

D-029

遺伝性疾患データベースを利用した 関連遺伝子検索システムの開発

潘 洪涛[†]
Hongtao Pan

宮崎 純[‡]
Jun Miyazaki

渡邊 日出海[†]
Hidemi Watanabe

植村 俊亮[†]
Shunsuke Uemura

1. はじめに

遺伝子とは遺伝情報の単位であり、遺伝形質を規定している。人間の遺伝子の総数は約4万個あると推定され、それぞれの研究グループによって複数のデータベースに蓄積されている。

本論文では、従来の人間遺伝子間の関連性の研究を分析し、生物学的に意味のある関連遺伝子を検索する手法を提案する。本手法はOMIM(Online Mendelian Inheritance in Man) データベース [2] を利用することにより、疾患を媒介して関連遺伝子の検索を可能にする。

本研究で提案するシステムでは、入力した検索遺伝子に対して、単に関連する遺伝子だけを列挙するのではなく、グラフの構造から各遺伝子の関連度の強さ順にランキングして出力することが可能となる。この手法は、以下の2つの点で従来の研究とは異なる。

- 文献中の無作為な単語の共起関係ではなく、専門家により解釈されたOMIMデータベースを利用するため、遺伝子間の関連性が明確である。
- 関連する遺伝子の関連度を、遺伝子と疾患からなるグラフ構造から容易にかつ高速に計算可能である。

本手法により、従来の検索手法と比較して、関連する遺伝子の検索精度が改善できる。

2. 従来の手法とその問題点

テキストデータで記録された生物学の知識を処理する多くのアルゴリズムは、文字列のパターンマッチや単語の出現頻度といった表面的な情報を解析するにとどまる。T.K.JenssenらはMedLineに記載された1000万以上の文献のタイトルとアブストラクトを解析し、タイトルとアブストラクト中に遺伝子名が共起した場合に、それらの遺伝子が潜在的に生物学的に関連性があると仮定している。それらの共起関係をリンクすることにより、人間の遺伝子間の関連データベースPubGene [1] を構築した。

PubGeneデータベース中では、各遺伝子をノードとして表し、関連する遺伝子間をリンクすることにより、遺伝子間の関連をグラフで表現している。更に、隣接するペアの遺伝子の関連の強さを評価するために、ペアの遺伝子が両方とも出現する文献の数を重みとしている。しかしながら、ある遺伝子とそれに関連する遺伝子はどのように関連があるのかについては表現できない。しかも、遺伝子間の関連の強さは生物学的な意味を考慮していない。即ち、この手法は遺伝子の表層的な情報を利用するにとどまり、そのため、ある遺伝子に対して、関連する遺伝子の検索精度が低かった。

一方、生物学的に意味づけられたデータベースとして、アメリカバイオテクノロジーセンター (NCBI) により提供されているデータベースOMIM(Online Mendelian Inheritance in Man) がある。OMIMは人間の遺伝性疾患に関連する遺伝子情報のデータベースであり、人間の遺伝性疾患を引き起こす遺伝子の機能、関連遺伝子の情報など最新の情報が提供されている。

しかしながら、OMIMデータベースのリンク情報を利用して関連する遺伝子を検索する場合、関連する遺伝子の情報は表示されるが、個々の遺伝子情報がテキストデータで記述されているため、遺伝子と遺伝子の間に直接関連が存在するかどうかについては、そのテキストデータを読まねばならず、関連する遺伝子の関連度の強さも判別できない。

3. 提案する手法

3.1 概要

本研究ではOMIMデータベースに記載されている人間の遺伝性疾患に関連する遺伝子情報を利用する。具体的には、OMIMデータベースのテキストデータを対象とし、遺伝性疾患を媒介して遺伝子間の関連性を求める。OMIMの1つのレコードの中に遺伝子と疾患が共起する場合には、それらの遺伝子と疾患が生物学的な関連性がある。これは、OMIMが専門家により遺伝子と疾患との関連を分類してデータベース化したものであるからである。

- 前処理: OMIMデータベースの「GENE FUNCTION」項目ごとに疾患と責任遺伝子を抽出する。疾患ごとに遺伝子をまとめ、疾患と遺伝子をリンクすることにより、遺伝子間の関連性をグラフで表現する。次に疾患と責任遺伝子からなるグラフから関連遺伝子の関連度の強さを求める。
- 検索処理: 入力した遺伝子に対して、関連する遺伝子だけを関連性の強さ順にランキングして出力する。

3.2 疾患と遺伝子の抽出

提案するデータベースを構築するには、OMIMデータベースのテキストデータから疾患と遺伝子を抽出し、グラフへ変換する。今回の実験では、OMIMデータベース中の全てのレコードの「GENE FUNCTION」項目のみに注目する。なぜならば、この「GENE FUNCTION」項目の中に遺伝子の変異および引き起こす疾患の情報が記述されているからである。

具体的には、OMIMデータベースの[GENE FUNCTION]項目中の共起するすべての疾患と遺伝子を抽出し、抽出された疾患と遺伝子は、それらをノードとして遺伝子から疾患に向けて、リンクを張ることによって有向グラフを構築する(図1の点線の右側参照)。

[†]奈良先端科学技術大学院大学 情報科学研究科

[‡]北海道大学大学院 情報科学研究科

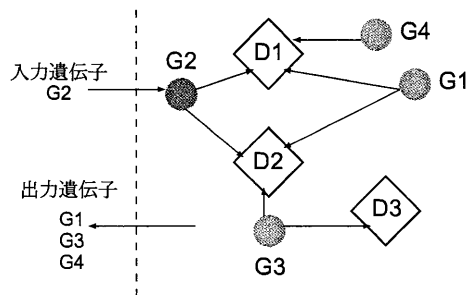


図 1: 遺伝子と疾患の関連グラフ

3.3 遺伝子間の関連度の強さの計算

入力した遺伝子と関連する遺伝子間の関連度の強さの計算方法はウェブ・リンクページの関連性を求める既存研究 HITS アルゴリズム [3] の Hub と Authority の概念を利用する。

本研究の場合では、Hub はある疾患に関連する遺伝子に、Authority はその疾患に相当する。Hub 値の高い遺伝子は Authority 値の高い疾患にリンクしており、Authority 値の高い疾患は Hub 値の高い遺伝子からリンクされている。従って、ある疾患と関連の強い遺伝子が上位にあって、遺伝子の変異によってよく引き起こされる疾患も上位にある。この HITS アルゴリズムの手法を利用することにより、より良い関連遺伝子のランキングが可能であると考えられる。

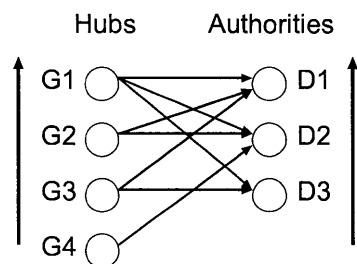


図 2: HITS アルゴリズムの原理

例えば、図 2 の左側で示すように、 G_2 と G_3 は共に 2 つのリンクを持つが、リンク先疾患のランクがより高い遺伝子 G_2 の方が高くなる。一方、右側の疾患の D_1 と D_2 は共に 3 つの遺伝子からリンクされているが、被リンク先遺伝子のランクがより高い疾患 D_1 のほうが高くなる。

3.4 関連遺伝子の検索

本研究の目的は、入力した遺伝子に関連する遺伝子をすべて検索し、入力した遺伝子との関連度の強さ順でランキングすることである。

図 1 のグラフでは、遺伝子 G_2 を検索遺伝子として入力し、入力した遺伝子 G_2 と関連する全ての遺伝子を検索し、この検索された遺伝子を入力した遺伝子 G_2 との関連度の強さに基づいてランキングして出力する。

結果として、入力した遺伝子 G_2 に関連する全ての遺伝子が疾患を媒介して検索される。

出力される関連遺伝子は以下のメリットがある。

- 関連遺伝子の間に生物学的に明確な意味がある
- 関連遺伝子の関連度の強さでランキング可能である

4. 実験結果と評価

提案する手法に基づくデータベースを構築し、検索プログラムを実装した。OMIM データベースの 2479 個の [GENE FUNCTION] テキストデータに基づいて、2614 個の疾患と 23299 個の遺伝子を出し、その中に、1817 個の遺伝子が関連遺伝子を持っていることが分かった。それらの疾患と遺伝子からなるグラフから関連遺伝子の関連度の強さを HITS アルゴリズムにより計算し、データベースに蓄積した。出力された関連遺伝子の数は最小 1 個、最大 1281 個あり、平均 297 個関連遺伝子があった。例えば、

- 遺伝子 BCR は 158 個の関連遺伝子を出力したが、OMIM 本体では 62 個の関連遺伝子を出力し、8 個の遺伝子のみが一致した。
- 遺伝子 SERCA1 は 2 個の関連遺伝子を出力したが、OMIM 本体で 3 個の関連遺伝子を出力し、遺伝子漏れが発生した。

これらは原因を以下と考えられる。

- 1 つの遺伝子に複数の名前が存在し、その処理を行っていない
- 遺伝子と疾患の抽出方法に問題がある

5. まとめと今後の課題

本研究では、遺伝性疾患と疾患責任遺伝子との対応関係について情報科学の視点から遺伝性疾患を通じて遺伝子間の関連性を求める方法を議論した。

今後の課題として、4 節で説明した原因を詳しく分析し、問題を解決することである。更に、関連遺伝子の関連度の計算方法に HITS アルゴリズムを導入したが、直接関連のない遺伝子同士の相互作用により疾患を引き起こすことは表現できない。このため、遺伝子間の本質的な生物学意味付けの関連度の計算方法を開発することも今後の課題である。

参考文献

- [1] PubGene, <http://www.pubgene.org>
- [2] OMIM, <http://www.ncbi.nlm.nih.gov/omim/>
- [3] Jon M. Kleinberg, 'Authoritative Sources in a Hyperlinked Environment', Journal of the ACM, Vol.46, No.5, pp604-632, 1999.