ゲノム医療を推進するAI技術
—文献検索AI，病原性を予測するXAI（説明可能なAI）—

村上　勝彦，森田　一，光石　豊，馬場　謙介，マルティネス　アンデル，
グェン　レ　アン，多湖　真一郎，小林　賢司，阿部　修也，富士　秀

本稿ではがんゲノム医療に向けて開発した二つのAIを紹介する．一つ目は文献検索AIで，たとえば「遺伝子変異Aをもつタイプのがんに効く薬剤は何か」といった論文中の記述表現を直接検索できる．また，画像分類により特定の図を優先して探すことができる．本技術は，実証実験で検索時間の半減を達成した．また有料で利用可能である．二つ目は遺伝子変異ががんの原因であるかという病原性の予測および予測説明可能なAIである．予測だけでは思考過程が見えないという課題を解決するため，予測の理由と根拠を提示する仕組みをつくり，説明ができるAIを開発した．

キーワード：自然言語処理，説明可能AI，がんゲノム医療

1. はじめに

背景知識としてがんゲノム医療の課題と研究の問題設定について紹介する．古くからあるがん医療の薬物療法では，特定のがんのタイプを臓器などの発症部位などで診断し，決められた薬剤を用いるが，それでは30%の患者にしか効果がみられないといわれる．近年では，遺伝子検査でがん患者のがん組織を個別に調べ，どの遺伝子変異があるかによって薬剤を使い分ける治療が行われる．それにより高い奏効率を達成する（あるがんでは75%）．

しかしながら，現状ではどの遺伝子変異でも薬剤があるわけではなく，使える薬剤が見いだせるのは，遺伝子変異を調べた患者の約1割(9%)程度に過ぎない．そこで新たな薬剤の開発や，新たな治療法（既存の薬の別のがんでの有効性がわかるなど）の研究が盛んに行われている．

医師は最新の知見に基づき治療方針を決めるために，しばしば「この遺伝子変異Aをもつ“がんB”には，どの薬が有効か」を確認する目的の文献検索を行う．ここに，研究論文の数が年々指数関数的に増加しているため，文献検索に非常に時間がかかるという課題が生じている．たとえばがん関連だけの論文をcancer（がん）とtumor（腫瘍）だけで検索したとき2018年だけでも20万件もの論文がヒットする．

そこでわれわれはがんゲノム医療を推進するために二つのAIを開発した．一つ目は，効果的な薬剤について論文報告がある遺伝子変異を検討するためのAIで，その情報に効率よく辿り着くための文献検索AIである（2節）．二つ目は，薬剤報告がない遺伝子変異を検討するためのAIで，変異が「がん」の原因であるのか（病原性の程度）を予測するAIである（3節）．

これらのAI技術の開発にあたっては，東京大学医科学研究所の医師らとともに議論をしながら進めた．

本稿の最後に，結言と今後の課題を述べる．

2. 文献検索AI

2.1 文献検索システム

本節ではわれわれが開発したブラウザベースの文献検索システムについて，画面のスナップショットを用いて紹介する．全機能の説明はせず，基本的な機能，および特徴的な機能に絞って説明する．

むらかみ　かつひこ，もりた　はじめ，
みついし　ゆたか，ばば　けんすけ，
マルティネス　アンデル，グェン　レ　アン，
たご　しんいちろう，こばやし　けんじ，
あべ　しゅうや，ふじ　まさる
富士通株式会社 富士通研究所
〒211–8588 川崎市中原区上小田中4–1–1
murakami.kt@fujitsu.com, hmorita@fujitsu.com,
mitsuishi-y@fujitsu.com, ander@fujitsu.com,
nguyenle.an@fujitsu.com, s-tago@fujitsu.com,
kobayashi_kenji@fujitsu.com, abe.shuya@fujitsu.com,
fuji.masaru@fujitsu.com
ばば　けんすけ
岡山大学サイバーフィジカル情報応用研究コア（現）
〒700–8530 岡山市北区津島中3–1–1
kenbaba@okayama-u.ac.jp

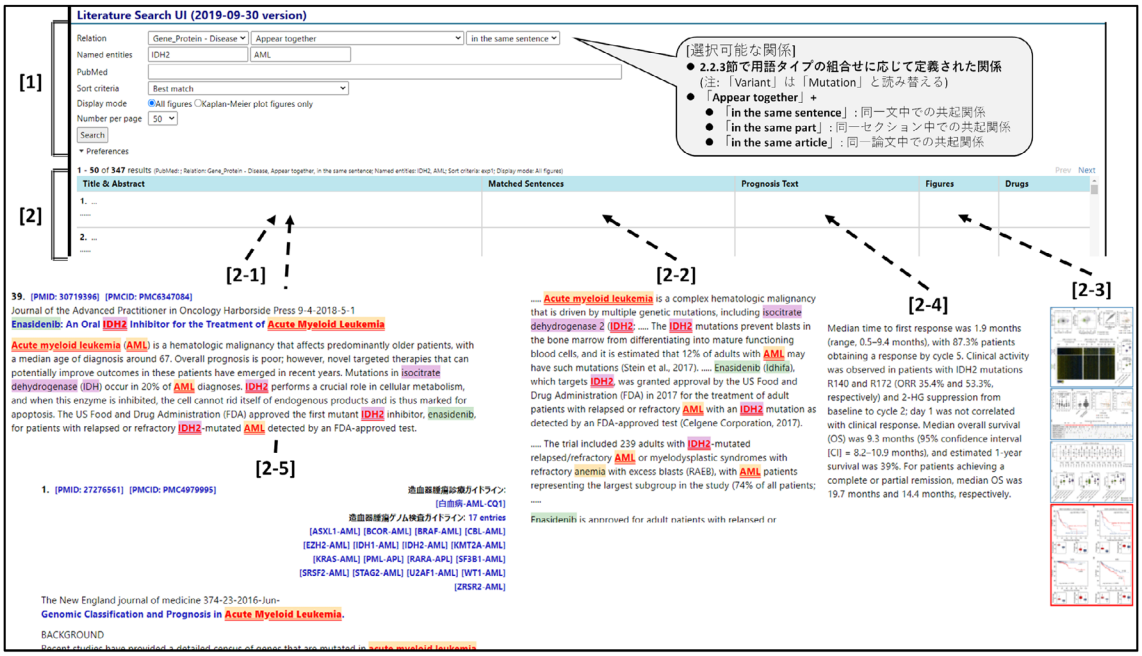

図 1　文献検索システムの検索例

本文献検索システムでは，メタ情報を除くコンテンツに関する情報としてはタイトルとアブストラクトのみを保有する MEDLINE [1] という医療生命科学分野の文献データベースのデータと，PMC [2] という本文まで含めた医療生命科学分野の文献情報を集めたサイトのデータのうち，白血病をはじめとする造血器腫瘍に関する文献を試験的に検索対象とした．文献数は，MEDLINE が 831,720 件，PMC が 60,693 件で，合わせて 892,413 件である．

図 1 は文献検索システムの検索例である．【1】は入力部，【2】は出力部である．

「【1】入力部」では，「Named entities」の欄に検索したい用語タイプの用語二つと，メニュー「Relation」からそれらの用語間に成り立つ「関係」を指定する．【1】では，用語タイプ「遺伝子」(Gene_Protein) の用語「IDH2」と用語タイプ「疾患」(Disease) の用語「AML」（白血病の一種）が「同一文中で共起している」(Appear together in the same sentence) という関係を検索条件としている．

用語タイプには，遺伝子，疾患，変異 (Mutation) および薬剤 (Drug) の合計四つがある．用語タイプの組合せで指定可能なものは，疾患と薬剤，遺伝子と疾患，遺伝子と薬剤，変異と疾患，変異と薬剤の 5 通りのうち一つである．関係 (Relation) で指定可能なものは，2.2.3 節で用語タイプの組合せに応じて定義される関係（ただし，「Variant」は「Mutation」と読み替える）か，「同一文中での共起」，「同一セクション中での共起」，「同一論文中での共起」である．用語については，2.2.1 節で述べる固有表現抽出技術と 2.2.2 節で述べる同一性判定技術により同定された用語を検索対象としており，関係のうち Positive と Negative については，2.2.3 節で述べる関係抽出技術により抽出された関係を検索対象としている．

こうした関係の検索により，たとえば患者の疾患と患者がもつ遺伝子変異の関連を調べたり，患者の疾患と投与したい薬剤の関連を調べたりすることができる．二つの用語のうち片方を空欄にするとワイルドカードの扱いとなり，たとえば一つの疾患に対する薬剤を網羅的に調べることも可能である．

「【2】出力部」には検索でヒットした論文一覧が検索結果として表示される．説明のため，各列に表示される情報の例を拡大して，実際の出力画面外に【2-1】～【2-5】として配置している．

【2-1】にはタイトル，アブストラクトが，【2-2】には入力にマッチした本文中の文が表示される．四つの用語タイプである疾患，遺伝子，変異，薬剤に対応する用語は，異なる色でハイライトされる．【2-1】や【2-2】では，「Acute myeloid leukemia」や「AML」が疾患として，「IDH」が遺伝子として，「enasidenib」が薬剤としてハイライトされている．図 1 には表示されて

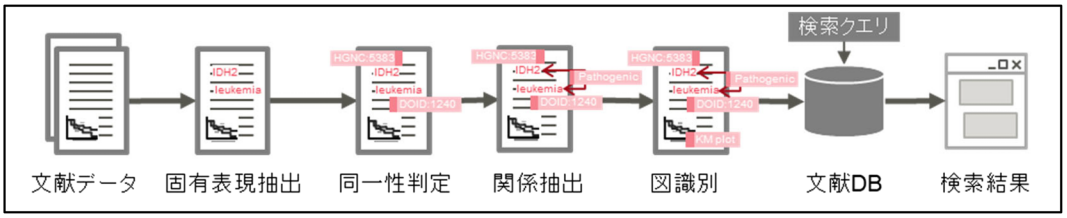

図 2　文献検索 AI の自然言語処理パイプライン概要図

いないが，たとえば「R140Q」や「R172K」などがあれば変異としてハイライトされる．これらの用語の同定は，2.2.1 節で述べる固有表現抽出技術により行われている．

さらに，【2-1】や【2-2】を見てわかるように，入力にマッチした用語「IDH2」と「AML」には二重線（実際には赤色の二重線）が引かれている．後者の正式名称である「acute myeloid leukemia」にも二重線が引かれているが，これは 2.2.2 節で述べる同一性判定技術により「acute myeloid leukemia」が「AML」と同義語であると判定された結果である．

ゲノム医療においては，治療やゲノム情報を評価する際，それらの違いによって患者の状態がどれだけ違うかという予後に関する情報が重要である．そこで，【2-3】の最後の図に示すように，予後に関する図をほかの図とは違い太線（実際は赤色の太線）で囲んで表示することによって，ユーザに一目でわかるようにしている．本機能を実現する技術については 2.2.4 節で説明する．また，「overall survival」などの予後に関する用語のパターンマッチにより本文中の記述を抽出し，【2-4】に示すように表示しユーザが確認できるようにしている．

【2-5】の右上に示すように，日本血液学会が発行する，造血器腫瘍診療ガイドライン [3] と造血器腫瘍ゲノム検査ガイドライン [4] で参照されている論文については，ガイドライン中の該当部分に逆リンクを張っている．たとえば「[白血病-AML-CQ1]」や「[ASXL1-AML]」は実際にはリンクとなっている．これにより，論文がガイドラインで参照されるほど重要であることがわかり，またガイドラインを容易に参照できるようになっている．

2.2　要素技術

図 2 は，システムのパイプラインと利用されている自然言語処理 (Natural Language Processing; NLP) [5] の要素技術の概観図である．文献データに対して，固有表現抽出，同一性判定，関係抽出，図識別の各技術を適用し，解析済みの文献データを文献抽出データベース (DB) へと格納する．文献の検索を行う際には，検索クエリ（問い合わせ式）に対して，解析結果を元に，文書ごとにスコアをつけ，検索結果として出力する．

2.2.1　固有表現抽出技術

文を単語に分割し，エンティティ（用語，固有表現）にあたる単語列を抽出する技術を，固有表現抽出と呼ぶ．処理の初めには，スペースや記号などを手掛かりに，文を区切り，単語に分割する．辞書を利用して固有表現を抽出するため，事前に用意した辞書と完全一致する単語列に対して，あらかじめ辞書と一致したことを表す特徴量を単語に付与し，その後に，単語系列をみて，単語ごとにタグを決定する．

本技術の訓練・評価用のデータとして，MEDLINE の「がん」に関する論文に，エンティティ（遺伝子，変異，疾患，薬剤）とその間の関係をアノテーションしたコーパスを作成した．コーパスは，文献から数文ずつ，遺伝子名や疾患名などを含む文を抽出し，その後人手で見直す手順を経て作成している．

固有表現抽出に用いる辞書として，次の公共 DB，オントロジーを元に辞書を作成した．

- 遺伝子：HGNC Gene name
- 遺伝子変異：HGVS Sequence Variant Nomenclature
- 疾患：Disease ontology
- 薬剤：DrugBank

2.2.2　同一性判定技術

固有表現に対して，辞書を用いて，DB 上の識別子 (ID) を付与する．曖昧性のある語については，辞書に

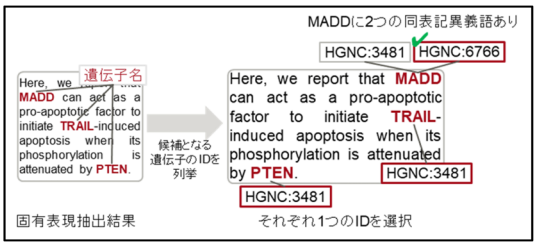

図 3　同一性判定の概要

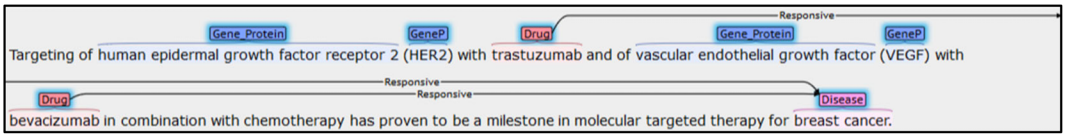

図 4　固有表現抽出，関係抽出結果の例

表 1　コーパスサイズ

	遺伝子	疾患	薬剤	合計
項目数	41,375	11,033	11,882	64,290

表 2　構築辞書サイズ

文数（文）		20,692
抽出元文献数（件）		12,103
エンティティの異なり数	遺伝子（件）	30,364
	遺伝子変異（件）	20,037
	疾患（件）	20,762
	薬剤（件）	18,023

ある候補から文脈上適切な ID を判定し，付与を行う（図 3）．同一性判定に用いる辞書として，固有表現抽出の辞書と同じリソースから，遺伝子名，薬剤名，疾患名について，ID，エンティティタイプ，代表名称，シノニムをもつ辞書を作成した．表 1 に，それぞれのコーパスの項目数を示す．

この辞書には，同形異義のエンティティが存在しており，文脈によって，それらを判別する必要がある．辞書には，2,217 個の一意でない名称があり，それらの ID をアノテーションした，約 3,000 万文のコーパスを，MEDLINE の論文のアブストラクトから辞書を利用して，自動作成した．また，自動作成では十分な事例が作成できなかった 1,690 件のエンティティについて，アノテーションしたコーパスを人手で作成して合わせて利用した．結果として構築された辞書のサイズを表 2 に示す．

2.2.3　関係抽出技術

関係抽出は，文章を主な入力として，文に書かれた固有表現と固有表現の間にある関係を，あらかじめ定義された関係の候補から選択する問題である．入力としては文と共に固有表現の組も与えられ，関係を出力する，多値分類の問題であり，一般的に言われていることと，文に書かれていることが真逆であった場合でも，文に書かれている関係を抽出できているかどうかで正解／不正解が決定される．

抽出した事例を図 4 に示す．ここでは，“Targeting of human epidermal growth factor receptor 2 (HER2) with trastuzumab and of vascular endothelial growth factor (VEGF) with bevacizumab in combination with chemotherapy has proven to be a milestone in molecular targeted therapy for breast cancer.” という文中の Drug（薬剤）のエンティティ trastuzumab が，Disease（疾患）のエンティティ breast cancer に奏効するという関係と，Drug（薬剤）のエンティティ bevacizumab が，同じく breast cancer に奏効するという関係を，文内の単語列から判定している．

固有表現抽出技術の評価で作成した学習・評価データに対して，次の 8 個の関係を定義し，付与した．

1. Gene-Disease Positive
   遺伝子と疾患の間に定義され，遺伝子に変異が起こることにより疾患の原因となる関係．
2. Gene-Drug Target
   遺伝子と薬剤の間に定義され，タンパク質が薬剤のターゲットとなる関係．
3. Variant-Disease Pathogenic
   遺伝子変異と疾患の間に定義され，遺伝子変異が疾患に対する原因となり，疾患を引き起こすと書かれている関係．
4. Variant-Disease Benign
   遺伝子変異と疾患の間に定義され，遺伝子変異が疾患の原因とはならないと書かれている関係．
5. Drug-Disease Responsive
   薬剤と疾患の間に定義され，薬剤が疾患に対して奏効すると書かれている関係．
6. Drug-Disease Insensitive
   薬剤と疾患の間に定義され，薬剤が疾患に対して効果がないと書かれている関係．
7. Variant-Drug Responsive
   遺伝子変異と薬剤の間に定義され，遺伝子変異があることによって，薬効が増大すると書かれている関係．
8. Variant-Drug Resistant
   遺伝子変異と薬剤の間に定義され，遺伝子変異があることによって，薬効が減少すると書かれている関係．

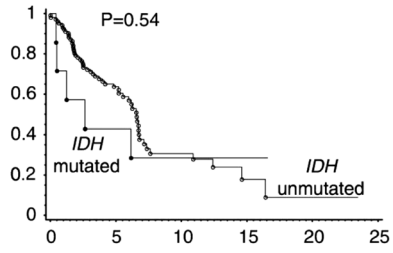

図 5　Kaplan-Meier plot（文献 [6] より引用）

表 3　KM plot 分類コーパス

文献数（論文）		9,600
画像数（件）	KM plot	4,034
	その他	33,231

2.2.4　図識別技術

生命医学論文には細胞画像などの写真や模式図，グラフなどの図が含まれ，論文の内容を具体的にわかりやすく説明する重要なパーツとなっている．その中でも特に，Kaplan-Meier plot (KM plot) は予後を調べるうえで非常に重要な情報源となる．KM plot は特定の疾患に対して遺伝的要因や投与する薬などの条件などで比較し，患者の生存曲線をプロットしたものである．図 5 は一般な KM plot の例で，縦に生存者数（率），横に期間をとり，どの程度の期間生存することができたかを表現している．

大量の論文の中から KM plot を載せている論文だけを選択することや，論文のサマリとして KM plot を表示するには，論文に存在するさまざまな図から KM plot を識別・抽出することが必要となる．そこで，言語処理と画像認識技術を組み合わせ，キャプションと画像を両方利用して図の種類を識別するモデルを構築した．モデルの学習のために作成したデータの詳細を表 3 に示す．

2.3　実証実験

開発した文献検索 AI での効率改善効果を調べるために，実証実験を行った [7]．実験では，文献を検索するための課題を設定して，本システムを用いてユーザが課題解決にかかる時間を測定した．測定時間の範囲は，「被験者の医師に患者の遺伝子変異と癌腫などの周辺情報が提示されてから，文献に書かれた証拠が見つかって診断と治療方針が決まるまで」とした．実験の結果，通常の方法では 30 分程度この検討に時間がかかるところ，このシステムを使うと平均半分以下であるという結果が得られた．これは，推定年間 1 万 2,000 人以上といわれる日本の白血病の罹患者全員を対象にした場合に，医師による検討作業 6,000 時間が 3,000 時間以下に短縮される計算になる．

以上の文献検索技術をベースにして，インターフェースを強化したサービスが，ジー・サーチ社から提供されている [8]．

2.4　文献キュレーションシステム

文献からの自動的な知識抽出は，知識ベースの拡充に役立つが，抽出した知識に誤りや取りこぼしがあると知識ベースの品質に問題が生じる．この観点から，専門家による抽出知識のチェックや修正を行う文献キュレーションが重要となるが，この作業は負荷が大きいという課題がある．本節では，文献検索 AI で使われる自然言語処理技術で抽出された知識に対するキュレーションを，効率的に行えるよう支援する文献キュレーションシステム [9] を紹介する．

文献キュレーションでは，まずキュレーションする文献を検索することになるが，膨大な文献すべてをキュレーションすることは現実的ではなく，特に信頼度の高い文献からキュレーションすることが望ましい．そのため，文献キュレーションシステムでは，文書分類技術を活用し，信頼度の高い文献を選択できるようにしている．文書分類技術とは，文書全体の文を読み込んで，内容に合ったラベルを付与できる技術である．本システムでは，MEDLINE で公開されている論文 2,600 件のアブストラクトに対し，解析手法（例：統計的解析，遺伝子解析）や実験対象種（例：人，マウス・ラット）など，文献の信頼度に関わるラベルを人手で付与し，それを学習データとすることで，自動でラベルを付与できるようにしている．ラベルは文献の信頼度に関わるため，付与されたラベルを元に文献に優先度が付けられる．たとえば実験対象種は「マウス・ラット」よりも「人」の方が信頼度の高い実験となるため，そのような実験について記述される文献を優先する．本システムではラベルごとに信頼度を示す係数を割り振り，文献に付与されるラベルからその文献の信頼度を算出している．

文献キュレーションシステムの文献検索画面では，文献検索システムと同じように疾患名や薬剤名といった固有表現の名前やタイプを指定して検索し，文献を一覧表示する．前述の文書分類によりラベル付けされた結果は，文献ごとに図 6 のようにハイライトした箇所で確認できる．また，信頼度は星印により 5 段階で表現している．

文献検索画面から文献を選択すると文献キュレーション画面に遷移する．この画面は，2.2.3 節の図 4 と同様

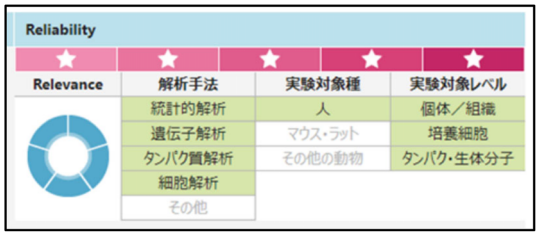

図 6　文献キュレーション 文献検索画面

図 7　文献キュレーション編集画面（文献 [9] より引用）

に，固有表現名をハイライトし，疾患や薬剤といったタイプ名や，それらの関係を矢印で確認できる．キュレーターは誤りや漏れがあれば，固有表現名やタイプ，関係を指定することで編集できる．たとえば，図 7 は固有表現を選択した場合の編集画面である．ラジオボタンで選択されている項目（図 7 では Drug）が固有表現のタイプであるが，誤りがあれば別のタイプを選択して修正できる．

3.　説明可能 AI (Explainable Artificial Intellignce, XAI)

3.1　背景

個々の患者の体質や病状に合わせた治療を行うために，患者の遺伝子に着目したゲノム医療が行われている．ゲノム医療では，患者の組織を採取し，シーケンサーにより遺伝子配列を決定し，遺伝子変異の解釈と調査を経てエキスパートパネルで患者の治療方針を決定する．遺伝子変異の調査では，公共データベースや論文などを調査して該当の遺伝子変異が病原性をもつか（臨床的意義の解釈）などを確認するが，この調査は多大な労力が必要であり，また調査者の熟練度などにより結果が異なる場合があるという問題がある．

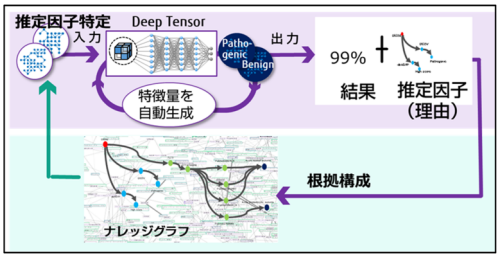

図 8　推定根拠を提示する説明可能 AI

この問題を解決するために，公共データベースや論文などから遺伝子変異の病原性を機械学習で予測する手法が試みられている．しかし，病原性の予測結果のみから，その正誤を人間が判断するのが困難な場合がある．機械学習による予測結果の正誤を人間が判断するためには病原性の予測結果に加え予測結果を導いた理由が必要である．このような機械学習による予測結果の理由を提示する技術は「説明可能 AI (XAI)」と呼ばれ，近年集中的に取り組まれている技術である [10].

3.2　開発技術

われわれは，グラフデータ向けブラックボックス AI の一つである Deep Tensor [11] の予測結果に推定根拠を付与することができる「説明可能 AI」を開発した（図 8）[12]．Deep Tensor は深層学習をベースとした機械学習技術である．一般的な深層学習の特長の一つである特徴量設計（入力のうち予測に有効な部分の特定と組合せ計算）の自動化をグラフデータに対して実現する．Deep Tensor では，グラフデータをテンソル分解という分析により特徴量を取り出しつつテンソルという表現形式に変換し，そのテンソルをニューラルネットワークに入力するという構成をとる．学習時にはテンソルとニューラルネットワークを共に修正しながら学習モデルを構築する．この仕組みにより，学習はテンソルに含まれる特徴量の取り出し方も修正しながら進むため，特徴量設計が自動化できる．

しかし，説明性の観点からすると，後段にはニューラルネットワークを用いており，入力されたもとのグラフのどの部分がその事例における予測に寄与したかを判断することはそのままでは難しい．そこでわれわれは深層学習をはじめとするブラックボックス AI を説明可能にする技術である Local Interpretable Model-Agnostic Explanations (LIME) [13] を応用した，Deep Tensor の予測に寄与したグラフデータの要素（推定因子）を特定する推定因子特定技術を利用し，説明性を高めることに取り組んだ．推定因子特定技術のポイントは，

Deep Tensor のニューラルネットワーク部分を人間にとってわかりやすく説明が容易な線形回帰モデルで近似することにある．線形回帰モデルであれば，テンソルの各要素の寄与度が計算できる．寄与度の高いテンソル要素がわかれば，グラフデータとテンソルの関連度を用いて，元のグラフの要素（ここでは辺）においてもそれ自体の寄与度を算出できる．すなわち，Deep Tensor の予測結果に対する寄与度がグラフの辺ごとに得られる．

しかし，人が膨大な寄与度の付加されたグラフを見てもすぐに理解することはできない．抽出したグラフの意味を説明することが次の課題である．

部分グラフの意味の説明を可能にするために，以下の仕組みを構築した．その前にまず，知識がグラフデータとしてどう表現されているかを具体的に説明しておく．グラフデータは点と辺からなる．点に概念を対応づけ，辺に関係を対応づけることで概念間の関係を表現する．たとえば，ある点に「治療薬 Cytarabine」という概念を対応づけ，別の点に「遺伝子 FLT3」という概念を対応づける．次に，その点「治療薬 Cytarabine」から点「遺伝子 FLT3」に向きのある辺で 2 点をつなぎ，その辺に「作用する」という概念を対応づける．このグラフで，「治療薬 Cytarabine は遺伝子 FLT3 に作用する」という知識が表現できる．このようなグラフで表現された知識の集まりは「ナレッジグラフ」と呼ばれる．

Deep Tensor の予測の際，入力には予測したい遺伝子変異に加えて，全ナレッジグラフ（図 8 下）のうち一部（図 8 左，入力部）を加える．すなわち，予測をするのに重要と思われる知識のみからなる「特徴グラフ」を入力する．推定因子特定技術により，その特徴グラフの各辺に寄与度が付与される．寄与度が高い部分は「推定因子グラフ」（図 8 右上，図 9 左下）として出力される．そのグラフ内の各点について，概念の説明，その概念が関連する知識，およびほかの関連知識を全ナレッジグラフから取り出して，「補完グラフ」として付与する（図 9 右）．こうして，元々の寄与度の高い辺が表す意味を補足できるうえ，関連情報を説明として追加することができる．「推定因子グラフ」に「補完グラフ」がついたものを「根拠説明グラフ」（図 9 全体）と呼ぶ．このように，寄与度のついたグラフから適切な部分を抽出し，理解に必要なグラフとテキスト情報を付与することで，グラフの意味を説明できる．

3.3　遺伝子パネル検査への適用

開発した技術をがんゲノム医療へ適用した事例を紹介する．患者の遺伝子パネル検査を行うといくつかの変異が報告される．すでに病原性が確認されている変異に対しては治療方法を検討できるが，そうでない変異は困難である．その場合に，説明可能 AI を用いて病原性を予測すれば，あわせて病原性の根拠を確認できて，治療法の検討につなげられる．

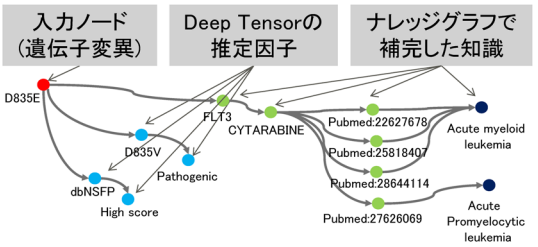

図 9　根拠を説明するグラフ

ゲノム医療データはオープン化 [14] により，公共データベースとして，日々蓄積，公開されている．ナレッジグラフを作成するために，われわれは MED2RDF [15] というプロジェクトを京都大学，ライフサイエンス統合データベースセンター (Database Center for Life Science; DBCLS) と共同で立ち上げた．MED2RDF プロジェクトでは公共データベースを RDF に変換するツールを公開している．本研究では，そのツールを用いて，複数のデータベース（実際の病原性が変異と共に登録されている ClinVar，さまざまな既存アルゴリズムによる変異の病原性予測結果が登録されている dbNSFP，治療薬の情報が登録されている DGIdb）からナレッジグラフを作成した．これは図 8 下の全ナレッジグラフに相当する．

このグラフを学習する際，正解データが書かれた ClinVar には信頼性の低い情報も載っているため，メタ情報を使ってそれらは利用対象から排除した．また，学習や予測の手がかりとなる「特徴グラフ」としては，既存アルゴリズムによる病原性予測結果と，該当変異と関連が深いほかの変異の有無，および，その病原性判定結果などを使用した．

学習したモデルがテストデータに対して行った予測と説明の例を紹介する（図 9）．遺伝子 FLT3 の変異 D835E を予測させた場合，病原性が高いと予測された．その根拠（推定因子グラフ）をみると，dbNSFP では病原性ありの予測があり，また周辺変異としては D835V が ClinVar 情報で病原性変異という情報が出力されており，妥当な説明が提示されることが確認できた．また，補完グラフとしては，DGIdb の登録情報から治療薬 Cytarabine と関連文献情報，および，対象

となる疾患名「Acute myeloid leukemia」と「Acute promyelocytic leukemia」が出力された．根拠として出力された文献を読むことで予測の正しさを検討できる．すなわち AI の予測結果をやみくもに信用するのではなく，根拠を確認しながら，治療を検討することが可能となる．

遺伝子パネル検査において，病原性が確認されていない大量の変異を人が調査・判断するのは大変な作業となる．今回開発した技術は，病原性の高さを表すスコアでより検討すべき重要な変異に絞れるなど，調査を効率化できる．

今後は，予測精度の向上，根拠の妥当性の検証を行い，本手法の実用化を目指したい．

4. おわりに

本稿では，ゲノム医療を推進するために開発した二つの AI を紹介した．

一つは，自然言語処理による知識抽出を行い，遺伝子変異に結びつけて薬剤の奏功情報を効率的に探せる文献検索 AI である [7]．新規の論文には辞書にない新規の薬剤や遺伝子変異が出現するが，本システムでは自然言語処理技術によって文脈を判定することで同定できる．しかし，精度は完全ではないため，現場で満足される程度にまで改良することが課題である．

二つ目の AI は，文献からの知識ベースをもとに，まだ遺伝子変異に対する薬剤が報告されていない変異について，遺伝子変異ががんの原因か（病原性の程度）を予測する AI である [12]．この AI においては，予測の理由と根拠を出力する「説明可能な AI (XAI)」を構築した．

謝辞　文献検索 AI の課題設定と実証実験においては，東京大学医科学研究所 井元清哉教授，横山和明助教に貴重なアドバイスをいただき，また同附属病院の 4 人の医師の方々に実証実験に参加していただいた．

本研究の一部は AMED の課題番号 JP20kk0205013 の支援を受けた．

参考文献

[1] U.S. National Library of Medicine, MEDLINE, https://www.nlm.nih.gov/medline/index.html（2021 年 4 月 5 日閲覧）
[2] U.S. National Library of Medicine, PMC, https://www.ncbi.nlm.nih.gov/pmc/（2021 年 4 月 2 日閲覧）
[3] 日本血液学会，「造血器腫瘍診療ガイドライン」，http://www.jshem.or.jp/gui-hemali/table.html（2021 年 4 月 2 日閲覧）
[4] 日本血液学会，「造血器腫瘍ゲノム検査ガイドライン」，http://www.jshem.or.jp/genomgl/（2021 年 4 月 2 日閲覧）
[5] 黒橋禎夫，『自然言語処理〔改訂版〕（放送大学教材）』，放送大学教育振興会，2019.
[6] A. Tefferi, T. L. Lasho, O. Abdel-Wahab, P. Guglielmelli, J. Patel, D. Caramazza, L. Pieri, C. M. Finke, O. Kilpivaara, M. Wadleigh, M. Mai, R. F. McClure, D. G. Gilliland, R. L. Levine, A. Pardanani and A. M. Vannucchi, "IDH1 and IDH2 mutation studies in 1473 patients with chronic-, fibrotic- or blast-phase essential thrombocythemia, polycythemia vera or myelofibrosis," *Leukemia*, **24**, pp. 1302–1309, 2010.
[7] 富士通，「新しい AI によるがんゲノム医療の効率化を東大医科研との共同研究で実現」，https://pr.fujitsu.com/jp/news/2019/11/6.html（2021 年 3 月 25 日閲覧）
[8] ジー・サーチ，「AI を活用した論文調査サービス「JDream SR」の提供開始」，https://www.g-search.jp/release/2020-10-14-000693.html（2021 年 3 月 25 日閲覧）
[9] 小林賢司，阿部修也，森田一，吉川和，岡嶋成司，富士秀，"エビデンスに基づく医療のための文献キュレーションシステム"，研究報告グループウェアとネットワークサービス (GN), 2019-GN-107, pp. 1–8, 2019.
[10] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti and D. Pedreschi, "A survey of methods for explaining black box models," *ACM Computing Surveys*, **51**, Article 93, pp. 1–42, 2018. doi:10.1145/3236009.
[11] K. Maruhashi, M. Todoriki, T. Ohwa, K. Goto, Y. Hasegawa, H. Inakoshi and H. Anai, "Learning multi-way relations via tensor decomposition with neural networks," In *Proceedings of Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.
[12] 富士通，「AI の推定理由や根拠を説明する技術を開発」https://pr.fujitsu.com/jp/news/2017/09/20-1.html（2021 年 3 月 25 日閲覧）
[13] M. T. Ribeiro, S. Singh and C. Guestrin, "Why should I trust you?: Explaining the predictions of any classifier," In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1135–1144, 2016. doi:10.1145/2939672.2939778.
[14] Linked Open Vocabularies, Linked Open Vocabularies, https://lov.linkeddata.es/dataset/lov/（2021 年 3 月 30 日閲覧）
[15] MED2RDF, MED2RDF, http://med2rdf.org/（2021 年 3 月 30 日閲覧）