自 然 言 語 処 理 154－16
（２００３．３．７）

生物医学質問応答システム (bio-QA) の提案

平 博順 † 平尾 努 † 泉谷 知範 † 鈴木 穣 ‡ 前田 英作 †
†NTT コミュニケーション科学基礎研究所
〒 619-0237 京都府相楽郡精華町光台 2-4
{taira,hirao,izumi,maeda}@cslab.kecl.ntt.co.jp
‡ 東京大学 医科学研究所
〒 108-8639 東京都港区白金台 4-6-1
ysuzuki@manage.ims.u-tokyo.ac.jp

概要

本稿では，生物・医学分野の研究をすすめる上で欠かせない、効率のよい生物情報取得のための質問応答システム bio-QA の提案を行う．bio-QA は、生物・医学系の研究者が新規の遺伝子や蛋白質に関する研究を始める前に、対象の遺伝子や蛋白質に関する情報を分かりやすく提示するシステムである。システムは過去に発表された論文のアブストラクトの中から対象の遺伝子、蛋白質等に関するテキスト情報を抽出、統合し、研究者の質問に答える。本稿では、生物・医学の質問で頻出する関係性を問う質問に着目し、関係抽出パターンの拡張技術についての提案も行う。小規模な評価実験を行ったところ、興味深い関係を抽出することができることが分かった。

キーワード: バイオインフォマティクス，質問応答，情報抽出，情報検索

Bio-medical Question Answering System - bioQA

Hirotoshi Taira†, Tsutomu Hirao†, Tomonori Izumitani†,
Yutaka Suzuki‡ and Eisaku Maeda†
†NTT Communication Science Laboratories
2-4, Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0237 Japan
{taira,hirao,izumi,maeda}@cslab.kecl.ntt.co.jp
‡The Institute of Medical Science, The University of Tokyo
4-6-1, Shirokanedai, Minato-ku, Tokyo 108-8639, Japan
ysuzuki@manage.ims.u-tokyo.ac.jp

Abstract

This paper describes a new question answering system for bio-medical text data(bio-QA). It is a system which presents the information about target genes and target proteins intelligibly, before the biomedical researchers begin research on new genes or protein functions. This system extracts biological consequences about the target genes, proteins, etc. from the past paper abstract and replies to the user's questions. In this paper, we also present a new technique to extended a related extraction pattern for replying to a question about the relation between object to object which occurs frequently in the questions of bio-medical domain. We did a small-scale experiment and evaluated it. As a result, it has turned out that our technique can extract interesting relations from text.

Keywords: Bioinformatics, Question Answering, Information Extracstion, Information Retrieval

1 はじめに

ヒトゲノムプロジェクトをはじめとする遺伝子解析プロジェクトの進展、計算機の速度向上、記憶媒体の大容量化等に伴い、生物・医学系の情報が大量に蓄積、検索されるようになってきた。もともと、生物・医学分野は広範囲の各論的な小分野を多く持つ研究分野であったが、現在さらに多くの小分野に細分化されて研究されている。多くの研究者にとって、自分の関わっている小分野の情報を収集、管理することすら難しく、ましてや異なる小分野の情報を把握するのはさらに難しい。しかし、ヒトゲノムプロジェクトなどの全塩基配列決定プロジェクトの進展とともに、一つの遺伝子、一つの蛋白質を軸として小分野をまたがる形で、総論的に情報を把握したい、蓄積のある他分野の豊富な知見も手に入れて研究の効率化を図りたい、という要求が高まっている。

DNA の塩基配列に関しては、シーケンシングした配列を GenBank [2] [2] 等の公開データベースに登録する体制が整っており、生物・医学系の研究者の間ではこれらを配列検索ツール BLAST [1] で検索することが、一般的である。

一方、実験で得られた配列以外の知見に関しては、主に論文から得ることになる。生物・医学分野の主要な論文に関しては PubMed [3] に論文のタイトル、著者、要約等が登録される枠組みが作られており、このデータベースは生物・医学系の研究者から非常に重要視されている。

PubMed に登録されている文献は現時点で約 1200 万件と大量であるため、それらすべてに目を通すことは困難である。そこで、現状では、PubMed に対して検索を行なうシステム ENTREZ(アントレ) などの全文検索システムがよく用いられている。しかしながら、全文検索の場合、検索結果は大量であり、かつ自分が調べているものと無関係な文献が含まれることも多い。ユーザは大量の検索結果に目を通し、そこから自分で欲しい情報を取捨選択する作業が必要であり、効率が悪い。

そこで我々は情報の取捨選択を容易にするために生物医学分野のテキストを対象とした質問応答システムの提案を行う。質問応答システムとは、自然文で入力された質問に対し、回答を文字列で返すシステムであり [7]、近年、情報検索に関する評価型ワークショップ TREC(Text REtrieval Conference) [4] においても注目を浴びている。ただし TREC で対象とする質問応答はオープンドメインタスク、つまり、質問の対象領域を限定しない質問応答であり、このような設定ではドメイン知識を扱うことが非常に難しい。しかし我々の提案する bio-QA では対象を生物・医学分野に限る。対象分野を限定することによって、ドメイン知識（オントロジー）の利用が容易となり、より高度な言語処理を行うことで、応答の精度向上に大きく寄与をすることが期待できる。本稿では、bio-QA の全体像を提案し、その実現のために必須の技術の一つである関係抽出パターンの拡張手法について述べる。また、小規模な評価実験を行ったのでその結果も報告する。

本論文の構成は以下の通りである．次章で bio-QA の全体構成について述べる．3 章では、生物・医学分野での質問にはどのようなものがあるかを例示し、関係性を問う質問が割合として多いことを示す。4 章で、本稿で対象とする「関係性」について説明し、テキストからの関係抽出パターン拡張手法について述べ、実験結果を示し、考察を行う。最終章で結論を述べる。

2 bio-QA システムの概要

bio-QA の概要を図 1 に示す。まず、ユーザからの質問を自然文（テキスト）で受取り、質問解析を行なう。質問解析では質問タイプが 5W1H、さらにはもっと細分化された質問タイプのどれに当てはまるかを解析する。次に質問文から対象となる文献を探し出すのにキーワードとなるような単語を抽出し、文献検索を行う。検索された文献の中から、質問解析結果とドメイン知識を用いた解析結果をもとに最終的な回答を決定する。さらに、回答の根拠を示すための回答を含む文献を要約したあと、回答、およびその根拠をユーザに提示する。

[2] http://www.ncbi.nlm.nih.gov/Genbank/
[3] http://www4.ncbi.nlm.nih.gov/PubMed/
[4] http://trec.nist.gov/

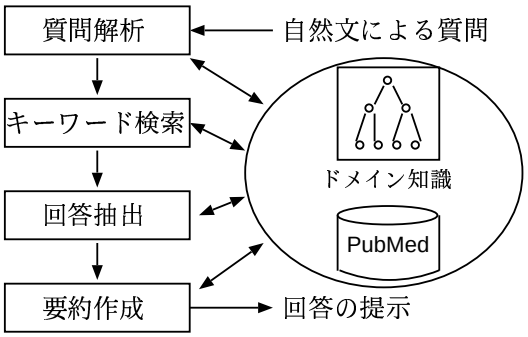

図 1: bio-QA の概略.

3　生物・医学分野での質問の特性

生物・医学分野での質問の特性を調べるため、実際に生物・医学分野の研究者に英文の質問を 22 問、作成してもらい、その分析を行なった。

表 1 にそれらの質問を掲げる。質問は大きく以下の 3 つのタイプに分類することができた。

1) ある物質の性質を問う質問
（質問番号 4,9,17,18,19,21）
2) 2 つの物質同士の関係を問う質問
（質問番号 3,7,12,13,15,16,20,22）
3) その他の質問
（質問番号 5,6,8,10,11,14）

この中で、1) のタイプの質問に精度良く応答するには、質問対象に対するドメイン知識が必要になると考えられる。2) のタイプの質問に精度良く応答するには、ドメイン知識に加えて、関係を表す言語情報が必要になると考えられる。今回は 2) に注目し、関係性についての情報を得るにはどうすればよいかを検討した。

4　関係性抽出

先の質問から分かるように物質（あるいは遺伝子、蛋白質）の関係性を同定することは重要であり、bio-QA の実現には欠かせない。ここでいう、関係性とは、物質、現象、機能同士が、どのような関係にあるか、さらにそれがどのような文脈、環境、条件、実験方法において得られたものか、という情報である。例えば、“What gene is associated with drosophila neurogenesis?”（ショウジョウバエの神経発生に関わる遺伝子は何ですか？）という質問に対し、その遺伝子は、“Groucho” で、“the interaction trap, a yeast two-hybrid system” という手法によってそのことが分かった、ということを答えたい。

関係性と一口に言ってもさまざまな種類のものが存在する。例えば、“A binds B”（A は B と結合する）、“A inhibits B”(A は B を抑制する) のような関係がある。BIND 関係については文献 [6]、INHIBIT 関係については文献 [5] において、関係の抽出が試みられている。これらは人手で書いた抽出パターンを用いて、関係を抽出する考え方に基づいている。こうしたパターンマッチングによる抽出は MUC [3] などの情報抽出コンテストにおいても多く用いられている。

しかし、このように一つ一つの関係について抽出器を作成する方法は、非常に手間がかかる。関係性には様々なものがあり、それらの抽出パターンをすべてリストアップすることは困難である。また、生物・医学分野のように研究スピードが速く、新技術、新発見が多い分野では、関係性を表す言い回しも年々増加していく。これらを人手でカバーするのは困難である。パターンを人手で書き上げることはなく、自動的に関係を抽出する汎用性の高い技術が求められている。

単純に、質問文中に出現したキーワード及び動詞をキーとして検索する方法もあるが、質問文中の動詞句、キーワードがそのままの表現で検索対象のテキストに出現することは少ない。そこで、検索語の拡張が必要となる。キーワードの名詞に関しては、オントロジーなどを利用して検索語の拡張が可能であるが、動詞の拡張に関しては難しい。そこで我々は、一つの動詞から出発し、関連名詞句を抽出し、そこからさらに動詞を検索することで検索に必要となる動詞句を増やしていくことにより関係性を抽出する方法を提案する。この方法に対して、簡単な予備実験を行ない、その結果について考察を行なう。

4.1　関係性抽出パターン拡張アルゴリズム

上で述べたように、ブートストラップ的に関係性抽出パターンを拡張することを考えた。ア

表 1: 質問の例.（「」内は日本語訳）

1. Show me the physiogoical role of IKK epsilon in the NFkB signaling pathway.
「NFkB シグナル伝達経路における IKK イプシロンの生理的役割を述べよ」
2. Show me the physiological role of Ets1.
「Ets1 の生理的役割を述べよ」
3. What molecules does TRAF6 interact with?
「TRAF6 が相互作用する分子は何？」
4. Show me the list of the target genes of TCF.
「TCF のターゲット遺伝子のリストを示せ」
5. Show me the list of the target genes of STAT6 and thier biological consequences.
「STAT6 のターゲット遺伝子のリストとその生物学的な重要性を示せ」
6. Show me the bilogical context in which the Wnt pathway is activated.
「Wnt 伝達経路が活性化されるのは、どういった生物学的な文脈のときですか」
7. What causes the activation of Hes-1?
「何が Hes-1 を活性化させますか？」
8. Show me the outline of the p38 signaling cascade.
「p38 のシグナル伝達の流れの概略を示せ」
9. What molecules are reported to show oscillation?
「どんな分子が振動を起こすと報告されていますか？」
10. Show me the recent advance in RNAi technology.
「RNAi 技術における最近の進展を教えてください」
11. Is there any knock out mouse study of p16?
「p16 欠損マウスの研究は過去にありますか？」
12. Is there any report that describes the interaction of SERF2 with HYPK?
「SERF2 と HYPK の間に相関があるという報告は今までにありますか？」
13. What molecule phorphorylates Src?
「Src をリン酸化する分子を示せ。」
14. How is p53 targetted for degradation?
「p53 を分解に導く機構を示せ。」
15. What molecule regulates the transcription of p53?
「p53 の転写を調節している分子は何ですか？」
16. What kind of cytokines invoke the translocation of NF-kB?
「どんなサイトカインが NF-kB の核移行を引き起こすか？」
17. What kind of cytokine receptors transmit the signals to JAK-STAT pathway?
「JAK-STAT パスウェイに対してシグナル伝達しているサイトカインレセプタにはどのようなものがありますか？」
18. What molecules are responsible for subcellular translocation of 14-3-3 proteins?
「14-3-3 蛋白質の細胞内局在を変化させる分子を挙げよ。」
19. Show me the list of transcription factors that are regulated by acetylation.
「アセチル化によって調節される転写因子のリストを示せ」
20. Show me the list of genes whose transcriptions are regulated by E-box?
「E-box によって転写が調節される遺伝子のリストを示せ」
21. Show me the list of papers describing the knockout phenotypes of p16.
「p16 欠損（マウス）の表現型を記載した論文を列挙せよ。」
22. What causes the degradation of IkB?
「何が IkB の分解を引き起こしますか？」

ルゴリズムの詳細は以下の通りである。
(Step 1) 対象となるテキストから、種となる動詞句＋前置詞 (VP+P) パターンの前後に来る名詞句（NP) で閾値以上の組を検索
(Step 2) 検索された名詞句の組 (NP1,NP2) が前後に出現するような VP+P パターンで頻度が閾値以上のパターンを検索
(Step 3) Step 2 で抽出された VP+P パターンの前後に来る名詞句 (NP) の組で閾値以上の組を検索

このようなアルゴリズムを用いることで、関係性として少量のデータを与えるだけで、多くの関係情報を得ることが期待できる。

4.2 実験結果

4.2.1 実験設定

実験には，PubMed に登録されている文献アブストラクトのうち、Swiss-Prot (Release40) [5] にも登録されている約 10 万件を用いた．SwissProt は蛋白質の配列（2 次構造まで）のデータベースである。このデータに対し、Brill タガーを使って形態素解析、自作の chunker で chunking をおこなった。最初に与える動詞句のパターンは、“is associated with” とした。

4.2.2 評価

評価は、以下のような基準で人手で行った。抽出された (NP1,NP2) の組について、相関があり、意味のある情報を含む場合を Y、誤っている場合を N、間違いではないが情報をほとんど含まない場合、または判定することが困難あるいは文脈による場合を NA とした。

初めに “is associated with” で検索を行った場合に抽出された (NP1,NP2) の組を表 2 に示す。頻度の閾値は 5、つまり 5 回以上出現する (NP1,NP2) の組が示されている。そのような組は全部で 9 組であったが、これらの評価はすべて NA であった。この結果は、動詞句をそのまま検索語に用いても、有用な情報を抽出するのが困難であることを示唆している。その内訳は、チャンキングの処理が不十分なため、NP1 や NP2 に that,the など真の名詞句が来なかったものが 3 組、照応関係が分からないために関係が判定できなかったものが 3 組、NP1 と NP2 に来たものが一般的過ぎて情報持たなかったものが 3 組であった。例えば、H1.4 遺伝子と H2B 遺伝子はヒストンの遺伝子で遺伝子名の最初にくる H の文字は Histon の頭文字から取られたものである。DNA は細胞の核の中に折り畳まれているときには、このヒストンと呼ばれる蛋白質を中心にして巻かれる構造になっている。このヒストンはいくつかの部分からなっており、それら一つ一つの蛋白質をコードしている遺伝子に H+（数字）という形で名前が付けられている。このように機能等の名前の頭の部分を取ったアルファベットと数字（あるいは大文字のアルファベット一文字）で遺伝子名をつけることが非常に多い。よって、逆に言えば関係があるものとして H+（数字）のようなものが出てきても当然でほとんど情報がない。

次に、これら 9 つの (NP1,NP2) を使って、二つの名詞句に挟まれる VP+P パターンを抽出した。2 回以上出現した VP+P パターンは 25 種類であった。その結果を表 3 に示す。種パターン “is associated with” に対して、“is” が “are”,“may be” と変化したパターンのみならず、“correlates with”,“interacts with” など表現は異なるが、意味的に似通った表現が多数抽出されている。また、“is phosphorylated by”（リン酸化されている）など、一般にはほとんど出現しないが、生物・医学分野では出現する関係表現も抽出できている。我々の提案手法が有効に働いていると言える。

さらにこれら 25 種類の VP+P に隣接する 2 つの NP の組を検索し、5 回以上出現するものを抽出した。その結果の一部を表 4 に示す。得られた (NP1,NP2) の組は全部で 246 組で、そのうち Y は 11 個、N は 1 個、それ以外の 234 個は NA であった。

Y と判定されたものを具体的に見てみると、AapJ 遺伝子が輸送に関係する遺伝子、Groucho 遺伝子がショウジョウバエの神経発生に関係する遺伝子、であるなど 1 回目の検索では得られなかった有用な関係が得られている。また、“all exon-intron junction sequences” と “the GT rule” との関係が抽出されている。これは

[5] http://kr.expasy.org/sprot/sprot-top.html

表 2: 抽出された関連候補（1 回目）.

NP1	NP2	評価
The H1.4	an H2B gene	NA
Another island	two genes	NA
(もう一つのアイランド)	(二つの遺伝子)	
The ndhD gene	a gene	NA
ovine trophoblast protein	the maternal recognition	NA
(羊の発生栄養膜蛋白)	(その母性認識)	
higher primates	increased variation	NA
	(増加した変異)	
that	initiation	NA
	(開始)	
a result	the	NA
the E coli uncI gene	this locus	NA
(その大腸菌の uncI 遺伝子)	(この座位)	
that	virulence	NA
	(毒性)	

DNA 塩基配列の中で遺伝子をコーディングしている部分をエキソン、そうでないものをイントロンと呼ぶが、その境界の配列が GT ルールに関係あるということが示されている。GT ルールとは、もともと GT-AG ルールと一般に呼ばれているもので、エキソンの先頭の配列は “GT”, 末尾の配列は “AG” であるという、生物の研究者であれば誰でも知っているような非常に一般的なルールである。生物学においては一般的な関係が成り立つことが少ない中で、このような一般的な関係がテキストベースで抽出されていることは、非常に興味深い。このように全体としては、提案手法が有効であったと言える。

ただし、検索語拡張を行うときに、よく起こる無意味な関係も多く抽出されてきている。NA と判定されたものを見てみると、まず、表 2 で取り上げた問題、つまり H1.4 遺伝子と H2B 遺伝子、nodM 遺伝子と nodN 遺伝子が各々関係するのは命名上当然であり、有用な情報が含まれていないという問題がある。また、NP1,NP2 の一方に which,that のような語が来てしまうという問題、the biosynthesis（その生合成）というように照応関係が解決されていない問題、different genes というように対象が一般的過ぎて有益な情報にならないという問題が見られる。

この他にも生物・医学分野のテキストを扱う際の問題として、同じ物質に対し、複数の名称がしばしば存在することが挙げられる。例えば、p21 遺伝子は人によっては、 waf-1 または cip-1 とも呼ばれている。さらにそれ以前にどこからどこまでが専門用語であるかを抽出することも大変難しい。

こうしてみると、専門用語抽出処理、照応関係処理は応答精度の向上には不可欠であることが分かる。最近、GENIA プロジェクト [4] など、生物・医学分野のオントロジ、タグ付きコーパスの利用が可能になりつつある。このようなドメイン知識、タグ付きコーパスを利用すれば、さらに深い処理を行うことができ、応答精度の向上につながると考えられる。

5 結論

本稿では，生物･医学分野の研究をすすめるための効率のよい生物情報提示システム bio-QA の提案を行った。bio-QA は、生物・医学系の研究者が新規の遺伝子や蛋白質に関する研究を始める前に、対象の遺伝子や蛋白質に関する情報を分かりやすく提示するシステムであるが、過去に発表された論文のアブストラクトの中か

表 3: 抽出された VP+P パターン.

VP+P	頻度
belongs to	2
hybridized to	2
correlates with	2
is required for	2
were cloned from	2
is highly expressed in	2
conform to	2
occurs at	2
are located on	2
may be of	2
is phosphorylated by	2
are encoded by	2
formed with	2
is found in	2
interacts with	2
had been isolated from	2
are expressed in	2
corresponds to	2
is involved in	2
is expressed by	2
may be involved in	2
is associated with	8
are associated with	3
may be associated with	2

ら対象の遺伝子、蛋白質等に関する情報を抽出統合し、ユーザの質問に答えるために必要となる関係性抽出パターンの拡張技術について提案を行った。興味深い関係性を得ることができたが、今後、さらに高度な自然言語処理を利用して抽出精度を高めるとともに、bio-QA を構築し、評価を行っていきたい。

参考文献

[1] Altschul, S., Gish, W., Miller, W., Myers, E. and Lipman, D.: Basic local alignment search tool, *J. Mol. Biol.*, Vol. 215, No. 3, pp. 403–410 (1990 Oct 5).

[2] Benson, D., Karsch-Mizrachi, I., Lipman, D., Ostell, J., Rapp, B. and Wheeler, D.: Genbank, *Nucleic Acids Research*, Vol. 30, No. 1, pp. 17–20 (2002 Jan 1).

[3] DARPA: *Proc. of the 7th Message Understanding Conference(MUC-7)*, Fairfax, VA, USA, Morgan Kaufmann (1998).

[4] Ohta, T., Tateisi, Y., Kim, J. and Tsujii, J.: The GENIA Corpus: an Annotated Corpus in Molecular Biology Domain, *Proc. of 10th International Conference on Intelligent Systems for Molecular Biology (ISMB 2002)* (2002).

[5] Pustejovsky, J., Castano, J. and Zhang, J.: Robust Relational Parsing over Biomedical Literature: Extracting Inhibit Relations, *Proc. of Pac. Symp. on Biocomputing* (2002).

[6] Rindflesch, T. C., Rajan, J. V. and Hunter, L.: Extracting Molecular Binding Relationships from Biomedical Text, *Proc. of 6th Applied Natural Language Processing Conference and 1st Meeting of the North American Chapter of the Association for Computational Linguistics (ANLP/NAACL-2000)*, pp. 188 – 195 (2000).

[7] 佐々木裕, 磯崎秀樹, 平博順, 平尾努, 賀沢秀人, 鈴木潤, 前田英作: SAIQA: 大量文書に基づく日本語質問応答システム, 情報処理学会研究報告 NL145/FI64-12. IPSJ (2001).

表 4: 抽出された関連候補（2 回目）.

NP1	NP2	評価
ctaA	heme	Y
neurexophilin	neurexins	Y
the pul cluster	secretion. outC	Y
the Ght3p function (Ght3p の機能)	D-gluconate transport (D-グルコン酸の輸送)	Y
MTM1	severe hypotonia (重度の低血圧)	Y
AapJ	transport (輸送)	Y
the flgA gene (flgA 遺伝子)	P-ring formation (P リングの形成)	Y
ARI-1	a novel ubiquitin-conjugating enzyme (一つの有名なユビキチン結合酵素)	Y
all exon-intron junction sequences (すべてのエクソン・イントロン境界配列)	the GT rule (GT ルール)	Y
Groucho (Groucho 遺伝子)	Drosophila neurogenesis (ショウジョウバエの神経発生)	Y
cul-1	cell cycle exit	Y
the minCDE genes (minCDE 遺伝子)	pSymB	N
Two periplasmic amino acid-binding proteins (二つの周辺質アミノ酸結合蛋白)	the livJ (livJ 遺伝子)	NA
the nodM gene	the nodN gene	NA
The H1.4 gene	an H2B gene	NA
DoxH	the conversion (その変換)	NA
III intron	ycf12	NA
S. cerevisiae Orc2p	Arabidopsis thaliana	NA
which	the biosynthesis (その生合成)	NA
the IIB,IIC and IID domains (その IIB と IIC と IID 領域)	different genes (異なる遺伝子)	NA
replicatiors	the initiation (その開始)	NA
operons (オペロン)	strains	NA
epsilon-subunits	the end-plate channel	NA
alpha 1-PI	a small gene family (一つの小さな遺伝子ファミリ)	NA
ORC (複製開始因子 ORC（origin recognition complex）)	various eukaryotes (様々な真核生物)	NA
that	Thizobium meliloti	NA
the enzyme (その酵素)	the structural genes hoxK and hoxG (その構造遺伝子 hoxK と hoxG)	NA