# 複数の疾患との共起性に着目した 生物医学文献からの創薬標的遺伝子候補の抽出

權 娟大 †　　清水 將吾 ‡　　宮崎 智 †

† 東京理科大学 薬学部

‡ 産業技術大学院大学 産業技術研究科

†{yekwon, smiyazak}@rs.noda.tus.ac.jp

‡shimizu-syogo@aiit.ac.jp

## 1　はじめに

本研究では，生物医学文献から創薬標的遺伝子の候補を抽出することを目的とする．生物医学文献から疾患関連遺伝子を抽出する研究は多数行われている [1] が，これらは特定の疾患に対する原因遺伝子としての関連の強さの評価に重点を置いており，創薬標的遺伝子の候補となるために必要な，他の疾患との関連によって引き起こされる副作用の可能性については考慮されていない．近年，遺伝子，疾患，薬品用語間の関係を抽出する手法がいくつか提案されている [2] が，辞書が必要であり，文献中に薬品に関係する用語が明示的に出現しなければ，関係性を抽出することは困難である．

そこで，本研究では，遺伝子と複数の疾患との共起関係によって，薬品辞書を用いずに副作用の可能性を評価し，創薬標的遺伝子の候補を抽出する手法を提案する．更に，本手法による創薬標的遺伝子候補の抽出結果を薬理ゲノミクス知識ベース PharmGKB [3] を使って評価する．

## 2　提案手法

疾患と遺伝子の関連性の評価基準として，文献中での用語の共起頻度を使用する．特定の疾患に対して，ある遺伝子を標的遺伝子としたときに副作用が起こる可能性を，その遺伝子が関連をもつ異なる疾患の数によって評価する．標的疾患との関連が強く，かつ，他の疾患との関連も少ない場合にその遺伝子を創薬標的遺伝子の候補として抽出する．関連する疾患の数は，標的疾患との関連性を考慮して定量化する．

### 2.1　用語間の共起関係の抽出

生物医学文献集合として米国国立医学図書館から入手した PubMed アブストラクト，遺伝子名辞書として NCBI Gene のヒト遺伝子，疾患名辞書として CTD（Comparative Toxicogenomics Database）と MeSH（Medical Subject Headings）を使用する．遺伝子名辞書は同義語を含め 117,170 エントリ，疾患名辞書は同義語を含め 48,480 エントリからなる．

これらの辞書を用いて，PubMed アブストラクトから遺伝子名と疾患名の共出現を抽出した．同義語は対応する公式名に置き換えることで同じ実体として処理した．また，遺伝子シンボル名が一般的な英単語と同じ綴りをもつことがある（例えば，LARGE，IMPACT）ため，遺伝子名から作成した構成語が同時に出現しているかどうかを追加で検査することによって誤検出を減らした．共出現の範囲は同一文中とした．

### 2.2　標的遺伝子としての指標値

上で抽出した共起関係に基づき，疾患に対する遺伝子の関連度を tf-idf と同様の手法により評価する．まず，標的疾患 $d$ と共起する遺伝子群を $g_1, \ldots, g_n$ としたとき，$d$ に対する遺伝子 $g$ の共起出現頻度 $gtf_d(g)$ を次式で定義する．

$$gtf_d(g) = \frac{n_d(g)}{\sum_{i=1}^{n} n_d(g_i)}$$

ここで，$n_d(g)$ は $d$ と $g$ が共起する文献数である．$gtf$ は $d$ に対する $g$ の関連度合いを表す．

次に，$g$ を 薬の標的遺伝子とした場合に副作用が起こる可能性を，次式によって評価する．

$$adf_d(g) = \log \frac{M}{ad_d(g)}$$

表 1: 癌に対する標的遺伝子の抽出結果

| Disease | Drug | Target | Rank by $as$ |
|---|---|---|---|
| Breast neoplasms | capecitabine | DPYD | – |
| | cetuximab | EGFR | 52 |
| Colonic neoplasms | cetuximab | EFGR | 37 |
| Colorectal neoplasms | capecitabine | DPYD | 9 |
| | cetuximab | EGFR | 37 |
| Gastrointestinal neoplasms | capecitabine | DPYD | 5 |
| Head and neck neoplasms | capecitabine | DPYD | 43 |
| | cetuximab | EGFR | 1 |
| Lung neoplasms | cetuximab | EGFR | 7 |
| Pancreatic neoplasms | capecitabine | DPYD | 88 |
| | cetuximab | EGFR | 25 |

ここで，$M$ は全疾患の数，$ad_d(g)$ は $g$ と共起する疾患の数である．但し，疾患同士が互いに関連している場合は，それらがまったく別の疾患である場合と比較して副作用が起こる可能性が高くなると考えられるため，疾患間の関連性の有無を考慮して $ad_d(g)$ の値を定量化する．互いに関連している疾患同士は似たような遺伝子群と関連するという仮定に基づき，二つの疾患 $d_1, d_2$ の関連度を次のように定義する．

$$drel(d_1, d_2) = \frac{||G(d_1) \cap G(d_2)||}{||G(d_1) \cup G(d_2)||}$$

ここで，$G(d)$ は疾患 $d$ と関連する遺伝子の集合であり，$||G||$ は集合 $G$ の基数を表す．$ad_d(g)$ は $g$ と関連をもつと判定された各 $d_i$ に対する $drel(d, d_i)$ の合計値として定義する．

標的疾患 $d$ に対する $g$ の創薬標的遺伝子としての良さ $as_d(g)$ を次式で定義する．

$$as_d(g) = gtf_d(g) \cdot adf_d(g)$$

## 3 実験

提案手法によって，既知の標的遺伝子が抽出されるかどうかを検証する．疾患と標的遺伝子に関する情報は人手によって作成された薬理ゲノミクス知識ベースである PharmGKB を利用する．薬品が実在する標的遺伝子は副作用が少ないと考えられるため，同じ標的遺伝子を上位に抽出することができれば，標的遺伝子が未知である疾患に対しても本手法による創薬効率化支援が期待できる．

表 1 に，多くの癌の治療薬である capecitabine と cetuximab の既知の標的遺伝子と当該標的遺伝子の $as_g(d)$ での順位を示す．これらはそれぞれ標的遺伝子として DPYD (dihydropyrimidine dehydrogenease) と EGFR (epidermal growth factor receptor) を用いている．DPYD は gastrointestinal neoplasms で，EGFR は head and neck neoplasms でそれぞれ 5 位と 1 位にランクしており，この結果から capecitabine は他の癌より gastrointestinal neoplasms に，cetuximab は head and neck neoplasms に有効な薬であることが予想できる．

## 4 おわりに

本稿では，生物医学文献データベースから，遺伝子/疾患用語の共出現や疾患間の類似性に基づいて創薬標的遺伝子候補を抽出する手法を提案した．今後は，副作用データベースの情報を組み入れ，標的遺伝子抽出精度の改善を試みる予定である．

## 参考文献


[1] Y. Garten, A. Coulet, and R.B. Altman, “Recent progress in automatically extracting information from the pharmacogenomic literature,” Pharmacogenomics 11 (10), pp.1467–89, 2010.

[2] F. Rinaldi, G. Schneider, and S. Clematide, “Mining complex drug/gene/disease relations in PubMed,” Proc. of the workshop ”Mining the Pharmacogenomics Literature”, Pacific Symposium on Biocomputing, 2011.

[3] E.M. McDonagh, M. Whirl-Carrillo, Y. Garten, R.B. Altman, and T.E. Klein, “From pharmacogenomic knowledge acquisition to clinical applications: the PharmGKB as a clinical pharmacogenomic biomarker resource.” Biomarkers in Medicine 5(6), pp.795–806, 2011.