10

PubMedテキストマイニングにおける遺伝子等の語句判定手法の開発

○平田誠、坂手龍一、木村友則

国立研究開発法人 医薬基盤・健康・栄養研究所　難治性疾患研究開発・支援センター

PubMedなどのテキストマイニングにおいて、遺伝子名（Gene symbol）等を検索しても関係無い語句にヒットすることがある。こうした場合は、目的とする論文の抽出に不都合となる。例えば、NMRは核磁気共鳴装置の略称として一般的だが、Gene symbolにも登録されている（LINC01672のAlias）。

そこで、検索でヒットした語句が、どの属性（遺伝子、薬剤、疾患）に近いかを判定する手法を開発した。この手法は、まず、①各属性に関係性の強い論文を教師データとして全語句の出現頻度を調べる。それをもとに、②対象となる語句を含むテキスト（文）の構成語句を評価することで、その語句の属性を判定する。教師データとして、遺伝子は HUGO、薬剤はDrugBankの各々の参照論文、疾患はDDrare（指定難病）の疾患名の検索でヒットした論文を用いた。

最近の15万件の論文（タイトル、要旨）をテストした結果、例えば、遺伝子については検索でヒットした語句の半数ほどがFalse-positiveであり、薬剤や疾患と比べてもその率が高いことがわかった。遺伝子は、論文要旨にGene symbolのみの記載が多いことも影響している。この手法は、文章を人が読解する方法をモデルとしており、より正確な文章の解釈に活用が可能と考えられる。

1. 語句判定としての属性（疾患、薬剤、遺伝子）評価

■ 目的・解析手法

テキスト中の遺伝子などをサーチしても、関係のない語句にヒットすることがある。そこで、本研究では、疾患、薬剤、遺伝子の３つの属性について、関係性の強いテキスト集団（教師データ）と、偏りのないテキスト集団（テストデータ）との間において、文ごとの構成語句（動詞、名詞、形容詞、副詞、外来語）の出現頻度を比べ、その文がどの属性に近いかを判定することを試みた。TreeTaggerで品詞付けし、文ごとに単語を正規化（appleとapplesを同じ語句とするなど）して出現頻度を計算した。属性スコアが0より大きいほど、その文はその属性に強く関係し、よって、文中の語句（検索でヒットした遺伝子名など）がその属性である可能性が高いと考えた。教師データをもとに、近年のPubMed論文（file No. 740-1238）をテストした。

教師データ：PubMed論文

	DB	論文数	対象論文
疾患	DDrare	178,100	疾患名で検索してヒットした論文
薬物	DrugBank	9,316	各DrugのGeneral References
遺伝子	HUGO	21,291	各遺伝子のReferences

■ 結果１

疾患（Disease）と薬剤（Drug）は、”属性スコア=0”を中心とする、ほぼ左右対称な分布なのに対し、遺伝子（Gene）は”属性スコア<0”への偏りが見られる 。遺伝子はGene symbolでの表記が多く、検索でヒットしたものの、遺伝子ではない語句が多い（False positive）と推察される。この遺伝子の分布も考慮して、我々は、暫定的に±0.2において、属性スコアの判定基準を3段階に定めた（最終スライド参照）。

図１．疾患、薬剤、遺伝子の属性スコア分布

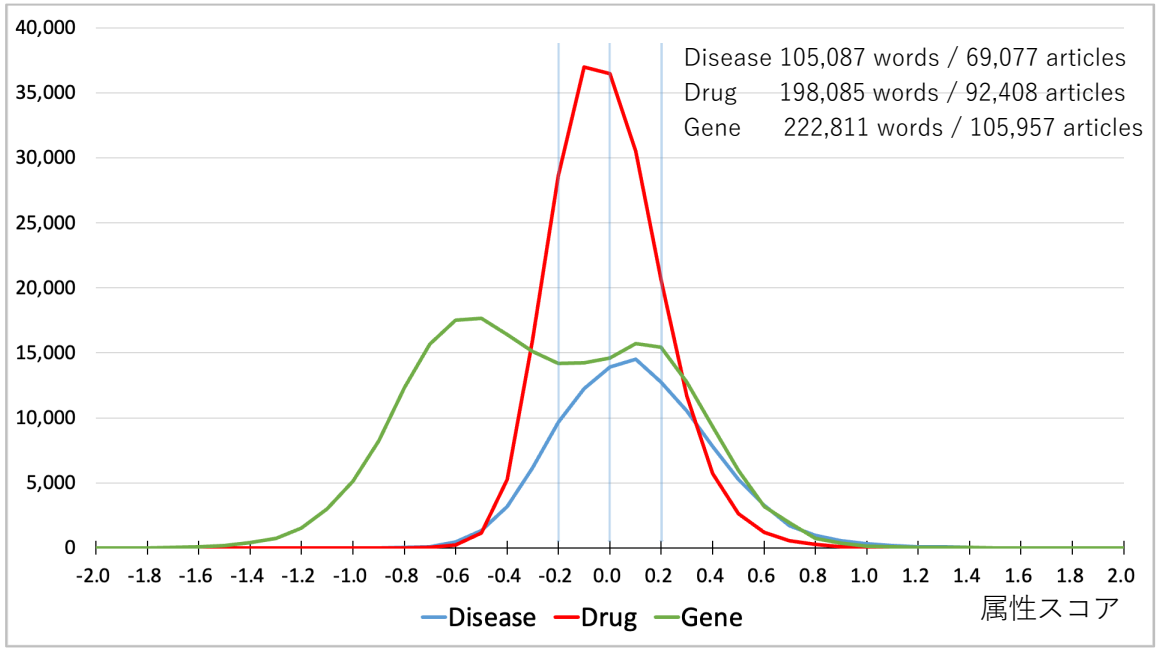
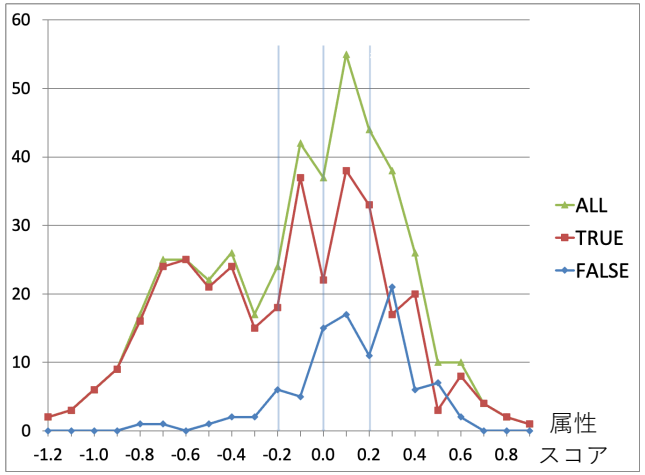


図２．キュレーション（Gene）



PubMed235論文（タイトル、要旨）の445遺伝子について、目視で属性の正解（TRUE）／不正解（FALSE）を確認した結果、TRUE: 21.8%、FALSE: 78,2%であった。

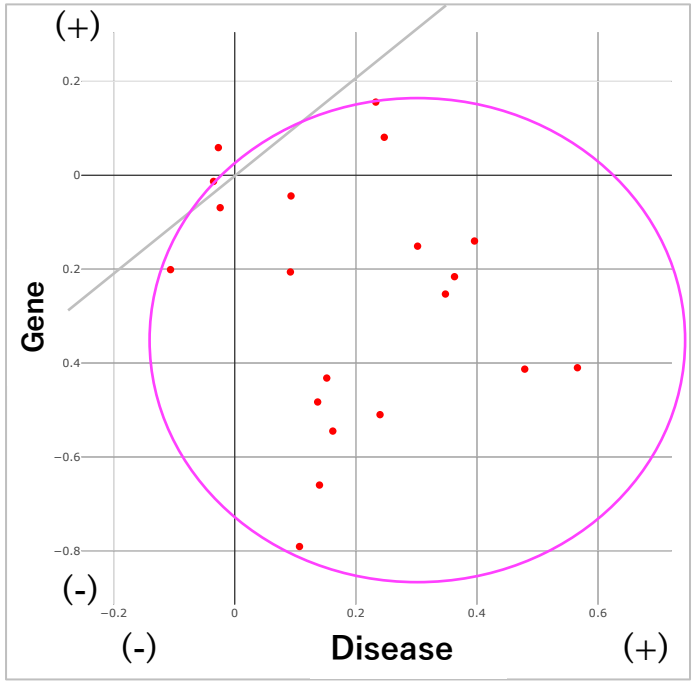
属性スコアが負（< 0）の時、正解率が高い傾向が見られた。実際のテキスト検索において、特に属性スコアが負の場合に、その語句を検索結果から除外するなどの改善が可能と考えられる。

2. 属性スコアによる評価結果の例

■ 結果２

属性スコアによる語句判定については、興味深いケースが見られたので紹介する。

図３． ２つの属性スコアを持つ語句（Gene – Disease: CF）



CFは一般にCystic Fibrosis（嚢胞性線維症）の略語であるが、遺伝子名としても存在する。

左のグラフの20件のデータでは、全体として疾患（Disease）としての属性スコアが遺伝子（Gene）より大きく、疾患として判定されていることがわかる。

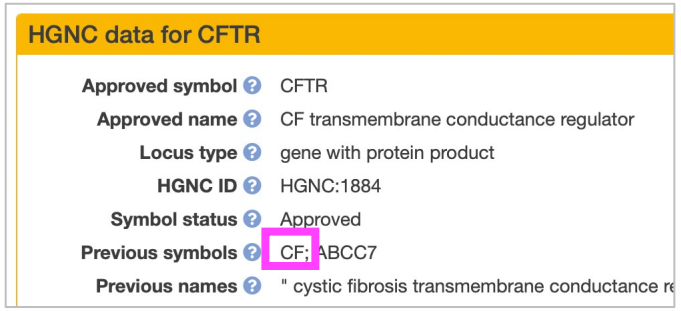
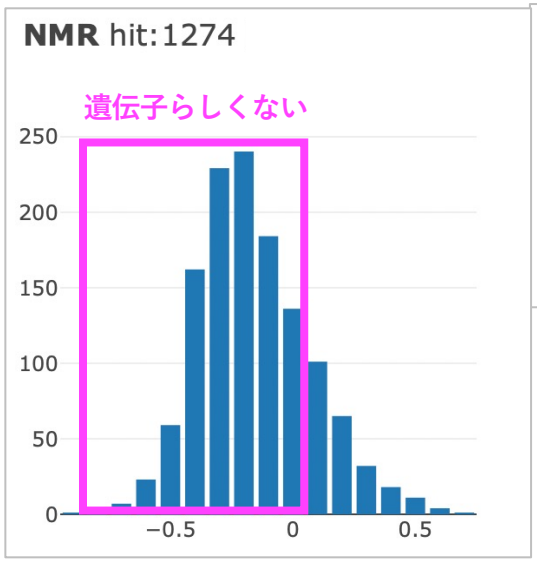
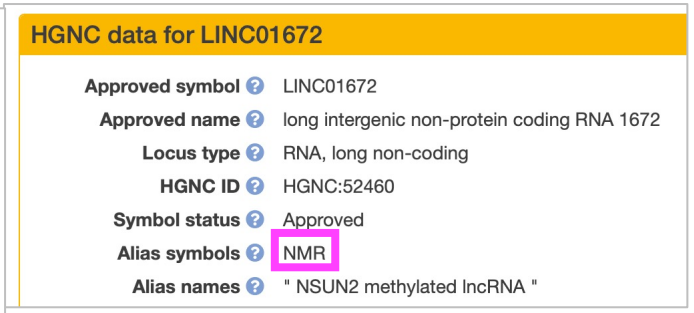


図４． 複数回出現する単語の属性スコア分布（Gene: NMR）





“NMR”は遺伝子としても登録されている（LINC01672のAlias）が、核磁気共鳴装置の略称として一般的であり、出現頻度も高い。属性スコア（Gene）も負のケースが多い。

■ まとめ・考察

本研究の語句判定手法は、PubMed以外にも応用可能であり、日本語でも分かち書きや、辞書、オントロジーが適正に利用できれば、有用な判定が可能であると考えられる。また、属性としても、疾患、薬剤、遺伝子以外にも応用可能である。本手法は機械学習ではないが、教師データの作成が同様に重要である。

本研究ではデータドリブンな形での語句判定として、属性スコアを考案した。文章を人が読解する方法をモデルとしており、より正確な文章の解釈に活用が可能であるとの予測のもとに実施した。一方、「属性」の概念については、文字通り、文脈による属性の定義、属性間の距離といった概念が存在する。

例えば、本研究はHUGO（遺伝子）のデータを用いたが、これはPDB（タンパク質）とも結果として近い属性である。実際に、要旨中で、遺伝子なのかタンパク質（酵素）なのかといった、判定の揺らぎがみられた。

また、PubMed論文（タイトル、要旨）では、独自の省略語や、同一文章中で同じ単語が複数の意味を持つような記載、あるいは、時系列での意味の変化なども想定され、より正確な語句判定を行うツールとして、改良していく計画である。

3. PubMedのテキストマイニング例

PubMed論文タイトルと要旨を対象に、疾患名、薬物名、遺伝子名*を検索して、それらのヒットした語句をスコア**に応じてハイライト表示するシステムを開発。

* 疾患名（DDrare, MalaCards）、薬物名（DrugBank）、遺伝子名（HUGO）

** スコア
（同じ語句が複数回出現する場合は平均）

小 ← その属性らしさ → 大

Disease	~ -0.2	-0.2 ~ 0.2	0.2 ~
Drug	~ -0.2	-0.2 ~ 0.2	0.2 ~
Gene	~ -0.2	-0.2 ~ 0.2	0.2 ~

A phase 1 healthy male volunteer single escalating dose study of the pharmacokinetics and pharmacodynamics of risdiplam (RG7916, RO7034067), a SMN2 splicing modifier

Risdiplam (RG7916, RO7034067) is an orally administered, centrally and per
(SMN2) mRNA splicing modifier for the treatment of spinal muscular atrophy
were to assess the safety, tolerability, pharmacokinetics (PK) and pharmacody
CYP3A inhibitor itraconazole on the PK of risdiplam in healthy male volunte
design with 25 subjects receiving single ascending oral doses of risdiplam (ra
Bayesian framework was applied to estimate risdiplam effect on SMN2 mRNA
PK of risdiplam was also assessed using a two-period cross-over design (n =
tolerated. Risdiplam exhibited linear PK over the dose range with a multi-pha
Food had no relevant effect, and itraconazole had only a minor effect on plas
metabolized by CYP3A. The highest tested dose of 18.0 mg risdiplam led to a
of the estimated maximum increase in SMN2 mRNA. Risdiplam was well tole
the intended shift in SMN2 splicing towards full-length SMN2 mRNA. Based or
with SMA are now ongoing.

Di/-0.2013/139 Dr/0.4139/142 Ge/-0.3841/147

Disease	~ -0.2	-0.2 ~ 0.2	0.2 ~	Di:Muscular Atrophy=0.01 Atrophy=0.015
Drug	~ -0.2	-0.2 ~ 0.2	0.2 ~	Dr:Itraconazole=0.506 , D
Gene	~ -0.2	-0.2 ~ 0.2	0.2 ~	Ge:SMA=-0.275 , Ge:SMN

Disease: Spinal muscular atrophy/SMA

PK ['pharmacokinetics']
SMA ['spinal muscular atrophy']
SMN2 ['survival of motor neuron 2']

Glucosylceramide synthase inhibition with lucerastat lowers globotriaosylceramide and lysosome staining in cultured fibroblasts from Fabry patients with different mutation types

Fabry disease is an X-linked lysosomal storage disorder caused by mutations
alpha-GalA). The deleterious mutations lead to accumulation of alpha-GalA s
globotriaosylsphingosine. Progressive glycolipid storage results in cellular dys
i.e. neuropathic pain, impaired renal function and cardiomyopathy. Many
infusions of replacement enzyme. While the only available oral therapy is an
number of patients with specific 'amenable' mutations. Lucerastat is an ora
(GCS) that is in late stage clinical development for Fabry disease. Here we ir
globotriaosylsphingosine and lysosomal staining in cultured fibroblasts from
different pathogenic variants, with 13 cell lines harboring GLA mutations as
dose dependently reduced Gb3 in all cell lines. For 13 cell lines the Gb3 data
[interquartile range (IQR)] = 11 μM (8.2-18); the median percent reduction
also dose dependently reduced LysoTracker Red staining of acidic compartme
to those with current treatments-agalsidase alfa and migalastat. Consequ
mechanism to reduce Gb3 accumulation and lysosome volume, suitable for a

Di/0.0248/175 Dr/0.1283/169 Ge/-0.0864/166

Disease	~ -0.2	-0.2 ~ 0.2	0.2 ~	Di:Cardiomyopathy=0.46
Drug	~ -0.2	-0.2 ~ 0.2	0.2 ~	Dr:Agalsidase alfa=0.340 Dr:Lucerastat=0.219 , D
Gene	~ -0.2	-0.2 ~ 0.2	0.2 ~	Ge:GCS=-0.290 , Ge:GLA

GCS ['glucosylceramide synthase']
IQR ['interquartile range']

Immune-mediated inflammatory polyneuropathy overlapping Charcot-Marie-Tooth 1B

Charcot Marie Tooth (CMT) due to myelin protein zero (MPZ) mutations, may cause a wide variation of phenotypes, depending on the localization of the mutation within the gene. Among the most common phenotypes are: an infantile onset disease with extremely slow nerve conduction velocities (CMT1B) and an adult onset phenotype with nerve velocities in the axonal range (CMT2I). We reported a patient with CMT1B (MPZ p.Ser63del mutation) which developed an overlapping immune mediated polyradiculoneuropathy with recurrent episodes of quadriparesis and cranial nerve involvement. We observed reversible conduction block on serial neurophysiologic studies, non-uniform demyelination and good clinical response to prednisone and cyclophosphamide, as evidenced by objective functional recovery. Chronic inflammatory demyelinating polyradiculoneuropathy (CIDP)-like characteristics have not yet been described associated with a MPZ p.Ser63del mutation. This description adds evidence indicating that a defective structural myelin protein may predispose peripheral nerves to immune attacks.

Di/0.2417/102 Dr/-0.1198/99 Ge/-0.0311/103

Disease	~ -0.2	-0.2 ~ 0.2	0.2 ~	Di:CIDP=0.444 , Di:CMT=0.095 , Di:Charcot Marie Tooth=0.095 Di:Chronic Inflammatory Demyelinating Polyradiculoneuropathy=0.444 , Di:Polyradiculoneuropathy=0.418
Drug	~ -0.2	-0.2 ~ 0.2	0.2 ~	Dr:Cyclophosphamide=0.076 , Dr:Prednisone=0.076
Gene	~ -0.2	-0.2 ~ 0.2	0.2 ~	Ge:CMT1B=-0.134 , Ge:CMT2I=-0.165 , Ge:MPZ=0.058 Ge:myelin protein zero=0.323

Disease: Charcot Marie Tooth/CMT
GENE: MPZ/CMT1B ; CMT1B/CMT2I ; MPZ/myelin protein zero

CMT ['charcot marie tooth']
MPZ ['myelin protein zero']