DEWS2004 I-8-03

# 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提案

潘　洪涛† 宮崎　純† 渡邊日出海†,†† 植村　俊亮†

† 奈良先端科学技術大学院大学 情報科学研究科　〒 630–0192 奈良県生駒市高山町 8916–5
†† 国立情報学研究所　〒 101–8430 東京都千代田区一ツ橋 2–1–2
E-mail: †{koto-h,miyazaki,hwatanab,uemura}@is.aist-nara.ac.jp, ††watanabe@nii.ac.jp

**あらまし**　従来指摘されてきた遺伝病や遺伝性疾患の殆どは遺伝子の変異や発現異常により起こることが明白となってきた。変異や発現異常を引起す遺伝子は単一ではなく複数の遺伝子が関わっている。これらの関連する複数の遺伝子を検索できることは疾患の原因解明に不可欠である。しかし、従来の手法では同一文献内に現れる遺伝子同士が強い相関を持つと仮定されていたため、関連遺伝子の検索精度は低かった。本研究では OMIM データベースを利用し、そこに記録されている疾患ごとに、関連する遺伝子を全て抽出し、それらの遺伝子と疾患からなるグラフ構造から関連遺伝子を求める方法について議論する。この方法により、関連遺伝子の検索精度が改善できる。

**キーワード**　情報検索、遺伝子検索、バイオインフォマティクス、OMIM

# A Gene Retrieval System Using a Hereditary Disease Database

HongTao PAN†, Jun MIYAZAKI†, Hidemi WATANABA†,††, and Shunsuke UEMURA†

† Nara Institute of Science and Technology, Gradduate School of Information Science
Takayama 8916–5, Ikoma, Nara, 630–0192 Japan
†† National Institute of Information
2–1–2 Hitotsubashi, Chiyoda-ku, Tokyo, 101–8430 Japan
E-mail: †{koto-h,miyazaki,hwatanab,uemura}@is.aist-nara.ac.jp, ††watanabe@nii.ac.jp

**Abstract**　It becomes clear that most of genetic diseases and hereditary ones are caused by genetic variation and the abnormalities of gene expression. The genetic variation and the abnormalities are involved by one or more genes. In order to reveal the cause of these diseases, it is mandatory to enable us to retrieve the genes that are biologically related to such diseases. However, previous works considered that a pair of genes was assumed to be mutually related if an arbitrary pair of genes appears in the same literature. Therefore, the accuracy of retrieving related genes was low. In this paper, we discuss the way to retrieve biologically related genes by using OMIM database, in which functions of genes and their concerned genes are described for each genetic or hereditary diseases. We first generate a graph structure from OMIM database, in which diseases and genes are expressed as nodes and a gene node and a disease node are linked if these are appeared in the same OMIM database entry. Then, we obtain all of mutually and biologically related genes from the graph. This approach can improve the accuracy of retrieving related genes.

**Key words**　Information Retrieval,Gene Retrieval,Bioinformatics,OMIM

## 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る [1] 。正常であるべき遺伝子に、なんらかの原因によって変異が発生すると、異常なタンパク質ができてしまう。その異常なタンパク質では、生命活動の正常な反応が進まないため疾患が発生する。人間の遺伝病や遺伝性疾患のほか、癌を始めとする多くの疾患が遺伝子の変異や発現異常により起こることが明白となってきた。これらの遺伝子は疾患責任遺伝子と呼ばれる。

医療の場においても、多くの疾患の病因が遺伝子レベルで解明されてきており、すでに遺伝子診断及び遺伝子治療という新

しい医療技術も実施されつつある。しかしながら、遺伝子データベース中に大量に蓄積されたデータから疾患に関わる複数の責任遺伝子を特定することはまだ困難である。

人間の遺伝子間の関連づけの研究は多数あり、多様なデータベースが構築されてきた。PubGene データベース [2] はその１つである。PubGene では、複数の遺伝子が同一文献内に存在した場合は、それらの遺伝子が潜在的な関連性があるという仮定に基づいている。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、生物学的な関連性を考慮していない。そのため、ある遺伝子に関連する遺伝子の検索精度が低かった。

本論文では、遺伝子間の生物学的に意味のある関連遺伝子を検索するための新しい手法を提案する。本手法は OMIM(Online Mendelian Inheritance in Man) データベース [3] を利用する。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る。OMIM に蓄積されたデータを分析し、疾患ごとに関連する責任遺伝子を全て抽出することにより、疾患を媒介して生物学的に意味のある関連遺伝子間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性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正確である。
- 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

高次元的な生物学の知識がテキストデータで記録される場合が多い。ところが現在行われているテキストデータ処理に対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。福田らの生物学文献から生物学専門用語を自動抽出する手法 [4] はその１つである。しかし、抽出された遺伝子名、タンパク質名など専門用語の意味および作用、機能などを知りたいとき、対応する論文を読まざるを得ない。

この問題に対して、文献から抽出された遺伝子、タンパク質、遺伝子-遺伝子、遺伝子-タンパク質、タンパク質-タンパク質など、関連性の深いもの同士をリンク付けする研究がある。B.J.Stapley らは酵母の中の遺伝子の関連性ネットワークの構築方法を提案している [5] 。一方、人間の遺伝子間の関連性を求めるために、T.K.Jenssen らは MedLine [6] 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た [7] 。この手法は、タイトルとアブストラクト中に遺伝子略号が共起した場合に、それらの遺伝子が潜在的に生物学的関連性があると仮定している。それらの共起関係を利用して人間の遺伝子間の関連データベース PubGene 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、遺伝子の関係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使用している。

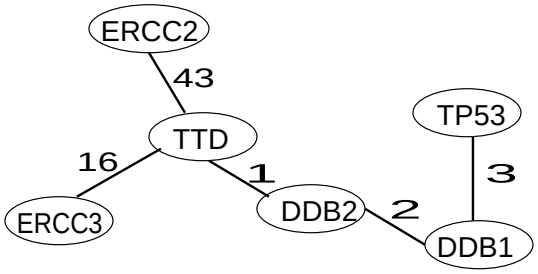


図 1　PubGene による関連遺伝子の検索

図 1 は、PubGene データベース中の関連する遺伝子のグラフの例を示している。この例では、入力した遺伝子 TTD と関連する遺伝子 ERCC2 の関連性の強さは 43 である。43 という重みは遺伝子 ERCC2 と遺伝子 TTD が 43 件の文献の中で共起したことを示している。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、遺伝子間の関連性の強さは生物学的な意味を考慮していない。

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) [8] により提供される人間の遺伝性疾患に関する遺伝子情報データベース OMIM がある。OMIM データベースは人間の遺伝性疾患を引き起こす遺伝子名、遺伝子機能、遺伝子地図、関連する遺伝子の情報など最新の情報が提供されている。しかし、それらの情報はすべてテキストデータである。

図 2 は OMIM データベースのデータ表現と関連する遺伝子を検索する様子を示している。遺伝子は全て疾患責任遺伝子であり、疾患を通じて遺伝子情報を表示できる。

例えば、慢性骨髄性白血病 (CML) を引き起こす責任遺伝子 ABL を入力すれば、遺伝子 ABL に関連する責任遺伝子と疾患を全部で 78 件表示する。その中で、関連する遺伝子 BCR を選ぶと、遺伝子 BCR に関する情報がすべて現れる。現れた情報の中で注目するのは「GENE FUNCTION」と「CYTOGENETICS」の項目である。「GENE FUNCTION」の項目では遺伝子 BCR の機能および関連する遺伝子の機能が説明され、「CYTOGENETICS」の項目では遺伝子 BCR の原因で引き起こされる疾患の情報が示される。この例で「CYTOGENETICS」の項目から遺伝子 BCR は CML だけではなく、急性リンパ性白血病 (ALL) も引き起こすことがわかる。

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には、関連する遺伝子情報はすべて表示されるが、個々の遺伝子情報がテキストデータで記述されるために、遺伝子と遺伝子の間に直接関連するかどうかについてはそのテキストデータを読まないと分からないし、関連する遺伝子の関連度の強さも判別できない。

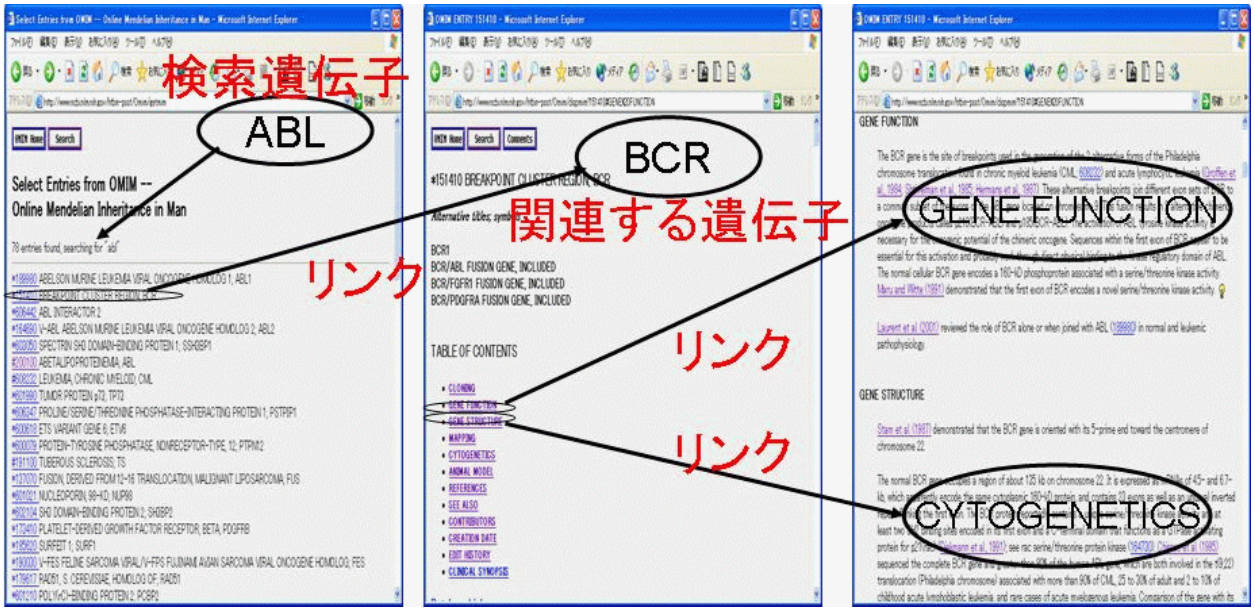


図 2　OMIM による関連遺伝子の検索

ここで、PubGene データベースと OMIM データベースの問題点をまとめておく。

PubGene データベースでは遺伝子を入力した場合、関連する遺伝子を表示できるが、

- 遺伝子の関連性が生物学的に意味を持つかどうか不明
- OMIM に存在するペアの遺伝子が PubGene には存在しない等、関連する遺伝子のリンクが不完全

という問題点がある。

一方、OMIM データベースは疾患と責任遺伝子を両方とも入力できる。いずれを入力した場合にも、関連する遺伝子および疾患をすべて表示できるが、

- 入力した遺伝子に直接関連しないが、間接的に関連する遺伝子も表示される
- 遺伝子間の関連度の強さが不明

という問題点がある。

PubGene データベースと OMIM データベースの問題点を解消するために、我々は新しい手法を次の節で提案する。

## 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。前節で説明したように、OMIM データベース中のすべてのレコードはテキストデータである。１つのレコードの中に複数の遺伝子と複数の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と仮定する。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ため、PubGene よりも正確な遺伝子の情報を与えていることによる。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。

- 前処理：疾患ごとに疾患責任遺伝子をグループ化する。

次に、疾患責任遺伝子間の関連性の強さを求める。

- 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。

疾患ごとに関連する疾患責任遺伝子をまとめ、疾患と遺伝子をリンクすることにより、遺伝子間の関連性がグラフで表現される。このグラフの構築手順を次節で説明する。

### 3.2　データベースの構築

OMIM データベースのテキストデータ中に含んでいるすべての疾患と遺伝子を抽出するために、疾患リストと遺伝子リストの情報が不可欠である。疾患リストは ICD-10 [9] と MeSH [10] から収集できる。ICD-10 は International Classification of Diseases Ten の略で、世界保健機関（WHO）が作成し、異なる国や地域から異なる時点で収集された死因と疾病の体系的なデータベースである。一方、MeSH は米国医学図書館（NLM）が作成したデータベースで、医学用語を分類したものである。

本研究での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することにより、テキストデータからグラフへの変換が必要である。今回の実験で、OMIM データベース中に全てのレコードの「GENE FUNCTION」と「CYTOGENETICS」の項目のデータに注目する。なぜならば、この 2 つの項目の中に遺伝子、疾患、関連する遺伝子、関連する疾患の情報がすべて記述されているからである。

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するには、形態素解析ツール（TreeTagger）[11] を利用する。具体的に、OMIM レコードごとに、「GENE FUNCTION」と「CYTOGENETICS」項目中のテキストデータを単語単位で切り出して、共起するすべての疾患と遺伝子を抽出する。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 3 の点線の右側参照)。そのグラフのデータ表現は表 1 の通りである。ある遺伝子 $G_i$ から疾患 $D_j$ へのリンクがあるとき、

$(G_i,D_j)$ を 1、そうでない場合、0 にしたものである。

表 1　有向グラフのデータの表現

| | $D_1$ | $D_2$ ⋯ | $D_n$ |
|---|---|---|---|
| $G_1$ | 1 | 1 ⋯ | 0 |
| $G_2$ | 1 | 1 ⋯ | 0 |
| $G_3$ | 0 | 1 ⋯ | 0 |
| $G_4$ | 1 | 0 ⋯ | 0 |
| ・ | ⋯⋯⋯⋯ | | |
| $G_m$ | 0 | 0 ⋯ | 0 |

形態素解析をするとき、多くの未知語 (原形を $<unknown>$ として表現する) が現れる。これらの多くは生物学の専門用語の可能性が高いと考えられる。未知語の処理、そして、OMIM データベースから遺伝子および疾患の抽出率の低下を避けるため、収集した遺伝子リストと疾患リストを専門用語として形態素解析ツールの辞書に加えることで対応可能である。

### 3.3　検索のアルゴリズム

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の強さ順でランキングすることである。

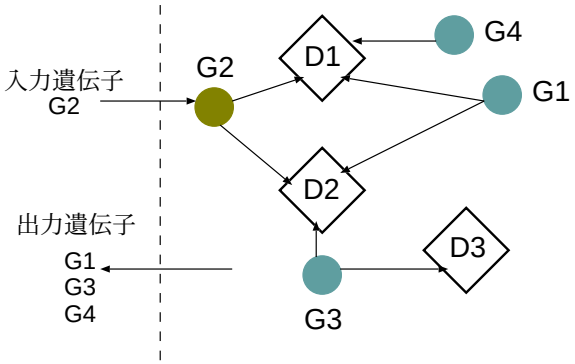


図 3　遺伝子と疾患の関連グラフ

図 3 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 $G_2$ との関連の強さに基づいてランキングして出力する。

具体的に遺伝子 $G_2$ と関連する遺伝子を求める方法は、まず、遺伝子 $G_2$ から関連する全ての疾患を探索する。探索した結果は疾患 $D_1$ と疾患 $D_2$ である。次に、疾患 $D_1$ と疾患 $D_2$ の原因となる全ての遺伝子を探索する。この結果、遺伝子 $G_1$、$G_2$、$G_3$ および $G_4$ が得られる。そして、遺伝子 $G_1$、$G_2$、$G_3$、$G_4$ 及びそれらの遺伝子から関連する全ての疾患を抽出し、サブグラフ (図 3 ) を構築する。

表 2　図 3 のグラフのデータ表現

| | $D_1$ | $D_2$ | $D_3$ |
|---|---|---|---|
| $G_1$ | 1 | 1 | 0 |
| $G_2$ | 1 | 1 | 0 |
| $G_3$ | 0 | 1 | 1 |
| $G_4$ | 1 | 0 | 0 |

抽出したサブグラフ (図 3 ) のデータ構造を表 2 で表現する。表 2 は $m*n$ の論理型行列であって、$m$ は遺伝子の個数、$n$ は疾患の個数とする。

入力した遺伝子 $G_2$ に対する $(G_2,D_j)$ を 1 にするとき、対応する疾患 $D_j$ を求める。そして、疾患 $D_j$ から $(G_i,D_j)$ を 1 にする遺伝子 $G_i$ を求める。この遺伝子 $G_i$ が入力した遺伝子 $G_2$ との関連する遺伝子である。この結果、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。

関連の強さの計算は次節で説明する。

### 3.4　遺伝子間の関連の強さの計算

入力した遺伝子と関連する遺伝子の関連の強さは、2 つの遺伝子に共通する疾患の数と定義し、重みと呼ぶことにする。重みの計算方法は次の通りである。

- 関連する遺伝子と入力した遺伝子が同じ疾患にリンクしている場合には、その関連する遺伝子の重みに 1 をプラスする。
- 関連する遺伝子と入力した遺伝子が異なる疾患にリンクしている場合には、その関連する遺伝子の重みに 0.5 をプラスする。

例えば、図 3 で入力した遺伝子 $G_2$ は疾患 $D_1$ を通じて遺伝子 $G_1$ と遺伝子 $G_4$ に関連しているので、$(G_1,D_1)$ と $(G_4,D_1)$ を 1 とする。次に、入力した遺伝子 $G_2$ は疾患 $D_2$ を通じて遺伝子 $G_1$ と遺伝子 $G_3$ にも関連しているので、$(G_1,D_2)$ と $(G_3,D_2)$ を 1 とする。この場合、遺伝子 $G_3$ は疾患 $D_3$ にも関連しているが、入力した遺伝子 $G_2$ は疾患 $D_3$ に直接関連していないため、$(G_3,D_3)$ を 0.5 とする。

表 3　遺伝子 $G_2$ に関連する遺伝子の重み

| | $D_1$ | $D_2$ | $D_3$ |
|---|---|---|---|
| $G_1$ | 1 | 1 | 0 |
| $G_3$ | 0 | 1 | 0.5 |
| $G_4$ | 1 | 0 | 0 |

以上の方法を利用して、入力した遺伝子 $G_2$ と関連する遺伝子の重みがすべて計算される (表 3 )。そして、遺伝子ごとにそれらの重みを加算する。この例では、入力した遺伝子 $G_2$ に関連する遺伝子の出力順序は表 4 の通りとなる。

表 4　$G_2$ に関連する遺伝子のランキング

| 遺伝子 | 重み |
|---|---|
| $G_1$ | 2 |
| $G_3$ | 1.5 |
| $G_4$ | 1 |

この例からわかるように、遺伝子 $G_1$ と遺伝子 $G_2$ は両方とも疾患 $D_1$ と疾患 $D_2$ に関連するので、この 2 つの遺伝子間の関連性は強く、これらは生物学的な関係が強いと解釈できる。

## 4.　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を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。この方法を利用することにより、ある遺伝子に関連する遺伝子の検索精度が改善できる。

しかし、関連する遺伝子の重みの計算方法は単に共通する疾患の数であり、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。遺伝子間の本質的な生物学意味を表現できるような重みづけの方法を開発することが必要である。

現在、Web の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 [14] の Hub と Authority の概念を導入することを検討中である。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 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。

さらに、提案した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て評価していきたい。

## 文　　献


[1] Arthur M.Lesk,“Introduction to Bioinformatics”, OXFORD University Press,p93-97,2002.

[2] PubGene, http://www.pubgene.org

[3] OMIM, http://www.ncbi.nlm.nih.gov/omim/.

[4] 福田賢一郎，角田達彦，田村あゆち，高木利久, “医学生物学文献からの専門用語の抽出に向けて－タンパク質名の自動抽出”，情報処理学会論文誌，Vol.39，No.8，pp2421-2430，1998.

[5] B.J.Stapley，G.Benoit，“Biobibliometrics:information retrieval and visualization from co-occurrencess of gene names in Medline abstracts”，Pac.Symp.Biocomput.5，529-540，2000.

[6] MedLine, http://www.nlm.nih.gov/

[7] T.K. Jenssen，A. Laegreid，J. Komorowski and E.Hovig，“A literature network of human genes for high-throughput analysis of gene expression”，Nature Genetics，Volume 28，May 2001，pp21-28.

[8] NCBI, http://www.ncbi.nlm.nih.gov/

[9] ICD-10, http://www.who.int/whosis/icd10/

[10] MeSH, http://www.nlm.nih.gov/mesh/meshhome.html

[11] TreeTagger http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html

[12] BIOME http://biome.ac.uk/biome.html

[13] Y.C.Tao and R.L.Leibel,“Identifying functional relationships among human genes by systematic analysis of biological literture”, http://www.biomedcentral.com/1471-2105/3/16.

[14] Jon M.Kleinberg，“Authoritative Sources in a Hyperlinked Environment”, Journal of the ACM，Vol.46，No.5，pp604–632，1999.