社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

生物学的パスウェイを用いたBioConceptVecにおけるアナロジータスク

山際　宏明† 　橋本　竜馬† 　荒金　　究†† 　村上　　賢†† 　大山百々勢†

下平　英寿†,††† 岡田眞里子††

† 京都大学 〒606-8501 京都府京都市左京区吉田本町
†† 大阪大学蛋白質研究所 〒565-0871 大阪府吹田市山田丘 3-2
††† 理化学研究所 〒351-0198 埼玉県和光市広沢 2-1
E-mail: †{hiroaki.yamagiwa,hashimoto.ryoma,oyama.momose}@sys.i.kyoto-u.ac.jp,
††{k.arakane,k-mrkm,mokada}@protein.osaka-u.ac.jp, †††shimo@i.kyoto-u.ac.jp

あらまし　自然言語処理は様々な応用分野で利用されており，skip-gram などのモデルを用いてテキスト中の単語を埋め込みと呼ばれる特徴ベクトルに変換することが一般的である．近年，生物学の分野でも自然言語処理の有用性が注目されており，概念の正規化を施した約 3000 万件の PubMed abstracts から学習した BioConceptVec が提案されている．一般に skip-gram では単語の埋め込みを加減算することによりアナロジータスクが解けるとされており，例えば king − man + woman から queen を予測できる．本研究では生物学的パスウェイの種類を関係性とみなし，薬剤と遺伝子の組についてアナロジータスクの実験を行った．その結果，同じパスウェイに属する薬剤と遺伝子の組についてパスウェイの関係性を表すベクトルを定義することで，アナロジータスクの高い精度が確認された．

キーワード　自然言語処理，分散表現，単語埋め込み，アナロジー，生物学，PubMed

Analogy Tasks in BioConceptVec using Biological Pathways

Hiroaki YAMAGIWA†, Ryoma HASHIMOTO†, Kiwamu ARAKANE††, Ken MURAKAMI††, Momose OYAMA†, Hidetoshi SHIMODAIRA†,†††, and Mariko OKADA††

† Kyoto University, Yoshidahonmachi, Sakyo-ku, Kyoto-shi, Kyoto, 606-8501, Japan
†† Institute for Protein Research, Osaka University, 3-2 Yamadaoka, Suita-shi, Osaka, 565-0871, Japan
††† RIKEN, 2-1 Hirozawa, Wako-shi, Saitama, 351-0198, Japan
E-mail: †{hiroaki.yamagiwa,hashimoto.ryoma,oyama.momose}@sys.i.kyoto-u.ac.jp,
††{k.arakane,k-mrkm,mokada}@protein.osaka-u.ac.jp, †††shimo@i.kyoto-u.ac.jp

Abstract　Natural language processing (NLP), often employing models like skip-gram, is widely utilized across numerous application domains to convert words in text into feature vectors known as word embeddings. The utility of this approach has recently been noted in the field of biology, with the introduction of BioConceptVec, a model trained on about 30 million PubMed abstracts using normalized concepts. In general, skip-gram can solve analogy tasks by manipulating word embeddings, such as predicting *queen* from *king* − *man* + *woman*. In this study, we applied this principle to biological pathways, conducting analogy tasks for pairs of drugs and genes, treating pathway types as relationships. Our results demonstrated high accuracy in these tasks when defining a vector to represent the pathway relationship for pairs of drugs and genes that belong to the same pathway.

Key words　natural language processing, distributed representations, word embeddings, analogy, Biology, PubMed

1. はじめに

自然言語処理とは，人間が日常的に使用する自然言語をコンピュータで処理するための技術である．自然言語処理は，機械翻訳 [1]，感情分析 [2]，文類似度の測定 [3] など，様々な応用分野で活用されているが，それらの多くで skip-gram [4], [5] や BERT [2] などのモデルを用いてテキスト中の単語を分散表現または埋め込みと呼ばれる数百次元の特徴量ベクトルに変換

する．ここでは単語 word の分散表現をベクトル v_{word} で表す．skip-gram は着目している単語の周辺の単語を予測することで，効率的に性能の高い分散表現を学習する手法であり，skip-gram で学習された分散表現では，$v_{\text{king}} - v_{\text{man}} + v_{\text{woman}}$ が v_{queen} に近いという性質が成り立つことが知られている．埋め込み空間において差分ベクトル $v_{\text{king}} - v_{\text{man}}$ と $v_{\text{queen}} - v_{\text{woman}}$ は *royalty* を意味する方向ベクトルであるとみなせるが，skip-gram の学習時にそのような関係性は教えていないため，skip-gram はこれらの性質を自然と獲得している．「man が king に対応する時，woman に対応するのは何か？」のような形式の問いを解くためには，単語間の関係性を理解する能力が必要となる．これはアナロジータスクと呼ばれる問題形式で，モデルの言語理解能力や推論力を評価する際に使用される．例えばこの場合では，man から king への *royalty* という関係性を理解し，これに類似する関係性を woman に適用することが求められる．skip-gram がベクトルの加減算でアナロジータスクをうまく解くことができる理由についても研究がなされている [6].

近年，生物学の分野でも自然言語処理を用いた手法が注目されている [7]〜[9]．しかし，通常の skip-gram では web 上のテキストデータなどで学習されているため，生物学分野の文中に出現する専門用語をうまく扱うことができない．とくに，複数の単語が同じ概念を表す場合は，あらかじめ概念の正規化処理を行って同じ分散表現を使用したほうがよい．Chen ら [10] は，概念を正規化した約 3000 万件の PubMed abstracts を学習データとして単語の分散表現を計算した．Chen らはこの正規化処理に PubTator [11] を用いた．PubTator は，生物医学文献の自動アノテーションを支援するオンラインツールであり，大量の科学的文献から特定の情報を効率的に抽出するための補助を目的としている．具体的には，特定の概念やエンティティを識別し，適切なカテゴリ（例えば，疾患，遺伝子，薬剤など）に分類することができる．そこで Chen らは PubTator の正規化を適用し，skip-gram などの 4 つのモデルで学習した BioConceptVec を提案した．

Chen ら [10] は BioConceptVec の有効性を内部評価と外部評価の 2 つの手法によって評価した．内部評価では薬剤と遺伝子（drug-gene）の相互作用と遺伝子間（gene-gene）の関連概念を特定するために，分散表現の cosine 類似度を使用して関連する遺伝子の集合の類似度を計算している．外部評価ではタンパク質間（protein-protein）の相互作用予測および薬剤間（drug-drug）の相互作用抽出のタスクにおいて，分散表現を特徴量としてニューラルネットワークの分類器を使用してクラス分類を行っている．しかし単語間の関係性を調べるアナロジータスクの性能評価は行われていない．

そこで，我々は BioConceptVec を用いて生物学におけるアナロジータスクの性能評価を行った．薬剤と遺伝子 (drug-gene) の関係性を推測し，薬剤からそのターゲットとなる遺伝子を予測するアナロジータスクである．我々は，薬剤と遺伝子の関係性はパスウェイ (pathway) の種類で定まるものとした．同じ pathway に属する薬剤と遺伝子 (drug-gene) からその関係性の決め方については，セクション 3. で説明する．ここで注意すべきは，$v_{\text{king}} - v_{\text{man}} + v_{\text{woman}} \approx v_{\text{queen}}$ の例では $v_{\text{king}} - v_{\text{man}}$ と $v_{\text{queen}} - v_{\text{woman}}$ の 2 つの差分ベクトルが *royalty* の関係性を表すのに対し，pathway 内に属する drug と gene は多くの場合 2 つ以上存在し，さらに 1 つの drug に対し複数の gene がターゲットとなる点である．ここで drug のターゲットとは，薬剤が作用する gene のことを指す．そのため我々は個別の drug と gene の組を比較するのではなく，同じ pathway に属する drug-gene の組でターゲットの関係をもつすべての組について同じ方向ベクトルが定まると考えた．そこで，同じ pathway に属する gene の平均ベクトルと drug の平均ベクトルを計算し，その差分を pathway を意味する方向ベクトルとしたアナロジータスクを考える．すなわち，pathway の集合を P とすると，pathway $p \in P$ ごとに drug の集合 D_p，gene の集合 G_p を定め，pathway p の関係性を表すベクトル v_p を $v_p := \bar{v}_{G_p} - \bar{v}_{D_p}$ で定義する，ここで集合 X について，$\bar{v}_X$ は X に属する単語の分散表現の平均である．

本研究の目的は，BioConceptVec を用いて，pathway の種類を関係性とみなし，drug と gene の組についてアナロジータスクを解くことである．そこで pathway 内のアナロジータスク (intra-analogy) と pathway 間でのアナロジータスク (inter-analogy) の 2 つの設定で実験を行い性能を比較した．ベースラインとしては，ランダムに gene を選択する場合と drug 全体の平均ベクトルと gene 全体の平均ベクトルの差分ベクトルを用いる場合を考慮した．実験では，BioConceptVec の skip-gram モデルと，BioConceptVec と同様の学習データを用いて我々が学習させた skip-gram モデルを用いた．結果として，pathway 内でのアナロジータスクにおいて，両方のベースラインよりも高い精度を達成した．また，pathway 間でのアナロジータスクにおいても，pathway p に属する drug へ，他の pathway p' から求めた $v_{p'}$ を足すことで，drug とターゲットの関係を持つ gene をある程度の精度で探すことができると示された．

本稿の構成は次のとおりである．まず関連研究について説明する．次に，本研究で使用する手法について詳細を説明し，intra-analogy および inter-analogy のタスクについて実験を行う．それぞれのタスクについて，cosine 類似度を用いた top k-accuracy で性能を評価する．その後考察と今後の展望について述べ，最後にまとめを行う．本研究の貢献は，pathway で表される関係性を考慮することで，通常の skip-gram と同様に BioConceptVec でもアナロジータスクを解くことができると示されたことである．

2. 関 連 研 究

特定の自然科学の分野に特化した単語ベクトルを学習し，その加法構成性を利用することで関係推論を試みる手法が存在する．Tshitoyan ら [12] はマテリアルサイエンス分野の膨大な文献の概要から単語の分散表現を計算し，「強磁性 – NiFe + IrMn ≈ 反強磁性」といった関係の類推を行っており，単語ベクトルの加法構成性の有効性を示している．

Word2Vec [4] 登場以前にもテキストマイニングによって生物医学文献から特定の知識を抽出する研究は行われている．

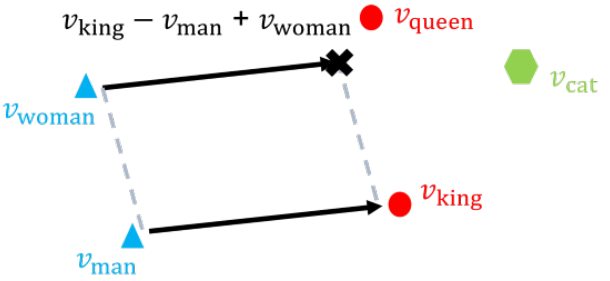


図 1: ベクトルの加減算を用いた通常のアナロジータスクで，「man が king に対応する時，woman に対応するのは何か？」を考える場合．

Müller ら [13] は特定の知識領域の概念とその関連性を定義した「オントロジー辞書」を用意することで，オンライン文献検索・キュレーションシステムを作成している．Friedman ら [14] は構文解析によって生物学的文献から細胞経路に関する情報を抽出・構造化するシステムを作成している．

タンパク質や薬剤などの専門的な概念に関する予測や関係性の抽出において，生物医学分野の文献から単語ベクトルを計算し活用する研究が存在する．Shtar ら [15] は新薬に対する drug-drug interactions の予測に対して，既存の薬剤の drug-drug の関係に加えて生物医学文献で学習した単語ベクトルを特徴量として与えている．Alachram ら [16] は単語ベクトルの cosine 類似度を利用して gene-gene ネットワークを作り，その特徴量を利用して Graph-CNN を学習することで，転移イベント予測タスクにおける性能が十分であることを示している．生物医学文献では，同一の概念であっても文献によって異なる用語を用いていることが多い．生物医学文献における同義語を検出する目的で単語ベクトルの cosine 類似度が頻繁に使用されている．Yeganova ら [17] は単語ベクトルの cosine 類似度による同義語検出能力を評価している．

単語のベクトル情報を明示的に使用しない場合でも，テキストから drug-drug や protein-protein などの生物医学的な概念の関係を抽出するためにニューラルネットワークを用いる研究は行われている [18]〜[22]．

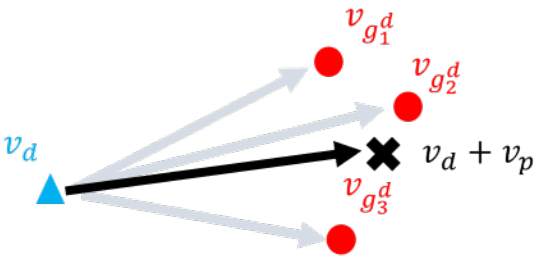
自然言語処理タスクにおいて優れた性能を達成している BERT [2] のような事前学習済みの言語モデルを生物医学のドメインに適応する研究も行われている [7], [23], [24]．

単語埋め込み手法は自然言語を対象にするものばかりではない．Du [25] らはトランスクリプトーム全体の遺伝子共発現を利用して遺伝子をベクトルで表現する手法を提案している．

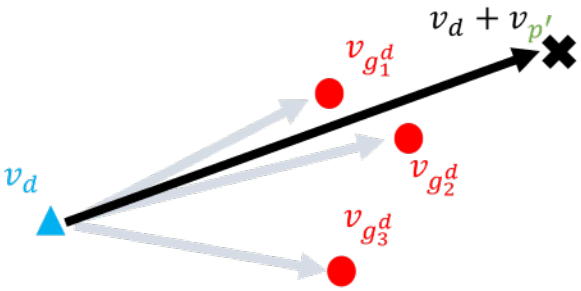
3. 手　　法

3.1 アナロジータスク

アナロジータスクの性能評価法について説明する．アナロジータスクでベクトル間の類似度を比較する際，cosine 類似度を用いる．これは分散表現の類似度比較で単純な Euclid 距離などよりも cosine 類似度が良い性能を示すことが知られているためである．ベクトルの加減算を用いた通常のアナロジータスクで



(a) intra-analogy



(b) inter-analogy

図 2: pathway p に関する drug $d \in D_p$ について，ターゲットの gene として $g_1^d, g_2^d, g_3^d \in G_p$ が定まっている状況を考える．(a) intra-analogy では，pathway p から計算した v_p を用いて，drug d の分散表現 v_d に v_p を足すことでアナロジータスクを解くため，正解となる $v_{g_i^d}$ を正しく選ぶことは比較的容易と予想される．(b) inter-analogy では，異なる pathway p' から計算した $v_{p'}$ を用いて，drug d の分散表現 v_d に $v_{p'}$ を足すことでアナロジータスクを解くため，intra-analogy と比べて，より難しい設定である．

「man が king に対応する時，woman に対応するのは何か？」を考える場合，$v_{\text{king}} - v_{\text{man}} + v_{\text{woman}}$ と語彙中の単語の分散表現との cosine 類似度を測り，v_{queen} が $v_{\text{king}} - v_{\text{man}} + v_{\text{woman}}$ と一番近い単語であれば，正しくアナロジータスクを解けている．正解となる単語が 1 番目ではなく k 番目までに現れるときの top k-accuracy などを用いて性能を評価する．図 1 に $v_{\text{king}} - v_{\text{man}} + v_{\text{woman}}$ と $v_{\text{queen}}, v_{\text{cat}}$ との比較の様子を示す．$v_{\text{king}} - v_{\text{man}}$ が *royalty* を意味する方向ベクトルであるとみなし，v_{woman} に $v_{\text{king}} - v_{\text{man}}$ を足すことで，$v_{\text{king}}, v_{\text{man}}, v_{\text{woman}}, v_{\text{king}} - v_{\text{man}} + v_{\text{woman}}$ を頂点にもつ平行四辺形ができる．ベクトル v_1 と v_2 の cosine 類似度を $\cos(v_1, v_2) = (v_1 \cdot v_2)/(\|v_1\| \|v_2\|)$ と書く．もし $\cos(v_{\text{king}} - v_{\text{man}} + v_{\text{woman}}, v_{\text{queen}}) > \cos(v_{\text{king}} - v_{\text{man}} + v_{\text{woman}}, v_{\text{cat}})$ であれば，$v_{\text{king}} - v_{\text{man}} + v_{\text{woman}}$ は v_{cat} よりも v_{queen} に近いとみなす．同様に他の単語の分散表現についても cosine 類似度を比べ，v_{queen} が $v_{\text{king}} - v_{\text{man}} + v_{\text{woman}}$ に何番目に近いかを調べる．

次に一般の skip-gram におけるアナロジータスクをもとに，BioConceptVec で drug と gene のアナロジータスクを考える．

そこで我々は pathway 内のアナロジータスク (intra-analogy) と pathway 間でのアナロジータスク (inter-analogy) の 2 つの設定で実験を行う．そのため pathway p について，drug の集合 D_p に属する drug $d \in D_p$ を考える．drug d が gene の集合 G_p のうち，n_d 個の gene とターゲットの関係を持つとき，これらの gene を $\{g_i^d\}_{i=1}^{n_d} \subset G_p$ と表記する．セクション 1. で述べたように，pathway p の関係性を表すベクトル v_p を $v_p := \bar{v}_{G_p} - \bar{v}_{D_p}$ と定義する，ここで集合 X について，$\bar{v}_X$ は X に属する単語の分散表現の平均 $\bar{v}_X = \sum_{x \in X} v_x / |X|$ である．v_p の定義から pathway 集合 P について，v_p の集合 $\{v_p\}_{p \in P}$ が求まる．

設定 1. intra-analogy では，pathway p から求まる v_p を用いて，$v_d + v_p$ と全ての gene の分散表現との cosine 類似度を測り，最も近い gene が $\{g_i^d\}_{i=1}^{n_d}$ に含まれれば，正しくアナロジータスクを解けているとみなす．ベクトルの加減算を用いた通常のアナロジータスクと同様に top k-accuracy などを用いて性能を評価する．図 2a に intra-analogy タスクの様子を示す．1 つの drug について，複数の gene がターゲットの関係を持つため，ベクトルの加減算を用いた通常のアナロジータスクとは異なり，集合に含まれるかどうかで正誤を判断している．また v_p は pathway p によって定まるので，図 1 で見られるような平行四辺形は図 2a には現れない．intra-analogy の accuracy が高い場合，ある pathway に属する drug の分散表現 v_d に v_p を足すことで，その pathway に属し drug とターゲットの関係を持つ gene を正しく探すことができる．

設定 2. inter-analogy では，異なる pathway p' から求まる $v_{p'}$ を用いて，pathway p の drug $d \in D_p$，gene $g_i^d \in G_p$ について intra-analogy と同様のアナロジータスクを考える．intra-analogy と比較し，v_p の代わりに $v_{p'}$ を用いる点が異なる．$v_d + v_{p'}$ と全ての gene の分散表現との cosine 類似度を測り，最も近い gene が $\{g_i^d\}_{i=1}^{n_d}$ に含まれれば，正しくアナロジータスクを解けているとみなす．intra-analogy と同様に top k-accuracy などを用いて性能を評価する．図 2b に inter-analogy タスクの様子を示す．v_p と $v_{p'}$ が異なる以外は intra-analogy と同じ実験設定である．pathway p と p' について inter-analogy の accuracy が高い場合，ある pathway に属する drug の分散表現 v_d に他の pathway p' から求めた $v_{p'}$ を足すことで，drug d とターゲットの関係を持つ，pathway p に属する gene を正しく探すことができるため，pathway p と p' は似た性質を持つと予測される．

中心化. アナロジータスクの性能評価をする際，gene 全体の集合を G として，あらかじめ gene 全体の分散表現の平均 $\bar{v}_G$ を計算しておき，cosine 類似度を測る際に，$\bar{v}_G$ で中心化した場合 (centering) も考える．設定 1 と設定 2 の両方について，中心化無しの場合はベクトル v_1 と v_2 の consine 類似度 $\cos(v_1, v_2)$ を用いるが，中心化有りの場合は consine 類似度 $\cos(v_1 - \bar{v}_G, v_2 - \bar{v}_G)$ を用いる．中心化する理由は，cosine 類似度は原点の位置の影響を受けるため，そのような影響を排除する狙いがある．

3.2 分散表現の設定

BioConceptVec では, CBOW [4], skip-gram [4], [5], Glove [26], fastText [27] の 4 つのモデルが公開されている[1]．我々は実験に用いる分散表現として BioConceptVec の skip-gram モデルを用いた．これは CBOW は skip-gram よりも単純なモデルであり，Glove よりも skip-gram でアナロジーの性質がうまく成り立ち [6]，fastText は $n$gram を用いた skip-gram であり実質的に skip-gram とみなせる事に由来する．BioConcepVec の skip-gram の次元は 100 であった．さらに BioConceptVec の論文中の手順に従い，公開されている PubMed abstracts データセットをもとに，我々が学習させた 100 次元の skip-gram モデルでも性能を比較する．これは公開されている BioConceptVec と我々が学習させた分散表現で大きな差がないことを確認するためである．具体的には，PubMed abstracts をコーパスとして学習に使用した．さらに学習に用いる文に対し，PubTator [11] を用いて 6 つの主要な生物学的概念 (genes, mutations, diseases, chemicals, cell lines, species) を指す単語の表記ゆれを統一し，NLTK [28] を用いてトークナイズを行った．学習時のハイパーパラメータを表 1 に示す．また，BioConceptVec と我々が学習させた分散表現の語彙に含まれる drug と gene の数を表 2 に示す．表 2 から，それぞれの skip-gram の語彙に含まれる drug と gene の数は異なることが分かる．

Hyperparameter	Values
training epochs	10
down-sampling threshold	10^{-5}
learning rate	0.025
window size	5
negative samples	5
minimal word occurrence	30
vector dimension	100

表 1: BioConceptVec をもとに skip-gram モデルを学習させたときのハイパーパラメータ.

	skip-gram	
Vocab Stats	BioConceptVec	Our calculations
Drugs	117282	23232
Genes	144584	45193

表 2: BioConceptVec と我々が学習させた分散表現の語彙中に含まれる drug と gene の数.

3.3 データセット

我々は実験に用いるデータベースとして KEGG [29] を用いた．KEGG (Kyoto Encyclopedia of Genes and Genomes) は，ゲノム，化学反応，pathway などの生命科学の広範な情報を総合的に提供するデータベースシステムである．我々は実験に用いる pathway として，人間に関する pathway のリストを KEGG API から取得した[2]．取得した pathway ごとに，再び KEGG API から drug と gene の集合を定める[3]．次に KEGG データベー

(注1)：https://github.com/ncbi/BioConceptVec
(注2)：https://rest.kegg.jp/list/pathway/hsa
(注3)：例えば ErbB signaling pathway の場合，ID は hsa04012 であるから https://rest.kegg.jp/get/hsa04012 のようにして取得する．

	skip-gram	
Pathway Stats	BioConceptVec	Our calculations
Pathways	189	154
Total Drugs	4500 (2126)	2273 (1388)
Total Genes	1492 (583)	1136 (532)
Total Targets	11715 (5865)	5402 (3617)
Avg Drugs/pathway	23.81	14.76
Avg Genes/pathway	7.89	7.38
Avg Targets/Drug	2.60	2.38

表 3: BioConceptVec の skip-gram と我々が学習させた skip-gram の語彙中に含まれる pathway の数，全体の pathway に含まれる drug, gene, target の数，pathway に含まれる平均の drug 数と gene 数，1 つの drug に対する平均のターゲット gene 数．全体の pathway に含まれる drug, gene, target の数について，括弧内は重複を取り除いた場合の数．

ス [30] から取得した ASURAT [31] の公開データ[4]から drug とそのターゲットである gene の関係性を定義する．KEGG から取得した drug データは KEGG ID を用いて表記されるため，biothings_client[5]によって Biothings API[6]を用いて，KEGG ID を MeSH [7] ID に変換する．これは PubTator で drug の表記の正規化に MeSH ID が用いられているため，BioConceptVec でも drug は MeSH ID を用いて語彙に登録されているからである．そのため，MeSH ID に変換できた drug のみを扱う．BioConceptVec の skip-gram と我々が学習させた skip-gram について，pathway の数，pathway に含まれる drug, gene, target の数，pathway に含まれる平均の drug 数と gene 数，1 つの drug に対する平均のターゲット gene 数を表 3 に示す．表 2 で見たように，それぞれの skip-gram の語彙に含まれる drug と gene の数が異なるため，表 3 では同じ手順を経た場合でも有効な pathway の数が異なり，それに伴い他の数値も異なる．また表 3 から分かるように，pathway に含まれる drug, gene, target の数は重複するため，同じ drug や gene が異なる pathway に現れる場合がある．

3.4 ベースライン

intra-analogy, inter-analogy の性能を測るために我々は 2 つのベースラインを考えた．

1 つ目のベースラインとして，全体の gene から候補となる gene をランダムに選ぶ場合を考える．この手法を random baseline と呼び，random baseline はアナロジータスクで当てずっぽうに gene を予測した場合に対応する．選択する gene のランダム性の影響を考慮して，実験を 100 回繰り返し，その平均から性能を評価する．

2 つ目のベースラインは drug 全体の集合を D，gene 全体の集合を G として，v_p の代わりに $v_{\text{shift}} := \bar{v}_G - \bar{v}_D$ を用いてアナロジータスクを解く場合を考える．この手法を shift baseline と呼び，v_{shift} は drug 全体の平均ベクトルと gene 全体の平均ベクトルの差であるから，全ての drug と gene が同じ pathway に属する状況とみなすことができ，drug から gene への大まかな方向を表すベクトルとなっていると分かる．このような性質から v_{shift} を用いた shift baseline は random baseline よりも良いベースラインであることが予測される．なお定義から intra-analogy, inter-analogy において，shift baseline のスコアは一致する．

intra-analogy, inter-analogy で用いる $v_p, v_{p'}$ による加算の方法を random baseline, shift baseline と対比し，pathway shift と呼ぶことにする. とくに，v_p による移動を p-shift，$v_{p'}$ による移動を p'-shift と呼ぶ.

4. 結　　果

BioConceptVec の skip-gram と我々が学習させた skip-gram について，intra-analogy, inter-analogy の設定で random baseline, shift baseline, pathway shift の手法で top1-accuracy, top5-accuracy, top10-accuracy のスコアを計算した結果を表 4 に示す．ただし shift baseline と pathway shift については中心化した場合のスコアも計算した．

random baseline では全ての場合で精度が低く，ほとんど 0 に近い数値を示しており，全くランダムに gene を予測することは難しい状況であると分かる．

shift baseline では inter-analogy, intra-analogy の両方で random baseline より良い結果となっている．中心化したときの shift baseline は，中心化をしない場合の shift baseline と比べ，BioConceptVec の skip-gram では top1-accuracy を除き性能がわずかに向上している一方，我々が学習させた skip-gram の場合，むしろ intra-analogy, inter-analogy の両方で精度が低下した．

pathway shift では両方の skip-gram で，ベースラインよりも十分高い精度を示している．中心化したときの pathway shift では，中心化をしない場合の pathway shift と比べ，BioConceptVec の skip-gram ではほとんどの場合でわずかな性能の向上がみられ一番良い結果となった．一方で shift baseline の時と同様に我々が学習させた skip-gram の場合，intra-analogy, inter-analogy の両方で精度が低下した．特に両方の skip-gram で pathway shift の intra-analogy の top1-accuracy のスコアが 64% 程度となっており，表 2 で見たように BioConceptVec の skip-gram の gene の数が約 14 万，我々が学習させた skip-gram の gene の数が約 5 万であることを考えると十分良いスコアであると考えられる．

5. 考　　察

セクション 4. で見たように，pathway p の drug と gene の分散表現の平均ベクトルから定義した v_p を用いて intra-analogy, inter-analogy の設定でアナロジータスクを解くことができた．そこで pathway ごとの drug と gene の分散表現の平均ベクトル間の関係を見るために，BioConceptVec の skip-gram について，これらの平均ベクトルについて PCA による変換を考え，第一主成分，第二主成分をプロットした様子を図 3 に示す．drug と gene が大まかにまとまって分布し，drug から gene への方向ベクトルが定まる様子が分かる．このことから shift baseline でも

(注4)：https://github.com/keita-iida/ASURATDB/blob/main/genes2bioterm/20221102_human_KEGG_drug.rda
(注5)：https://github.com/biothings/biothings_client.py
(注6)：https://biothings.io/
(注7)：https://www.nlm.nih.gov/mesh/meshhome.html

skip-gram	Method	Centering	intra-analogy			inter-analogy		
			top1-acc	top5-acc	top10-acc	top1-acc	top5-acc	top10-acc
BioConceptVec	random baseline		0.0000	0.0001	0.0001	0.0000	0.0000	0.0000
	shift baseline		0.0882	0.2039	0.3159	0.0882	0.2039	0.3159
		✓	0.0865	0.2346	0.3339	0.0865	0.2346	0.3339
	pathway shift		0.6427	**0.8117**	0.8488	0.1819	0.3851	0.4773
		✓	**0.6511**	0.8082	**0.8536**	**0.1869**	**0.3927**	**0.4847**
Our calculations	random baseline		0.0000	0.0001	0.0003	0.0000	0.0002	0.0005
	shift baseline		0.1396	0.3515	0.4467	0.1396	0.3515	0.4467
		✓	0.1035	0.2200	0.2674	0.1035	0.2200	0.2674
	pathway shift		**0.6484**	**0.8152**	**0.8525**	**0.1858**	**0.3800**	**0.4604**
		✓	0.5372	0.7120	0.7533	0.1346	0.2660	0.3241

表 4: BioConceptVec の skip-gram と我々が学習させた skip-gram について，intra-analogy, inter-analogy の設定で random baseline, shift baseline, pathway shift の手法で top1-accuracy, top5-accuracy, top10-accuracy のスコアを計算した結果.

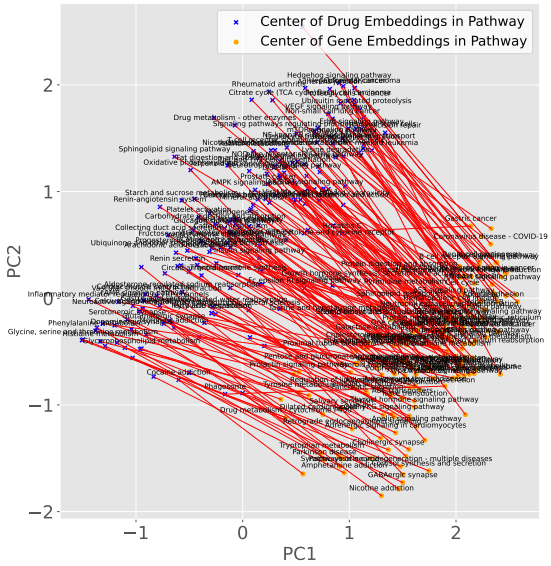


図 3: pathway ごとの drug と gene の分散表現の平均ベクトルについて PCA による変換を考え，第一主成分，第二主成分をプロットした図．対応する drug と gene を線分で結ぶと，線分 1 つが 1 つの pathway に対応し，pathway ごとの大まかな位置関係を把握することができる．

Fig. 4	pathway	Drugs	Genes	Targets	intra-top1
(a)	Tryptophan metabolism	15	4	24	0.8000
(b)	Non-small cell lung cancer	6	4	7	0.6667
(c)	ErbB signaling pathway	22	10	35	0.4091
(d)	pathways in cancer	119	67	235	0.3529

表 5: 図 4 の pathway に含まれる drug, gene, target の数，及び intra-analogy での top1-accuracy のスコア.

ある程度アナロジータスクを解くことができた理由が説明できる．ただし PCA は 100 次元を 2 次元に次元削減して描画しているため，ある 1 つの解釈に過ぎない点は注意が必要である．

次に，pathway ごとの drug と gene の分散表現の平均ベクトルだけでなく，いくつかの pathway について，pathway に属する drug と gene の分散表現についても PCA による変換を用いて第一主成分，第二主成分をプロットした様子を図 4 に示す．また図 4 の pathway に含まれる drug, gene, target の数，及び intra-analogy で top1-accuracy を計算したときの数値を表 5 に示す．pathway に含まれる drug と gene の分散表現の平均ベクトルを結んだ線分を v_p とみなすと，intra-analogy では，drug の分散表現から線分を伸ばして，ターゲットの gene の分散表現が近くにあるかを探す状況を考えている．図 4 を見ると，それぞれの pathway で大まかに線分がターゲットの gene が存在する方向を向いているのが分かり，intra-analogy で pathway ごとに v_p を用いてうまくアナロジータスクを解くことができた結果をある程度解釈することができる．また inter-analogy では v_p の代わりに v'_p を用いる分，タスクが難しくなり性能が下がる状況であると分かる．

最後に，同じ pathway 属する遺伝子は，細胞に対する機能が類似していると推測できる．そのため，これらの遺伝子に作用する薬剤も同様の細胞機能を持つと考えられる．BioConceptVec について，このような関係性が明確に観察できれば，生物学的な妥当性を示すことができると言える．

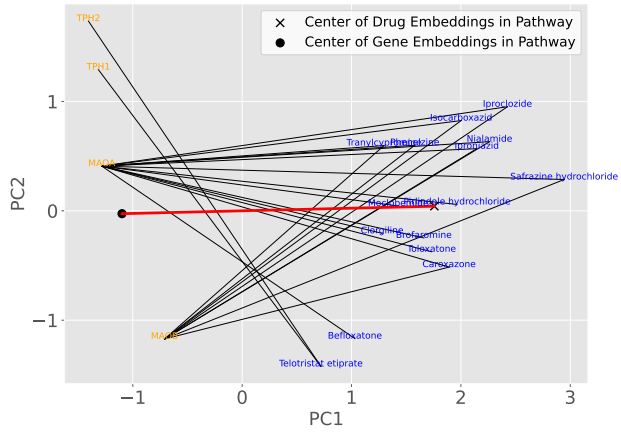
6. 今後の展望

今回の我々の実験では pathway を関係性とみなし，アナロジータスクを考えた．pathway の他にも drug と gene の間に共通の関係性を考えることができれば，アナロジータスクを設定することができるため，データセットが用意可能な他の関係性についても調べたい．

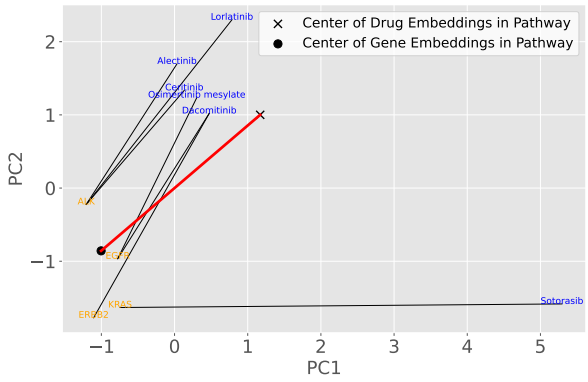
次に，今回の実験は drug から gene を予測するアナロジータスクを解いたが，gene から drug を予測するアナロジータスクを設定することもできるので，今回の結果との比較しどのような結果となるかを確認する予定である．

また inter-analogy については，具体的にどの pathway の組み合わせでアナロジータスクの性能が良いか，またそれらの pathway にどのような関係性があるかなどについても興味深い．

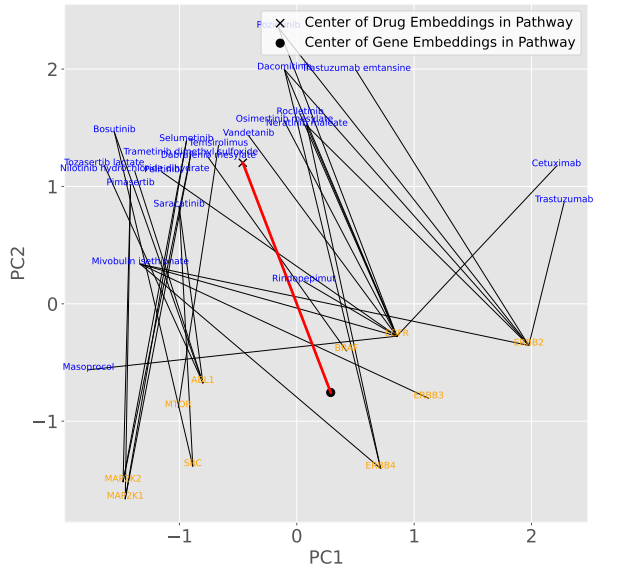
最後に，我々の動機は通常の skip-gram で成り立つアナロジー



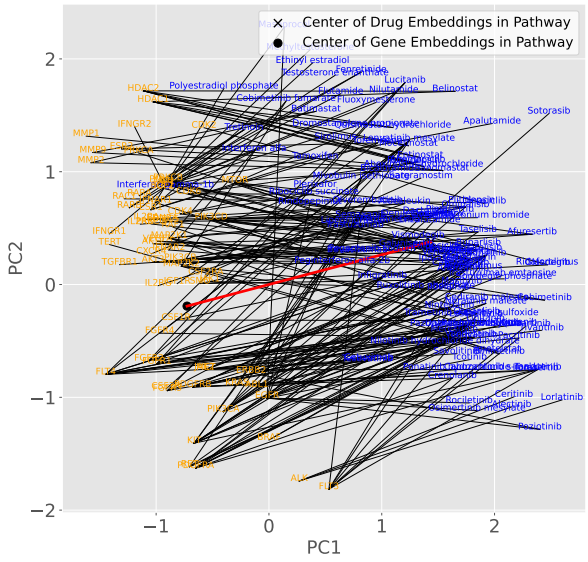
(a) Tryptophan metabolism



(b) Non-small cell lung cancer



(c) ErbB signaling pathway



(d) pathways in cancer

図 4: pathway ごとに drug と gene の分散表現，及びそれらの平均ベクトルについて PCA による変換を考え，第一主成分，第二主成分をプロットした図．pathway ごとの詳細な情報は表 5 に示す．

タスクが生物学の概念を埋め込んだ BioConceptVec の skip-gram でも成り立つかを調べることであった．通常の分散表現で成り立つ他の興味深い性質として独立成分分析 (ICA) [32] による白色化変換がある．通常の分散表現について ICA による変換を考えると，ある軸の成分の値が大きい単語は共通の意味を持ち，その軸ごとに意味を表すことが知られている [33], [34]．そこで BioConceptVec についても ICA による変換を考えた際に独立成分の軸がどのような意味を表すかを調べたい．

7. ま と め

本研究では，BioConceptVec を用いて，pathway の種類を関係性と見なし，drug と gene の組についてのアナロジータスクを解くことを目指した．BioConceptVec の skip-gram モデルと，我々が学習させた skip-gram モデルを用いて，pathway 内のアナロジータスクとして intra-analogy，pathway 間のアナロジータスクとして inter-analogy の 2 つの設定で実験を行い，その性能を比較した．結果として，pathway 内のアナロジータスクにお

いては，ベースラインよりもきわめて高い精度を達成した．また，pathway 間のアナロジータスクでも，drug の属する pathway p と異なる pathway p' から求めた $v_{p'}$ を足すことで，drug とターゲットの関係を持つ gene を十分に高い精度で予測することができた．

謝　　辞

本研究は，JST, CREST, JPMJCR21N3 (岡田, 下平) と JSPS 科研費 JP20H04148, JP20H04243, JP22H05106, JP23H03355 (下平) の支援を受けたものである．

文　　献

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, *Attention is All you Need*. In *Advances in Neural Information Processing Systems (NIPS)*, 2017.

[2] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2019.

[3] D. Cer, M. Diab, E. Agirre, I. Lopez-Gazpio, L. Specia, *SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Crosslingual Focused Evaluation*. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, 2017.

[4] T. Mikolov, K. Chen, G. Corrado, J. Dean, *Efficient Estimation of Word Representations in Vector Space*. In *1st International Conference on Learning Representations (ICLR)*, 2013.

[5] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, J. Dean, *Distributed Representations of Words and Phrases and their Compositionality*. In *Advances in Neural Information Processing Systems (NIPS)*, 2013.

[6] C. Allen, T. Hospedales, *Analogies Explained: Towards Understanding Word Embeddings*. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, 2019.

[7] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. So, J. Kang, *BioBERT: a pre-trained biomedical language representation model for biomedical text mining*. *Bioinformatics*, 2020.

[8] J. Giorgi, G. Bader, B. Wang, *A sequence-to-sequence approach for document-level relation extraction*. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, 2022.

[9] R. Luo, L. Sun, Y. Xia, T. Qin, S. Zhang, H. Poon, T. Liu, *BioGPT: generative pre-trained transformer for biomedical text generation and mining*. *Briefings in Bioinformatics*, 2022.

[10] Q. Chen, K. Lee, S. Yan, S. Kim, C. Wei, Z. Lu, *BioConceptVec: Creating and evaluating literature-based biomedical concept embeddings on a large scale*. *PLoS Comput. Biol.*, 2020.

[11] C. Wei, H. Kao, Z. Lu, *PubTator: a web-based text mining tool for assisting biocuration*. *Nucleic Acids Res.*, 2013.

[12] V. Tshitoyan, J. Dagdelen, W. Leigh, A. Dunn, Z. Rong, O. Kononova, K. A. Persson, G. Ceder, A. Jain, *Unsupervised word embeddings capture latent knowledge from materials science literature*, *Nature*, 2019.

[13] H.-M. Müller, E. E. Kenny, P. W. Sternberg, *Textpresso: An Ontology-Based Information Retrieval and Extraction System for Biological Literature*. *PLOS Biology*, 2004.

[14] C. Friedman, P. Kra, H. Yu, M. Krauthammer, A. Rzhetsky, *GENIES: a natural-language processing system for the extraction of molecular pathways from journal articles*. *Bioinformatics*, 2001.

[15] G. Shtar, A. Greenstein-Messica, E. Mazuz, L. Rokach, B. Shapira, *Predicting drug characteristics using biomedical text embedding*. *BMC Bioinformatics*, 2022.

[16] H. Alachram, H. Chereda, T. Beißbarth, E. Wingender, P. Stegmaier, *Text mining-based word representations for biomedical data analysis and protein-protein interaction networks in machine learning tasks*. *PLOS ONE*, 2021.

[17] L. Yeganova, S. Kim, Q. Chen, G. Balasanov, W. J. Wilbur, Z. Lu, *Better synonyms for enriching biomedical search*. *Journal of the American Medical Informatics Association*, 2020.

[18] S. Liu, B. Tang, Q. Chen, X. Wang, *Drug-Drug Interaction Extraction via Convolutional Neural Networks*. *Computational and Mathematical Methods in Medicine*, 2016.

[19] S. K. Sahu, A. Anand, *Drug-drug interaction extraction from biomedical texts using long short-term memory network*. *J Biomed Inform*, 2018.

[20] Z. Jiang, L. Li, D. Huang, *A general protein-protein interaction extraction architecture based on word representation and feature selection*. *International Journal of Data Mining and Bioinformatics*, 2016.

[21] C. Quan, Z. Luo, S. Wang, *A Hybrid Deep Learning Model for Protein-Protein Interactions Extraction from Biomedical Literature*. *Applied Sciences*, 2020.

[22] Y. Zhang, H. Lin, Z. Yang, J. Wang, S. Zhang, Y. Sun, L. Yang, *A hybrid model based on neural networks for biomedical relation extraction*. *Journal of Biomedical Informatics*, 2018.

[23] Y. Peng, S. Yan, Z. Lu, *Transfer Learning in Biomedical Natural Language Processing: An Evaluation of BERT and ELMo on Ten Benchmarking Datasets*. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, 2019.

[24] L. Fang, Q. Chen, C.-H. Wei, Z. Lu, and K. Wang, *Bioformer: an efficient transformer language model for biomedical text mining*, arXiv preprint, 2023.

[25] J. Du, P. Jia, Y. Dai, C. Tao, Z. Zhao, D. Zhi, *Gene2vec: distributed representation of genes based on co-expression*. *BMC Genomics*, 2019.

[26] J. Pennington, R. Socher, C. D. Manning, *Glove: Global Vectors for Word Representation*, In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014.

[27] P. Bojanowski, E. Grave, A. Joulin, T. Mikolov, *Enriching Word Vectors with Subword Information*, *Trans. Assoc. Comput. Linguistics*, 2017.

[28] S. Bird, E. Loper, *NLTK: The Natural Language Toolkit*. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, 2004.

[29] H. Ogata, S. Goto, K. Sato, W. Fujibuchi, H. Bono, M. Kanehisa, *KEGG: Kyoto Encyclopedia of Genes and Genomes*. *Nucleic Acids Research*, 1999.

[30] M. Kanehisa, S. Goto, M. Furumichi, M. Tanabe, M. Hirakawa, *KEGG for representation and analysis of molecular networks involving diseases and drugs*, *Nucleic Acids Research*, 2009.

[31] K. Iida, J. Kondo, J. N. Wibisana, M. Inoue, M. Okada, *ASURAT: functional annotation-driven unsupervised clustering of single-cell transcriptomes*, *Bioinformatics*, 2022,

[32] A. Hyvärinen, E. Oja, *Independent component analysis: algorithms and application*, *Neural Networks*, 2000.

[33] T. Musil, D. Mareček, *Independent Components of Word Embeddings Represent Semantic Features*. arXiv preprint, 2022.

[34] H. Yamagiwa, M. Oyama, H. Shimodaira, *Discovering Universal Geometry in Embeddings with ICA*. arXiv preprint, 2023.