遺伝子二重欠失研究のための関連論文検索手法

平野 颯　　野村 航　　進藤 裕之　　渡辺 太郎
奈良先端科学技術大学院大学 先端科学技術研究科

hirano.hayate.hc2@is.naist.jp　w-nomura@bs.naist.jp
{shindo, taro}@is.naist.jp

1 はじめに

生物学と情報科学の学際的分野であるバイオインフォマティクスにおいて，実験から得られた情報からの知識発見は課題の一つである．例えば大腸菌について，ゲノム（遺伝情報）の化学合成および最小化を目的とするゲノムデザインを目指し，遺伝子欠失株と呼ばれる特定遺伝子が機能しない細胞を用いて，「どの遺伝子が削れるのか」という観点から研究が進められている [1]．しかし遺伝的相互作用の組み合わせが非常に多く，生物学的知見がもっとも大きい生物である大腸菌でさえ，ゲノムデザインの達成には程遠い現状がある [1]．

本研究では，２種類の遺伝子を喪失させた遺伝子二重欠失株の生育実験をより効率的に行えるようにすることを目標とし，クエリとして与えられた２種類の遺伝子に対する相補的な関連性の検証に有用な論文を検索する方法を提案する．具体的には，木構造トピックモデルで得られる文書および単語のトピック分布を応用する．Latent Dirichlet Allocation (LDA [2]) に代表される一般的なトピックモデルと異なりトピック同士の類似性や粒度の違いを考えることができるため，別々のトピックに属する単語や文書に対しても類似するかどうかを考えることができる．実験により階層構造を持つトピック分布はそうでないものと比べ，遺伝子の相互作用をより適切に捉えられる可能性を示した．

2 関連研究

情報検索タスクには文書群からのデータ検索だけでなく，SQL をはじめとする構造化されたデータからの知識発見などさまざまな設定が存在する．このうちテキスト検索は，ユーザの入力したクエリに対してランク付けされたテキスト群を結果として返すものであり，情報検索における主要な問題である

Queries	Document	likelihood
b0720, b0002	PMC3753633	0.0156
	PMC3668280	0.00849
	PMC5001585	0.00409
b0928, b0231	PMC3753633	0.0156
	PMC3668280	0.00848
	PMC5001585	0.00408

図 1 クエリ：“b0720 b0002” および “b0928 b0231” の設定における取得された関連文献上位３件

[3]．情報検索も他の自然言語処理タスク同様，単語および文書のベクトル表現の学習に重点を置いており，BERT や GPT-2 などの事前学習済みの言語モデルを文書検索に利用する研究がある [4, 5]．

また，情報検索において文書の表現には Latent Semantic Indexing (LSI) をはじめとするベクトル空間モデルが広く用いられてきた．LDA [2] は LSI を生成モデルに拡張したものであり，潜在的な意味「トピック」からの単語の生成，および文書からのトピックの生成を行う際，Dirichlet 分布を事前分布に仮定する．代表的なトピックモデルとして様々なタスクへと応用されている．

ここで，LDA は大量文書に適応が可能な手法である一方，トピック同士の関連や粒度の違いを考慮することができない．トピック同士に階層構造を持たせる木構造トピックモデルはそれを可能にするが，大量文書への適用が困難であった．Isonuma ら [6] は木構造トピックモデルにおいて，文書からトピック分布への写像をニューラルネットワークで構成する手法を提案した．これにより大量文書へに対して木構造トピックモデルの構築，情報検索をはじめとする下流タスクへの応用が可能となった．

3 提案手法

本研究では，実験対象の２種類の遺伝子をクエリとして与えたときに互いの差異を含むような文献が

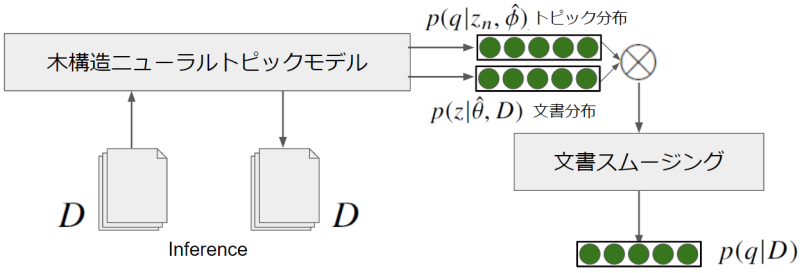

図 2　モデルの概略図

検索できるようにすることで，遺伝子欠失研究の仮説検証を補助することを目的とする．そこで生物医学分野の論文を対象に，木構造ニューラルトピックモデル [6] により獲得される文書のトピック分布およびトピックに紐づく単語分布を文書検索に利用する．

トピックモデル LDA [2] を情報検索に利用する場合，例えば文書分類では文書 d のトピック分布 θ_d およびトピック z に紐づく単語分布 ϕ_z がそれぞれパラメータ α, β である Dirichlet 分布から生成されるとした上で，文書（観測値である単語の集合 $\mathbf{w}$ で表現される）の生成確率

$$p(\mathbf{w}|\alpha,\beta) = \iint \prod_{z=1}^{X} P(\phi_z|\beta) \prod_{d=1}^{N} P(\theta_d|\alpha) \left(\prod_{i=1}^{N_d} \sum_{z_i=1}^{K} P(z_i|\theta) P(w_i|z,\phi) \right) d\theta d\phi \tag{1}$$

により各文書についてもっとも確率値の高いトピックに割り当てることで行うことができる．ここで，トピック数 K は開発データにおける perplexity と呼ばれる指標により事前に決定する必要がある．

一方，Isonuma ら [6] が提案するモデルでは事前分布から無限木上のトピック分布への写像をニューラルネットワークで構成することにより，トピック数を動的に学習させることが可能である．

図 2 は提案手法の流れを示している．クエリとして与えた 2 種類の遺伝子のそれぞれについて文書との対応の強さを推定する．具体的には，Wei ら [7] の手法に従いクエリ q に対して $p(q|D)$ を以下の式 (2) で表わされる文書スムージングにより計算する．

$$p(q|D) = \lambda \left(\frac{N_d}{N_d + \mu} p_{ML}(q|D) + \left(1 - \frac{N_d}{N_d + \mu} \right) p_{ML}(q|\mathrm{coll}) \right) + (1-\lambda) p_{tm}(q|D) \tag{2}$$

ただし $p_{ML}(q|D)$ は文書 D におけるクエリ q の最尤推定量，すなわち文書を構成する単語数 N_d に対する出現語彙数 $c(q, D)$ の割合 $\frac{c(q,D)}{N_d}$ を，$p_{ML}(q|\mathrm{coll})$ は全文書集合における クエリ q の最尤推定量，すなわち $\sum_{D \in coll} \frac{c(q,D)}{N_d}$ を表す．また，$p_{tm}(q|D)$ は文書モデルであり，各トピックに紐づく単語分布の推定値 $\hat{\phi}$ および文書のトピック分布の推定値 $\hat{\theta}$ を用いて以下のように書ける．

$$\begin{aligned} p_{tm}(q|D) &= \sum_{n=1}^{N} p_{tm}(q|z_n) p_{tm}(z_n|D) \\ &= \sum_{n=1}^{N} p(q|z_n, \hat{\phi}) p(z_n|\hat{\theta}, D) \end{aligned} \tag{3}$$

本研究では，クエリとして与える 2 種類の遺伝子に相補的な関連を持つ文書集合を獲得したい．クエリを各々 $q_1, q_2 \in Q$ とした場合に，2 つのクエリに対する文書との関連の強さは

$$p(Q|D) = \prod_{i=1}^{2} p(q_i|D) \tag{4}$$

と書けることから，式 (2) はクエリを構成する各々の単語について計算し，積を取ることで求められる．クエリである遺伝子双方に関連する文書群と，各々の遺伝子について関連する文書群との関連の大きさの差が大きかった文書を 2 種類の遺伝子に相補的な情報を持つ文書とする．

4　実験

4.1　データセット

生物医学系の研究論文が広く収集されているデータベースの一つである PubMed [1] から論文データを取得した．ここで，PubMed に収集されている論文には Medical Subject Headings (MeSH) [2] と呼ばれる

1）https://pubmed.ncbi.nlm.nih.gov/
2）https://www.nlm.nih.gov/mesh/meshhome.html

生物医学語彙データベース中の語句がラベルとして振られており，これにより的確な検索が可能となっている．本研究では，大腸菌（*Escherichia coli*）に関する研究論文のみを対象とするため，MeSH データベース中の小分類，*Escherichia coli* を利用した．下位分類として生物学の研究に主に用いられる大腸菌株 *Escherichia coli* K12 が存在するが，ラベルの網羅性の低さからこれは用いないこととした．検索された論文のうち，本文をプレーンテキストとして取得できた 16,740 件を実験の対象とした．

前述の論文本文データに加え，大腸菌遺伝子と同義な用語や概念，例えば遺伝子 *gltA* は，各大腸菌遺伝子に一意に振られる b number である b0720 と同義とみなされる．それらを同一視するため，各遺伝子の同義語を生物医学データベース群，EcoCyc [3], EcoGene [4], KEGG [5], RegulonDB [6] から収集した．

さらに，検索された文書群がどの程度クエリとして与えた遺伝子の相補的な情報を保持するかを検定するために，遺伝子二重欠失株生育データを利用する．これは大腸菌二重遺伝子欠失実験により得られた $4,000 \times 150$ 遺伝子対について，遺伝子欠失後の大腸菌の生育状況を生，死，不変にて 3 値化したものである．

4.2 実験設定

4.1 節で得られた論文本文のデータには NLTK を利用してストップワードの除去，および見出し語への変換（Lemmatization）を行った．加えて，遺伝子を表す単語は収集した遺伝子の同義語群を用いて b number に統一した[7]．低頻度語も除去し，遺伝子以外の語は 1,741 種類，遺伝子は 552 種類を利用する．これら文書群は評価用とそれ以外に 6,652 件，10,088 件に分割し，更に評価用でない文書群は訓練用と検証用に 9,031 件，1,057 件に分割した[8]．

加えて遺伝子二重欠失株生育データに対して，遺伝子を欠失順序ごとに Head 遺伝子，Tail 遺伝子と呼び，それらを入れ替えた場合に生育状況が変化するもの，変化しないものの二値化を行った．入れ替えによって生育状況が変化した遺伝子対は 1,982 組，変化しない，もしくは実験が行われていないものは 214,990 組得られた．このうち前述の 552 種類の遺伝子に含まれる各 8 組を検定に利用した．[9]

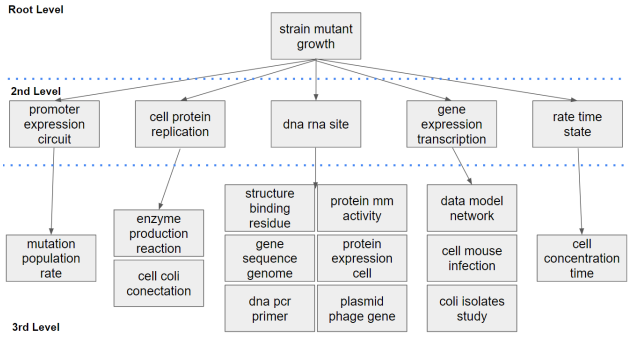

図 3 木構造ニューラルトピックモデルにより学習されたトピックの階層構造：各トピックの上位 3 語を表示

また文書モデルの計算のための木構造ニューラルトピックモデルのハイパーパラメータは Isonuma ら [6] の設定に従った．また文書検索については，Wei ら [7] のパラメータに従い，$\mu = 1000, \lambda = 0.7$ を用いた．

4.3 文書検索の結果

本研究では，式 (2) で計算される文書モデルを利用し，各文書 D に対して $\sum_i p(q_i|D) - p(Q|D)$ の降順に各 10 文献を取得した．得られた結果に対して Mann-Whitney の U 検定を用いて，遺伝子二重欠失実験において生育状況の変化のあったもの，なかったものの各遺伝子組について検索順序の中央値の差の検定を行った．

また，大腸菌二重遺伝子欠失実験により得られた欠失順序の違いにより生育状況に変化のあった遺伝子，および変化のなかった遺伝子，各 8 組について文書検索を行った．5 組について帰無仮説は棄却され，中央値に差があることが示された．結果の一例は表 1 に示した．ここから，生育状況に変化のあった遺伝子となかった遺伝子の 2 群間で異なる検索分布が提案手法により得られたと言える．

図 1 に取得された文献の一例を示す．遺伝子 b0720，b0002 は遺伝子生育実験により，欠失順序によって生育状況が変化すると判断された組，b0928，b0231 は変化しないと判断された組である．検索された文書順序は実験対象の 8 件での上位 20 文献についてはほとんど変化がなかった．これは，トピッ

3） https://ecocyc.org
4） http://ecogene.org
5） https://www.kegg.jp
6） http://regulondb.ccg.unam.mx
7） 同義な語であっても一意に b number に置換できない場合がある．本研究では収集した同義語群において，当該の語から b number への変換が一意に定まるもののみを対象とした．
8） 分割は [8] が採用し，Isonuma ら [6] も同様にしている 20 News Corpus の分割比，訓練データ : 検証データ + 評価データ = 3 : 2，検証データ : 評価データ = 1 : 9 にしたがった．
9） ある Head の欠失遺伝子について，各 Tail 遺伝子すべてについて上位 5% を生，下位 5% を死とした．生育状況 LB 培地培地上での大腸菌コロニーサイズの増減によって計算される．

表 1　文書検索結果：b0720, b0002 ペアおよび b0928, b0231 ペア同士の代表値についての両側検定結果

	Tree-Structured Topic Model		LDA	
	b0720, b0002	b0928, b0231	b0720, b0002	b0928, b0231
median	1.93×10^{-3}	1.92×10^{-3}	1.91×10^{-3}	1.90×10^{-3}
p-value	$*6.80 \times 10^{-8}$		$*6.17 \times 10^{-11}$	

$*p < 0.05$

クモデルと文書検索の学習を別々に行ったことが一因と考えられ，一体での学習は今後の課題である．

4.4　木構造トピックによる効果

フラットなトピックモデルを利用した場合と比較して木構造トピックが文書検索性能に影響があるかを検証する．図 3 は木構造ニューラルトピックモデルにより獲得されたトピックの階層構造を図示したものである．上位には “growth” などの概念的な語が見られるが，下位には “enzyme” や “pcr” など具体的な語が含まれることがわかる．フラットなトピックモデルの一例として LDA [2] を利用する．提案手法との比較のためトピック数は同一の 19 に，訓練には訓練用文書 9,031 件を用いた．検証用文書 1,057 件での Perplexity が 10 反復連続で変化しなかったところでモデルの学習を打ち切った．

結果は表 1 に示した．LDA へと変更したことにより帰無仮説の棄却数は変わらず 5 件であった．しかしながら，実験の対象とした 8 件の遺伝子の組すべてにおいて中央値は木構造ニューラルトピックモデルの方が大きくなった．このことから実験的にトピックに階層構造を持たせることが，２種類の遺伝子に相補的な関連をもつ論文を検索することを目的とする本研究には役に立つ可能性が示唆される．

5　おわりに

本研究では，情報検索技術を背景にゲノムデザイン研究の支援を行うことを目的とした．他の生物と比較して研究の進んでいる大腸菌においても，いまだ知られていない遺伝子相互作用が人の手に余るほど存在することを鑑みると，実験結果に関連のある情報を取得する技術は遺伝子欠失，ひいてはゲノムデザイン研究の効率化の一助となる．

そこで論文を効率よく検索することを目指し，木構造ニューラルトピックモデルで得られるトピックに紐づく単語分布および文書のトピック分布を後続のモデルの推論に利用し，実験により階層構造をもたないトピックモデルよりもより効果的に論文を取得できる可能性を示した．今後は木構造ニューラルトピックモデルの強みの一つである，下流タスクとの一体的な学習を進める．

参考文献

[1]浩禎 森. ゲノムデザインに向けて (創立 90 周年記念特別企画 : バイオ技術 10 年の軌跡特集 大規模ゲノム改変技術と微生物育種工学 : バイオモノづくり技術と合成生物学の発想). **生物工学会誌**, 90(6):293–297, 2012.

[2]David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[3]Bhaskar Mitra, Nick Craswell, et al. An introduction to neural information retrieval. *Foundations and Trends® in Information Retrieval*, 13(1):1–126, 2018.

[4]Wei Yang, Haotian Zhang, and Jimmy Lin. Simple applications of bert for ad hoc document retrieval. *arXiv preprint arXiv:1903.10972*, 2019.

[5]Yosi Mass and Haggai Roitman. Ad-hoc document retrieval using weak-supervision with BERT and GPT2. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4191–4197, Online, November 2020. Association for Computational Linguistics.

[6]Masaru Isonuma, Junichiro Mori, Danushka Bollegala, and Ichiro Sakata. Tree-Structured Neural Topic Model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 800–806, Online, July 2020. Association for Computational Linguistics.

[7]Xing Wei and W Bruce Croft. Lda-based document models for ad-hoc retrieval. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 178–185, 2006.

[8]Akash Srivastava and Charles Sutton. Autoencoding variational inference for topic models. *arXiv preprint arXiv:1703.01488*, 2017.