社団法人　情報処理学会　研究報告
IPSJ SIG Technical Report

2003－ＮＬ－158　(21)
2003／11／7

生物医学文献からの遺伝子機能フレーズの抽出

平 博順　　泉谷 知範　　平尾 努　　磯崎 秀樹　　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,izumi,hirao,isozaki,maeda}@cslab.kecl.ntt.co.jp

概要

本稿では、生物医学文献から、目的とする遺伝子に対し、その遺伝子の機能が記述されてある部分を抽出する手法について述べる。TREC Genomics Track 第２タスクを例に、Tidal PrefixSpan を用いたスキップを許す特徴的な単語列のパターンの抽出について説明し、得られた抽出パターンが遺伝子の機能について述べた文やフレーズの抽出に有効であることを示す。

キーワード: 生物医学文献、遺伝子、情報抽出、情報検索、TREC、Genomics Track、PrefixSpan、TidalSMP

An Extraction Method for Gene Functional Phrases from Bio-medical Documents

Hirotoshi Taira, Tomonori Izumitani, Tsutomu Hirao,
Hideki Isozaki 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

Abstract

This paper describes a extraction method for gene functional phrases from bio-medical documents. We explain the two tasks of TREC Genomics Track. We show that our method which mines patterns using Tidal PrefixSpan for the extraction of gene fuctional phrassses, is effective.

Keywords: Biomedical document, Gene, Information Extracstion, Information Retrieval, TREC, Genomics Track, PrefixSpan, TidalSMP

1　はじめに

現在、生物・医学の研究分野において様々な生物データのデータベース化が急速に進んでいる。生物・医学分野の文献についても例外ではない。生物・医学分野の文献データベースでは、MEDLINE が世界最大規模で、現在では MEDLINE に医学・生物学系以外の文献も加えた、PubMed [2] と呼ばれる巨大な文献データベースが発展し続けている。このデータベースには、文献のタイトル、著者、アブストラクト等が登録されており、特に生物・医学系の研究者から重要視されている。

PubMed の検索には、ENTREZ などの全文検索システムが用いられるが、文献数が約 1200 万件と膨大であるため、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となる。こうした手間を省くため、必要な情報を大量の文献情報の中から自動的に抽出する技術が強く求められている。

生物医学文献からの情報検索技術は TREC (Text REtrieval Conference) [3] からも注目され、2003 年には、生物医学文献を対象とした Genomics Track が新たなトラックとして加えられた。生物医学文献から情報抽出では、特に、遺伝子や蛋白質の性質、機能の抽出が重要な課題となるが [17]、TREC Genomics Track においても、遺伝子の機能の抽出をメインターゲットとしている。

我々は、TREC Genomics Track の 2 つのタスクについて取り組んだが、特に第２タスクについては Tidal PrefixSpan という新しいマイニング手法を取り入れて、遺伝子機能フレーズの抽出を行ったので、これを中心に紹介する。

本稿の構成は以下の通りである．次章では、TREC Genomics Track のタスクについて紹介する．３章では、我々が Genomics Track で用いた機能フレーズ抽出手法について述べる。４章では、実験結果を示し、考察を行う。最終章で結論を述べる。

[2] http://www4.ncbi.nlm.nih.gov/PubMed/

[3] http://trec.nist.gov/

2 TREC Genomics Track の概要

TREC は NIST (The National Institute of Standards and Technology; 米国立標準技術研究所) が主催し、1992 年から毎年開催されている情報検索に関する評価型ワークショップである。タスクの種類によって、Robust Retrieval Track、Question Answering Track、Web Track など約 10 のトラックに分かれている。参加団体は、まず、選択したトラックで設定されたタスクを自動的に処理するプログラムをコーディングする。そして、与えられたテストセットに対するプログラムの出力結果を期限までに TREC に提出する。TREC 側は、提出団体のスコアを採点し、順位を発表する。

毎年、トラックやタスクの見直しが行われる中、2003 年に Genomics Track が新たなトラックとして加わった。Genomics Track では第 1 タスクと第 2 タスクの 2 つのタスクが与えられた。以下、この 2 つのタスクの概要について述べる。

2.1 第 1 タスク

第 1 タスクは、指定された遺伝子について、その機能が記述されている文献を、大量の文献アブストラクトの中から検索するタスクである。検索対象は、生物医学文献データベース MEDLINE に 2002 年度新たに登録された 52 万 5938 件の文献アブストラクトである。検索要求は、ランダムに選択された 50 個の遺伝子である。また、これとは別に訓練用の 50 個の遺伝子が事前に与えられる。このタスクの一つの特徴として、検索要求が、複数の遺伝子名のセットで与えられることが挙げられる。一般に、遺伝子は発見者などが命名した名前や、機能から命名された名前など、多くの別名を持っているため、このような検索要求の形式を取っていると思われる。表 1 に、第 1 タスクの問題、すなわち検索要求の一部を示す。ここで、「生物名」は検索要求である遺伝子を含む生物種名であり、2003 年の問題では、Homo sapiens (ヒト), Mus musculus (マウス), Rattus norvegicus (ラット), Drosophila melanogaster (キイロショウジョウバエ) の 4 種が含まれていた。「遺伝子のタイプ」は、その右に記されている遺伝子名のタイプを表す。遺伝子名のタイプには、

- **OFFICIAL_GENE_NAME** : 公式遺伝子名
- **OFFICIAL_SYMBOL** : 公式略称
- **ALIAS_SYMBOL** : 別名略称
- **PRODUCT** : その遺伝子から生成される蛋白質名
- **ALIAS_PROT** : その遺伝子から生成される蛋白質名の別名
- **PREFERRED_PRODUCT** : RefSeq [4] で使われる蛋白質名

などがある。

遺伝子名は、一般的な語が連なった複合語、あるいは "TEL" など、一般語と紛らわしい遺伝子名が多く存在するため、複数の遺伝子名が検索要求として与えられるにも関わらず、高精度の検索は難しい。

評価には、各問題に対する Average Precision [1] の全 50 問の平均値である Mean Average Precision [14] が採用されている。

2.2 第 2 タスク

第 2 タスクは、指定された文献フルテキストを参照して、指定された 139 個の遺伝子の機能を記述する文またはフレーズを生成するタスクである。回答の評価は難しい問題であるが、2003 年度は暫定的に LocusLink [5] の GeneRIF の記述を正解と見做して、評価を行っている。LocusLink は、遺伝子データベースの一つであり、遺伝子の公式名称、別名、配列、発現型、関係するウェブサイトなど、多種のデータが登録されている。GeneRIF (Gene Reference into Function) は、LocusLink データベース中に存在する、遺伝子の機能がテキストで記述されているフィールドである。GeneRIF は、NCBI (National Center for Biotechnology Information; 米国バイオテクノロジー情報センター) の専門家が、人手で登録を行っている。参照対象となる文献フルテキストは、全部で 133 件である。2002 年後半に出版された 5 つの論文誌 (Journal of Biological Chemistry、Journal of Cell Biology、Nucleic Acids Research、Proceedings of the National Academy of Sciences、Science) の中から、各出版者から使用許諾を得たものである。

評価には、実際の GeneRIF と出力結果の間の、改良 Dice 係数が採用されている。正規の Dice 係数 [3] は、

$$Dice(A,B) = \frac{2Z}{X+Y}$$

で定義される係数である。ここで、A と B はある 2 つの文字列、X と Y はそれぞれ A と B に含まれる単語数である。また、Z は A と B の両方に出現した単語数である。Dice 係数には、

- ストップワードが考慮されていない
- ステミングなどの単語の正規化が考慮されていない
- 複数回現れた単語の回数や語順が考慮されていない

といった問題があるため、下記の 4 つの改良された Dice 係数が評価として採用された。

- **Classic Dice (CD と略す)** : 単語列に対しストップワードの除去を行い、Porter のステミングを行って得られた単語列に対する Dice 係数
- **Modified Unigram Dice (MUD と略す)** : 単語列中の異なり単語同士についての Dice 係数
- **Bigram Dice (BD と略す)** : 語順を考慮するために bi-gram を取った上での Dice 係数

[4] http://www.ncbi.nlm.nih.gov/RefSeq/

[5] http://www.ncbi.nlm.nih.gov/LocusLink/

表 1: TREC Genomics Track 第 1 タスク問題の一部.

問題番号	生物名	遺伝子名のタイプ	遺伝子名
3	Homo sapiens	OFFICIAL GENE NAME	ets variant g3n3 6 (TELoncogene)
3	Homo sapiens	OFFICIAL SYMBOL	ETV6
3	Homo sapiens	PRODUCT	ets cariant gene 6
4	Homo sapiens	OFFICAL GENE NAME	fibroblast growth factor 7 (keratinocyte growth factor)

- **Bigram Phrases (BP と略す)**：ストップワードの除去した後で Bigram Dice を計算したもの

3 機能フレーズの自動抽出

我々は、第 1 タスク、第 2 タスク両方について取り組んだが、本稿では、特に第 2 タスクに絞って、我々の取った手法について説明する。我々は、遺伝子の機能を記述するフレーズを完全に自動生成する方法として、遺伝子の機能を記述した文に特徴的な単語（正確にはステム）の並びのパターンを訓練集合から抽出し、パターンに情報量を用いたスコアを与える。そして文分割されたテスト集合に対して、各パターンの出現に応じて、スコアを加算し、合計のスコアの高かった文を出力する、という方法を取った。

3.1 訓練集合に対する正例、負例のラベル付与

まず、情報量を計算するための準備として、訓練集合に現れた文が正解に近いか否かを判断し、正例と負例のラベル付与を行う。簡単な文区切りプログラムで文区切りを行ったあと、訓練集合の中から、正解となる実際の GeneRIF と編集距離 (Edit Distance) が小さな文を抽出した。選び出された文を正例、それ以外の文を負例とする。

具体的には、編集距離が GeneRIF の文字数の 3 割以下の文、および編集距離が GeneRIF の文字数の 3 割より大きいが、その文献で一番編集距離が小さい文を正例とした。それ以外を負例とした。

3.2 物質名の判定

文中に出現する遺伝子についての情報は、その文が対象遺伝子の機能についての記述を含むか否かを判断する上で重要である。そこで、我々は、検索要求となる遺伝子およびその他の遺伝子を表す単語列を各文から特定し、それぞれ、“<QUERY_GENE>”、“<SUBSTANCE>” という文字列に置き換えた。

生物医学文献において、遺伝子は、“BMP2” や “CDKN1A” のようなアルファベットと数字からなる遺伝子略称で表現されるか、“BCL2-associated X protein” のような単語列からなる遺伝子名で表現される。これまで、遺伝子名、遺伝子略称を抽出する方法として、さまざまな提案があるが [11, 4, 6, 8, 2, 7, 5]、我々は次のような手法を用いた。

検索要求となる遺伝子名、遺伝子略称の抽出については、LocusLink に登録されているすべての別名を用いて、与えられた文献フルテキストすべてについて、大文字、小文字の区別なしに、全文探索し、一致した部分を “<QUERY_GENE>” という文字列に置き換えた。

また、検索要求以外の遺伝子名、遺伝子略称の抽出については、LocusLink と GOA データベースに登録されている遺伝子名や遺伝子略称の全文探索と、いくつかの経験的なルールを用いた。抽出された単語列を “<SUBSTANCE>” という文字列で置き換えた。

遺伝子略称の探索には、LocusLink に登録されている遺伝子略称の中で、3 文字以上で構成される略称を用いた。これは、データベースに “do”、“a” など頻繁に現れる単語が含まれるためである。また、遺伝子略称は括弧やスラッシュなどで区切られて記述される場合が多いので、文献中のすべての単語の先頭と末尾の記号は削除して探索を行った。

遺伝子名の探索には、GOA データベース [6]に登録されている遺伝子名を用いた。探索においては遺伝子コード、遺伝子名ともに大文字小文字の区別はしていない。

また、遺伝子名と遺伝子略称とは必ずしもデータベースに登録されている通りに表現されるとは限らないため、以下のような経験的なルールを導入し、当てはまるものも、遺伝子名や遺伝子略称と見做した。

まず、数字と大文字のみで構成される 3 文字以上 8 文字以下の単語で、以下の条件に当てはまらないものを遺伝子略称とみなした。

- 数字のみ、またはアルファベットのみで構成されているもの。
- “A”、“T”、“G”、“C” のみからなるもの (DNA 塩基配列が抽出されてしまうことを避けるため)。

次に、遺伝子名として、“the” で始まり、“子音+ase”、“子音+in”、“tor”、“ssor” で終わる単語列で以下の条件に当てはまらないものを遺伝子名とみなした。

- ストップワード (PubMed で定義されたもの[7]を元に独自に作成) を含むもの。

[6] http://www.ebi.ac.uk/GOA/

[7] http://www.ncbi.nlm.nih.gov/entrez/query/static/help/pmhelp.html#Stopwords

- それまでに探索された “<QUERY_GENE>”, “<SUBSTANCE>” を含むもの。
- 動詞の活用語尾 “-ing”、“-ed” やピリオド、コロンなどの記号を含むもの。
- 左括弧, 右括弧のどちらかのみしか含んでいないもの。

3.3 ステミング処理

表層の単語列からも、パターン抽出は可能であるが、例えば、意味的には同じ、“inhibition of A” と “A inhibits” という表現を同じもおんとしてカウントすることはできない。そこで、単語に対して Porter のステミングアルゴリズム [10] を用いてステミング処理を行った上で、パターン抽出を行った。

例えば、

Regulation of Fas-associated Death Domain Interactions by the Death Effector Domain Identified by a Modified Reverse Two-hybrid Screen

という文は、ステミング処理後

<regul> <fas-associ> <death> <domain> <interact> <SUBSTANCE> <domain> <identifi> <modifi> <revers> <two-hybrid> <screen>

というステム列に変換される。

3.4 Tidal PrefixSpan によるパターン抽出

得られたステム列に対し、正例に対して特異的に出現する、ステップを許すパターンを抽出する。様々な情報量の値が利用可能だが、我々は、超幾何分布を用いた指標 (hgs) を利用した。

久光ら [15] は、与えられた文書集合を特徴づける単語への重みづけの方法として、hgs を用いて重みを与える方法を提案している。記事内容を概観するのに有効な単語を選択できたかどうかで、その有効性を検証し、tf-idf などの指標に比べて有効であることを示している。また、機能フレーズを含む文の抽出では、「このパターンが入っていると、機能を述べている文ではない」というケースは稀であると考えられ、負例にどのようなパターンが出現するかより、正例にどのようなパターンが出てくるか、の方が重要であると考えられる。そのため、χ^2 値のように、正負例のどちらの特徴の偏りを見る情報量よりも、hgs のように正例に重きを置いた情報量を用いる方が適していると考えた。

ここで、この超幾何分布を用いたスコア (hgs) の定義は、「m 個の正例を含む n 個のサンプルの中から、重複なく x 個を取り出したときに、y 個以上が正例である確率」である。この hgs の逆数の対数、すなわち $-\log(hgs)$ を指標とする統計量として用いた。

高速化のため、Tidal PrefixSpan [13] を用いてパターンの抽出を行った。PrefixSpan [9, 16] は Pei らによって提案された、スキップを許す高頻度出現パターンの高速抽出方法である。PrefixSpan を使えば、例えば、

1. I should point out that we need ...
2. I must point out that it is important ...

という 2 つの文が与えられたとき、

“I”-“point”-“out”-“that”

という “I” と “point” との間にスキップを許したパターンが 2 つ出現したことを高速にカウントすることができる。“I”-“point”-“out”-“that” が 2 回出現するためには、1 つ短いパターンの “I”-“point”-“out” は 2 回以上出現することが必用条件である性質に着目し、前から順番に 1 つずつ高頻度に出現する単語をリストアップすることにより、カウントの高速化をはかっている。

ただ、オリジナルの PrefixSpan は高頻度パターンを取り出すにすぎないので、統計的に意味のあるパターンを取り出したい場合には、工夫が必要である。ここで、Tidal SMP (Tidal Statistical Metric Pruning) [12] が利用できる。Tidal SMP は、ある決まった数の統計量の多いパターンを抽出する場合に、高速化が行える手法である。PrefixSpan に Tidal SMP の手法を適用した Tidal PrefixSpan を利用し、統計的に有意なパターンの抽出を行った。指標とする統計量には、$-\log(hgs)$ の値をそのパターン長 ($= 1, 2, 3, ...$) で除した数とし、文に対するスコアとしても利用した。

3.5 機能フレーズ出力

テスト集合の各文に対し、訓練集合から得られたステムパターンの有無を調べ、パターンを含んでいれば、そのパターンのスコアを加算し、各文毎のスコアを計算した。次に、各文が出現した場所（タイトル、アブストラクト、本文、キャプション）毎に最高のスコアを持つ文を取り出し、適度な重み付けを行った後、最終的な出力文を選んだ。基本的には 1 文を出力し、文が長い場合は、先頭の 256 文字のみを出力した。

3.6 実験結果

Tidal PrefixSpan を用いて、パターン長が 3 以下、頻度 2 以上で hgs の値の高い上位 800 パターンを抽出した。初めに、2 頻度以上出現するパターンについて、Tidal PrefixSpan で抽出された上位のステムパターンを表 2 に示す。<crystallin>（crystallin; 水晶体）、<len>（lens; レンズ）など、一般的にはあまり出現しないパターンが、訓練集合にたまたま出現することから、抽出されてしまっていることが分かる。これは ($-\log(hgs)$ / パターン長) の値が低頻度のパターンでも大きくなることがあるためである。

次に、100 頻度以上出現するパターンについて、抽出された上位のパターンを表 3 に示す。テストデータにも出現しそうな、より汎化されたパターンが抽出できていることが分かる。このことから、頻度による足切りと、超幾何分布を用いた値の組合わせにより、汎化されたパターンが抽出できることが分かる。

表 2: 抽出されたステムパターン（上位 30 パターンまで）2 頻度以上.

パターン	正例頻度	負例頻度	$-\log(hgs)$ / パターン長
<crystallin>	13	28	44.40
<regul>	31	1095	29.25
<crystallin> <gene>	13	8	27.85
<len>	7	25	21.15
<crystallin> <express>	10	7	21.15
<human>	27	1264	19.45
<signal>	26	1180	19.37
<gene>	33	1887	18.76
<QUERY_GENE>	50	3818	18.71
<SUBSTANCE> <crystallin>	9	10	17.75
<crystallin> <gene> <express>	10	4	15.08
<pathwai>	19	826	15.01
<regul> <SUBSTANCE>	22	511	14.34
<recognit>	7	83	14.09
<SUBSTANCE>	135	18437	13.99
<suggest>	20	1022	13.29
<suffici>	8	139	13.20
<conclud>	6	61	13.08
<gene> <len>	5	0	13.00
<express>	46	4081	12.86
<SUBSTANCE> <crystallin> <gene>	8	4	11.82
<moieti>	4	19	11.78
<co-activ>	5	46	11.53
<pyrophosph>	4	21	11.43
<crystallin> <crystallin>	5	3	10.99
<gtp-bound>	4	27	10.55
<necessari> <gtp-bound>	4	0	10.39
<level> <crystallin>	4	0	10.39
<human> <moieti>	4	0	10.39
<gene> <crystallin>	4	0	10.39

出力結果に対する評価は、4 つの改良 Dice 係数で行った。全 139 問の平均では、CD: 49.03%, MUD: 50.74%, BD: 32.67%, BP: 35.54%という成績であった。具体的な結果の一部を表 4 に示す。これは、139 問中 Classic Dice 係数の成績でソートしたときの、上位 1、5、10、50、100 番の問題についての結果である。50 番目の問題のように、一見、非常に正解に近い出力が出ていても、せいぜい、bi-gram しか評価対象になっていないので、PrefixSpan でスキップを許したパターンを見つけても、改良 Dice 係数では低い評価しか得られないケースがあることが見て取れる。これらの評価方法についても、今後の課題であろう。

4 結論

本稿では、生物医学文献から遺伝子の機能について述べた文やフレーズを抽出する際、スキップを許した特徴的な単語列によるスコアリングが有効であることを示した。また、スキップを許した特徴的な単語列は、Tidal PrefixSpan を使うことにより高速に抽出できることが分かった。TREC Genomics Track の第 2 タスク関しては、評価手法の改善が大いに必要であると考えられる。

参考文献

[1] Buckley, C. and Voorhees, E.: Evaluatiing measure stability, *Proc. of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR-2000)*, pp. 33–40 (2000).

[2] Collier, N., Nobata, C. and Tsujii, J.: Extracting the names of genes and gene products with a hidden markov model, *Proc. of the 18th International Conference on Computational Linguistics(COLING'2000)*, pp. 201–207 (2002).

[3] Dice, L. R.: Measures of the Amount of Ecologic Association between Species, *Journal of Ecology*, Vol. 26.

[4] Fukuda, K., Tsunoda, T., Tamura, A. and Takagi, T.: Toward information extraction: Identifying protein names from biological papers, *Proc. of the Pacific Symposium on Biocomputing*, Vol. 3, pp. 705–716 (2003).

[5] Kazama, J., Makino, T., Ohta, Y. and Tsujii,

表 3: 抽出されたステムパターン（上位 30 パターンまで）100 頻度以上.

パターン	正例頻度	負例頻度	$-\log(hgs)$ / パターン長
<regul>	31	1095	29.25
<human>	27	1264	19.45
<signal>	26	1180	19.37
<gene>	33	1887	18.76
<QUERYGENE>	50	3818	18.71
<pathwai>	19	826	15.01
<regul> <SUBSTANCE>	22	511	14.34
<SUBSTANCE>	135	18437	13.99
<suggest>	20	1022	13.29
<suffici>	8	139	13.20
<express>	46	4081	12.86
<evid>	9	273	10.35
<function>	17	988	9.86
<gene> <express>	16	419	9.73
<regul> <cell>	12	208	9.60
<role>	15	835	9.30
<provid>	9	321	9.15
<transcript>	19	1267	9.09
<drosophila>	6	147	8.39
<necessari>	6	153	8.18
<novel>	6	153	8.18
<cancer>	8	294	8.07
<interact>	17	1177	7.82
<modul>	7	238	7.65
<taken>	5	110	7.64
<SUBSTANCE> <regul>	15	501	7.62
<SUBSTANCE> <SUBSTANCE>	82	8888	7.56
<essenti>	7	244	7.51
<SUBSTANCE> <express>	30	1901	7.45
<high>	9	408	7.43

J.: Tuning support vector machines for biomedical named entity recognition, *Proc. of the Workshop on Natural Language Processing in the Biomedical Domain*, pp. 1–8 (2002).

[6] Narayanaswamy, M., Ravikumar, K. E. and Shanker, V. K.: A biological named entity recognizer, *Proc. of the Pacific Symposium on Biocomputing*, Vol. 8 (2003).

[7] Nobata, C., Collier, N. and Tsujii, J.: Automatic term identification and classification in biology texts, *Proc. of the 5th Natural Language Processing Pacific Rim Symposium*, pp. 369–374 (1999).

[8] Olsson, F., Eriksson, G., Franzen, K., Asker, L. and Liden, P.: Notions of correctness when evaluating protein name taggers, *Proc. of the 19th International Conference on Computational Linguistics* (2002).

[9] Pei, J., Han, J., Mortazavi-Asl, B. and Pinto, H.: PrefixSpan: Mining sequential patterns efficienty by prefix-projected pattern growth, *Proc. of the International Conference on Data Engineering (ICDE)*, pp. 215–224 (2001).

[10] Porter, M. F.: An algorithm for suffix stripping, *Program*, Vol. 14, No. 3, pp. 130–137 (1980).

[11] Seki, K. and Mostafa, J.: A Probabilistic Model for Identifying Protein Names and their Name Boundaries, *Proc. of the 2003 IEEE Bioinformatics Conference (CSB2003)*, pp. 251–258 (2003).

[12] Sese, J. and Morishita, S.: Answering the most correlated N association rules efficiently, *Proc. of the 6th Europiean Conference on Principles of Data Mining and Knowledge Discovery (PKDD)*, pp. 410–422 (2002).

[13] 磯崎秀樹, 平尾努, 鈴木潤: 機械学習のための組み合わせ素性の選択基準について, 情報処理学会研究報告, 2003-NL-158 (2003).

[14] 岸田和明: 検索実験における評価指標としての Mean Average Precision の性質, 情報処理学会研究報告 2001-FI-63, pp. 97–104 (2001).

[15] 久光徹, 丹羽芳樹: 組み合わせ的確率モデルに基づく特徴単語選択方法, 情報処理学会研究報告 2000-NL-140, pp. 85–90 (2002).

表 4: 抽出された上位 1,5,10,50,100 番の結果.

成績 1 位（問題番号 9: CD:100.00%, MUD:100.00%, BD:100.00%, BP:100.00%）
正解フレーズ：Regulation of intracellular pH mediates Bax activation in HeLa cells treated with staurosporine or tumor necrosis factor-alpha
出力フレーズ：Regulation of Intracellular pH Mediates Bax Activation in HeLa Cells Treated with Staurosporine or Tumor Necrosis Factor-alpha

成績 5 位（問題番号 10: CD:100.00%, MUD:100.00%, BD:100.00%, BP:100.00%）
正解フレーズ：Apocytochrome c blocks caspase-9 activation and Bax induced apoptosis
出力フレーズ：Apocytochrome c Blocks Caspase-9 Activation and Bax-induced Apoptosis

成績 10 位（問題番号 90：CD:96.55%, MUD:95.24%, BD:94.74%, BP:92.31%）
正解フレーズ：Activity in the nucleus accumbens shell controls gating of behavioral responses to emotional stimuli.
出力フレーズ：CREB activity in the nucleus accumbens shell controls gating of behavioral responses to emotional stimuli

成績 50 位（問題番号 69：CD:51.28%, MUD:48.00%, BD:17.39%, BP:14.29%）
正解フレーズ：there is a mechanically coupled transcriptional circuit that promotes binding of p38 to Sp1 in the nucleus
出力フレーズ：Interaction of p38 and Sp1 in a Mechanical Force-induced, beta1 Integrin-mediated Transcriptional Circuit That Regulates the Actin-binding Protein Filamin-A

成績 100 位（問題番号 33 :CD:31.82%, MUD:32.26%, BD:13.79%, BP:23.53%）
正解フレーズ：the JH2 domain contributes to both the uninduced and ligand-induced Jak-receptor complex, where it acts as a cytokine-inducible switch to regulate signal transduction
出力フレーズ：The Pseudokinase Domain Is Required for Suppression of Basal Activity of Jak2 and Jak3 Tyrosine Kinases and for Cytokine-inducible Activation of Signal Transduction

[16] 工藤拓, 山本薫, 坪井裕太, 松本裕治: 言語情報を利用したテキストマイニング, 情報処理学会研究報告 2002-NL-148, pp. 65–72 (2002).
[17] 平博順, 平尾努, 泉谷知範, 鈴木穣, 前田英作: 生物医学質問応答システム (bio-QA) の提案, 情報処理学会研究報告 2003-NL-154, pp. 109–116 (2003).