いて得られた遺伝子発現データを体系化し、遺伝子同士の関連を示すことである．体系化に用いるデータは，DNA マイクロアレイの実験について記述された研究論文の序論を利用することを検討している．論文を解析するにあたっては，医療概念ベースの利用を検討しており，本発表ではその前段階として概念ベースを構築した．今後は医療概念ベースの構築，遺伝子発現データの体系化を行う．

参考文献

1) 角田慎一，ゲノムワイド DNA アレイによる癌診断技術,

2) マイクロアレイ
http://cdna01.dna.affrc.go.jp/RMOS/background.html

3) 眞鍋康人, 小島一秀, 渡部広一, 河岡司，概念間の関連度やシソーラスを用いた概念ベースの自動精錬手法, 同志社大学理工学研究報告

4) 大橋敬久, シソーラスを用いた意思決定支援のための文書の抽出, 平成 18 年度卒業論, 2007 年

5) NICT 独立行政法人 情報通信研究機構
http://www.nict.go.jp/