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

2004－ＦＩ－ 76　(4)
2004－ＮＬ－163　(4)
2004／9／16

一般語との曖昧性を持つタンパク質名の自動検出

大井洋子　大田佳宏　今一修　丹羽芳樹　久光徹
（株）日立製作所 中央研究所
〒185-8601 国分寺市東恋ヶ窪一丁目 280 番地
{h-ohi, yoh, imaichi, yniwa, hisamitu}@crl.hitachi.co.jp

本稿ではタンパク質名辞書に含まれるタンパク質名のうち一般語と曖昧性を持つタンパク質名（例えば “AND”, “CELL”, “SKI” など）を自動的に検出する方法を提案する。提案手法では，①公共データベースからタンパク質名を収集して初期辞書を作成し，次に，②初期辞書に含まれるタンパク質名からリファレンスコーパスにおける頻度が閾値より高いものを一般語と曖昧性の高いタンパク質名として検出する。閾値を変化させながら GENIA コーパスでのタンパク質名抽出を行って最適な閾値を決定した。性質の異なる 3 つのリファレンスコーパスで比較したところ，一般語の判定には AP 通信のような専門用語を含まないコーパスが適していることがわかった。また，MEDLINE における超高頻度語を（医学・生物学分野の）一般語として取り除くことでタンパク質名抽出の精度を，さらに向上させられることがわかった。また既存のタンパク質名抽出の方法と比較するために Yapex コーパスで評価した結果, 適合率 71.0%, 再現率 66.4%が得られ，提案手法のような簡便な方法で既存手法と同等の精度を達成することができた。

Detecting Common-Word Nicknames in a Protein-Name Dictionary

Hiroko Ohi, Yoshihiro Ohta, Osamu Imaichi, Yoshiki Niwa, and Toru Hisamitsu
Central Research Laboratory, Hitachi, Ltd.
Higashi Koigakubo, Kokubunji-shi, Tokyo, 185-8601, Japan,
{h-ohi, yoh, imaichi, yniwa, hisamitu}@crl.hitachi.co.jp

As well as their formal names, proteins are usually known by less formal names and acronyms, which are sometimes the same as very common words: these include ‘AND’, ‘CELL’, and ‘SKI’. Such terms create problems in the automatic extraction of information on protein-protein interaction and other automated text-processing tasks. In order to detect such troublesome common-word nicknames in a large-scale automatically constructed dictionary of protein names, we have developed a frequency-based method, in which a word having a frequency higher than a certain threshold within a corpus is considered a troublesome common word and marked accordingly. The effectiveness of the method was evaluated by using the dictionary to identify protein names in the GENIA test collection. And we compared the performance of our protein name detection method with that of a rule-based method, with the Yapex corpus as the target. We obtained 66.4% recall and 71.0% precision.

1. はじめに

医学・生物学分野も他分野同様，実験デザインや実験結果の新規性の確認，創薬ターゲットの絞込みに最新の知見を得るため文献を調査することが求められる。近年，実験技術の進歩に伴い，文献が指数関数的に増加し，現状のキーワード検索だけでは，求めている文献が検索されなかったり，求めていない文献が大量に検索されてしまい，必要な知見を得ることが困難な状況が起きている。これを解決するための技術として，情報抽出技術やテキストマイニング技術が脚光を浴びている。

そのような中，我々は文献からのタンパク質間相互作用情報の抽出を支援するために MEDLINE（米国 National Library of Medicine が 1960 年代から提供している約 1,300 万件の医学・生物学系の文献データベース）の抄録から，相互作用情報を自動抽出するシステムを開発してきた[1]。このシステムでは，文献から相互作用表現，例えば，「c-Jun N-terminal kinase ***phosphorylates*** peroxisome proliferator-activated receptor gamma1」という表現を抽出し，ユーザにタンパク質同士が相互作用しているという情報を提示する。図 1 に自動抽出システム全体の構成を示す。

このような情報抽出システムにおいては，タンパク質名抽出の精度が情報抽出システム全体の精度を左右する。タンパク質は，一般に正式名称やその同義語，略称など，平均 3～4 個の別名を持ち，多いものになると数十個の別名を持つ場合もある。例えば，正式名称が“Mitogen-activated protein kinase kinase kinase 7”であるタンパク質は，同義語として“Transforming growth factor beta activated kinase 1”，略称として“TAK1”，“MAP3K7”を持つ。

現在，このようなタンパク質名を認識する方法は，大きく分けて 3 つの方法がある。ルールによる抽出[2][3]，機械学習による抽出[4][5]，タンパク質名の辞書による抽出[6]である。

ルールによる抽出方法[2][3]は，タンパク質名の特徴から人手でルールを生成しタンパク質名らしき部分を抽出する方法である。ルールによる抽出では，ルール化を完全に行うことができれば網羅的にタンパク質名を抽出することができ，もしルールによる抽出で間違いがあったとしても構築したルールの調整は容易に行えるという利点がある。しかし，ルールの構築には労力がかかるという問題点や，どのタンパク質名が同義語関係であるかどうかを知ることができないという問題がある。

機械学習による抽出方法[4]では，特徴量として文字列のほかに品詞や共起する語の情報を使い，形態素に基づくチャンキングを SVM（Support Vector Machine）を用いて行いタンパク質名を抽出する方法を提案している。このような機械学習の方法は，ルール構築に比べ人手をかけずに行うことができるが，抽出が学習データに依存するため大規模な学習データを必要とすることやルールによる抽出と同様に同義語間の関係を知ることができないという問題がある。

タンパク質名辞書による抽出方法[6]では，大規模な公共のデータベースからタンパク質名を収集し，統合することによって辞書を構築し，抽出に用いている。辞書は新しい情報に合わせて更新する必要があるが，同義語間の関係情報を辞書に持てば，検索や情報抽出などでタンパク質に関する情報が網羅的に得られるため有用となる。これらのことから我々は辞書による方法に注目している。現在分かっているタンパク質は約 7 万個といわれており，各々が数

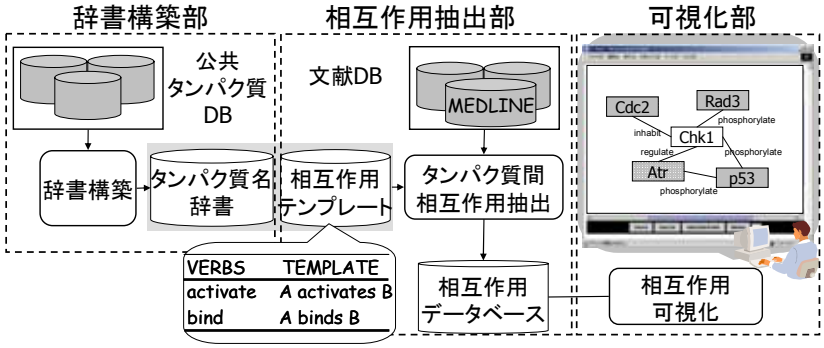

図 1：医学・生物学知識マイニングシステムの構成

個の別名を持つとすると，タンパク質名の総数は数十万語となる。したがって辞書を用いてタンパク質名を抽出するためには，名称を網羅したタンパク質名辞書が必要となる。しかし，まだ情報抽出の精度が十分なタンパク質名辞書は存在しないのが現状である。

ここで，精度が十分でない理由の１つとしては，タンパク質名辞書に，“WAS”，“AND”，“CELL”といった一般語がタンパク質名として含まれることが原因としてある。

一般語と曖昧性のあるタンパク質名（以下，このような語のことを要注意な多義語であるとして，要注意多義語と呼ぶ）は，タンパク質名としてよりも一般語として，医学・生物学文献に多く現れるため問題となる。

本稿ではタンパク質名辞書に含まれるタンパク質名のうち一般語と曖昧性を持つタンパク質名を自動的に検出する方法を提案する。提案手法では，①公共データベースからタンパク質名を収集して初期辞書を作成し，次に，②初期辞書に含まれるタンパク質名からリファレンスコーパスにおける頻度が閾値より高いものを一般語と曖昧性の高いタンパク質名として検出する。リファレンスコーパスの性質によって一般語の判定に差が出るかどうかを調べるために，性質の異なる 3 つのコーパスを用いる。各リファレンスコーパスに対して，頻度の閾値を変化させながら GENIA コーパス[7]でのタンパク質名抽出を行って最適な閾値を決定する。また既存のタンパク質名抽出の方法と比較するために Yapex コーパス[3]での評価を行う。

本稿では，2 章にタンパク質名辞書の構築について述べ，3 章では要注意多義語の検出方法とその評価結果について示し，4 章では考察を述べ，5 章で結論を述べる。

2. タンパク質名辞書の構築

2.1. 初期辞書の構築

本研究では，網羅的にタンパク質名を収集するために，複数の公共タンパク質データベースのタンパク質名が書かれている部分を参考にする。現在公共タンパク質データベースとして 5, 6 個のデータベースがあるが，これらの中で，信頼度が高く，多くの情報を持ったデータベースを情報の重なりが少ないように選択し，タンパク質名の収集を行った。選択したデータベースは，Swiss-Prot[1]，RefSeq[2]，PIR[3]である。

３つの公共タンパク質データベースから収集した名称数を表１にまとめる。

表 1: データベースから収集した名称の統計

データベース	エントリ数	名称数
Swiss-Prot	40,770	88,754
PIR	32,537	68,246
RefSeq	88,903	169,918

選択した３つの公共タンパク質データベースには，重複するタンパク質名が含まれている。そこで，データベース間のクロスリファレンス情報や正式名称の一致を参考にしてタンパク質名辞書エントリのマージを行い，71,518 エントリ，293,634 語のタンパク質名の初期辞書を構築した。

2.2. 初期辞書の問題

初期辞書を用いて MEDLINE abstract からのタンパク質名抽出の予備実験を行った。その結果，一般的に使われる語とスペルが一致するタンパク質名（例えば，“AND”や“CELL”）によってタンパク質名ではない文字列が誤って抽出されてしまうという問題がわかった。詳しく分析した結果このような名称には，二つのタイプがあることがわかった。一つは一般的に使われる語と一致する名称（“LIVER”，“DELTA”，“TOXIN”など），もう一つは，医学・生物学分野で一般的に使われる語と一致する名称（“binding protein”，“receptor”，“GTP”など）である。抽出の精度を上げるにはこれらの名称を検出あるいは除去する必要がある。

3. 一般語と曖昧性を持つタンパク質名の自動検出

初期辞書の問題を解決するためには，どのタ

[1] ジュネーブ大学，EBI（European Bioinformatics Institute）によって構築されたタンパク質アミノ酸配列のデータベース。http://us.expasy.org/sprot/

[2] 米国 NCBI（National Center for Biotechnology Information ）によって構築された参照配列のデータベース。http://www.ncbi.nlm.nih.gov/RefSeq/

[3] 米国ジョージタウン大学によって構築されたタンパク質アミノ酸配列のデータベース。http://pir.georgetown.edu/

ンパク質名が一般語として文献中で使われているかを知る必要がある。そこで，我々は大規模なコーパスに高頻度に現れる語を一般語であると考え，一般語と合致するタンパク質名を要注意多義語として検出することにした。

一般語の判定に用いるコーパス（以下，リファレンスコーパスと呼ぶ）として，

- BNC（British National Corpus）
  幅広い分野の書き言葉や話し言葉を集めたコーパス。約 1 億語。
- AP 通信（1989）
  新聞記事。約 5 千万語。
- MEDLINE abstract
  医学生物学文献の抄録を集めたデータベース。約 100 万語。

を用いる。これはリファレンスコーパスの性質によって一般語の判定に差が出るかどうかを調べるためである。また，どれくらいの頻度の語までを一般語とするかを決めるために，GENIA コーパスを用いてタンパク質名抽出の精度を測り，最も精度が高いところを頻度の閾値とした。

3.1. GENIA コーパスを用いた実験

実験の手順は以下のとおりである。

1. 各リファレンスコーパスから頻度付きの単語リストを作成
2. 頻度上位 x 語と辞書の語を照合
3. 合致した名称を辞書から除外
4. GENIA コーパスで精度評価

GENIA コーパスは MEDLINE で Mesh Term に“Human”，“Blood Cells”，“Transcription Factor”が含まれている 2,000 文書に，人手でタンパク質名や DNA, RNA 名，生物種・組織・器官などにタグを付与したタグ付きコーパスである。ここでは，“protein_molecule”や“protein_complex”のタグが付けられている語（延べ 23,908 語）を正解として評価実験を行った。

評価尺度には，Van Rijsbergen の E 尺度[8]を用いた。E 尺度の値は 0 から 1 の範囲にあり値が小さいほど良い。

$$E = 1 - \frac{1}{\alpha \frac{1}{P} + (1-\alpha)\frac{1}{R}}$$ (P:適合率, R:再現率)

α は再現率と適合率のどちらを重視するかを表すパラメタである。タンパク質名抽出では，

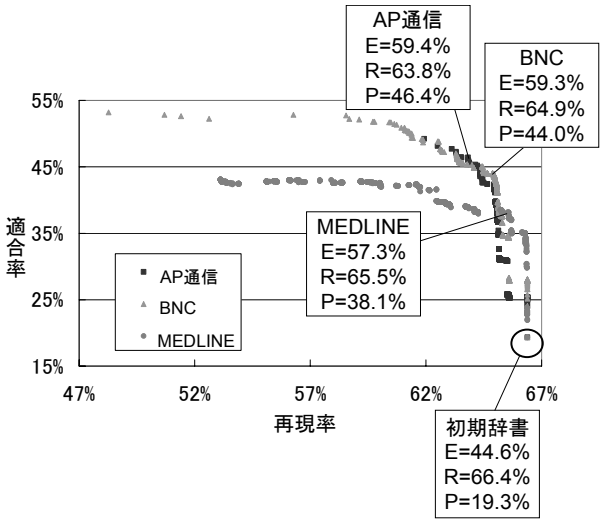

図 2：各リファレンスコーパスにおいて頻度の閾値を変化させた場合のタンパク質名抽出の精度

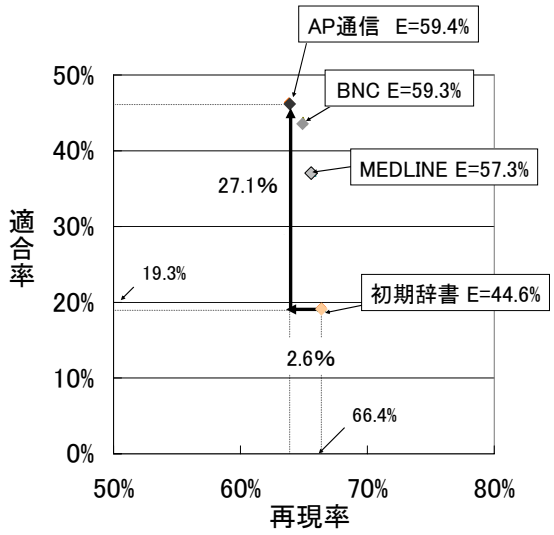

図 3：各リファレンスコーパスにおける最適点

検出漏れを防ぐことが重視されるので，再現率を適合率より重視するように，パラメタを 0.2 と設定した。

3.2. 各リファレンスコーパスの結果

図 2 は，BNC，AP 通信，MEDLINE において頻度の閾値を変化させた場合のタンパク質名抽出の精度である。右下に示された点が初期辞書の精度である。最適値は，AP 通信を用いた場合では，頻度 8 以上の語，BNC の場合では，頻度 164 以上の語，MEDLINE では，頻度 13,912 以上の語を除いた場合であった。

図 3 は各リファレンスコーパスにおける最適点をプロットしたものである。

3 つのリファレンスコーパスのうちＡＰ通信

を用いた場合に E 尺度で最も良い結果が得られた。この場合，初期辞書から再現率は 2.6%減少したが，適合率が 27.1%上昇した。

3.3. Yapex コーパスを用いた実験

上記の方法によって構築した辞書の有用性を確認するため, ルールを使ったタンパク質名抽出の方法である Kex[2]や Yapex[3]と性能比較を行った。評価には Yapex のテストコーパスを用いた。

Yapex コーパスは，200 の医学・生物学文献からなり, 文献中のタンパク質名部分にタグがつけられている。Yapex ではこのうちの 99 の文献をトレーニング用に用い, 101 の文献をテスト用に用いている。ここでは，Yapex でテスト用に用いられた 101 の文献を用いて評価を行う。このテスト用の文献のうち 48 文献は MEDLINE の文献で“interaction”，“molecular”の語と, Mesh タームに“protein binding”を含むものからなり，残りの 53 文献については GENIA コーパスからランダムに選択しタグを付け直したものである。これら 101 のテスト用文献には 1936 のタンパク質名が含まれている。評価は，文献[3]の方法に従い再現率，適合率，F 尺度を表 2 の評価基準に基づいて測定する。

表 2:タンパク質名抽出の評価基準

分類	評価基準
strict	システムの提示した答えが正解に正確にマッチした場合
left	システムの提示した答えの左の境界が正解の左の境界にマッチした場合
right	システムの提示した答えの右の境界が正解の右の境界にマッチした場合
sloppy	システムの提示した答えのどの一部でも正解のどこかにマッチした場合

表 3 に，AP 通信を使って改良した辞書によるタンパク質名抽出の結果とルールによるタンパク質名抽出の方法である Kex や Yapex の結果, また改良前の初期辞書の結果をまとめた。Kex と Yapex の評価値は文献[3]に記載されているものを用いる。

初期辞書ではどの評価基準においても再現率は高いが適合率が低くなった。提案した手法による改良を行うと, わずかに再現率は低くなるが，適合率は大幅に上昇した。F 尺度では初期辞書に比べ *strict* で 38.6%上昇した。また Kex や Yapex に比べ抽出精度が高く良い結果が得られた。

表 3: Yapex コーパスによる評価結果

	Kex	Yapex	提案手法	
			要注意多義語の除去前	要注意多義語の除去後
strict	R=41.1%	R=66.4%	**R=74.2%**	R=66.4%
	P=40.4%	P=67.8%	P=18.8%	**P=71.0%**
	F=40.7%	F=67.1%	F=30.0%	**F=68.6%**
left	R=62.6%	R=71.7%	**R=81.2%**	R=72.7%
	P=61.5%	P=73.2%	P=20.6%	**P=77.8%**
	F=62.1%	F=72.5%	F=32.9%	**F=75.2%**
right	R=49.9%	R=76.3%	**R=86.0%**	R=73.9%
	P=49.1%	P=77.9%	P=21.8%	**P=79.0%**
	F=49.5%	**F=77.1%**	F=34.8%	F=76.4%
sloppy	R=83.5%	R=82.1%	**R=89.2%**	R=80.5%
	P=82.1%	P=83.8%	P=22.6%	**P=86.1%**
	F=82.8%	F=82.9%	F=36.1%	**F=83.2%**

4. 考察

4.1. GENIA コーパスを使い各コーパスによって改良した辞書の結果の考察

図 4 に AP 通信を用いた要注意多義語除去によるタンパク質名抽出において要注意多義語数を変化させた場合の精度の変化を示している。第 2 軸は AP 通信の出現頻度となっている。点線と交差する部分が AP 通信での最適点を示している。出現頻度の閾値を 8 とすると 1371 語が要注意多義語となったことを示している。

AP 通信には専門用語はほとんど含まれておらずコーパスに含まれるほとんどすべての語は一般語として使われる。つまり AP 通信を使うことにより問題となる語を取り除くことができる。一方で AP 通信には一般的な医学・生物学用語はほとんど含まれていないため，医学・生物学用語と合致するタンパク質名については取り除くことができない。

図 4 に示した語は，除くことによって再現率や適合率に大きな変化をもたらした語である。例えば，“CELL”は 1815 回出現する語であり，“CELL” を除くことで大幅に適合率が上昇（約 10%）した。一方，“TAT”（軽い打撲という意味の語），“PLA”（Port of London Authority の略語），“DOD”（Department of Defense の略語）は MEDLINE では一般語として使われるより，タンパク質名として使われる場合が多くこれらの語を除いてしまうと再現率の低下が見られた。

次に図 5 に BNC を用いた要注意多義語除去によるタンパク質名抽出の精度を示す。

バランスドコーパスである BNC の高頻度語には一般語が含まれ，要注意多義語の除去が可能となる。また中頻度語には，医学・生物学分野の一般語が数少ないが含まれ，医学・生物学分野の要注意多義語除去も可能となっている。しかし低頻度語には，医学・生物学分野の専門語であるタンパク質名 （例えば“integrin”）が含まれてしまうため，低頻度語まで要注意多義語とすると必要なタンパク質名まで辞書から除いてしまい再現率が下がる。

図 6 に MEDLINE を用いた要注意多義語除去によるタンパク質名抽出の精度を示す。

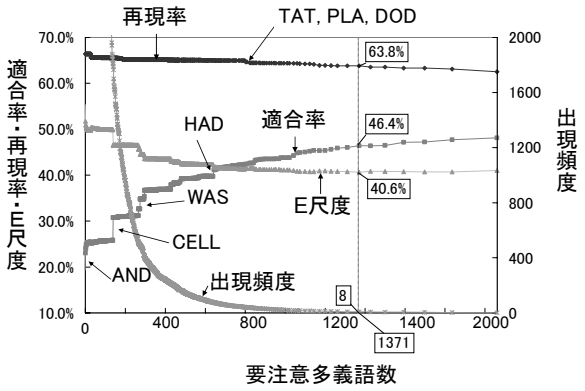

図 4：AP 通信による要注意多義語除去

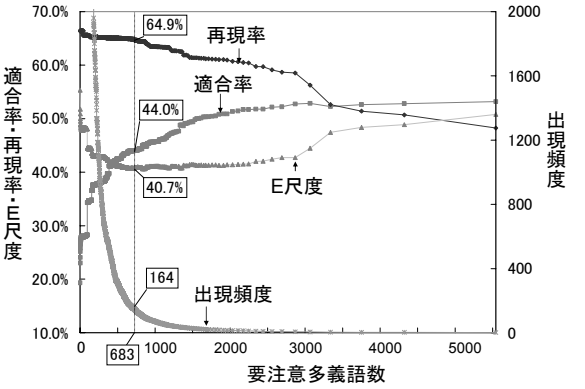

図 5：BNC による要注意多義語除去

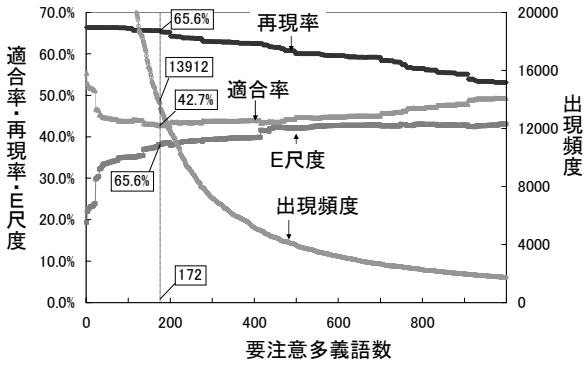

図 6：MEDLINE による要注意多義語除去

MEDLINE には一般語と専門用語が含まれている。超高頻度に表れる語を除くと AP 通信や BNC によるよりも効率の良い除去ができている。“PCR” や “antigen” や “peptide” など新聞には現れにくい語で特定のタンパク質名でもないような語が超高頻度領域に現れるからである。ところが，中頻度もしくは低頻度の語には多くの専門用語が含まれていて必要なタンパク質名を除くようになり再現率が下

がってしまう。

そこで AP 通信と MEDLINE の結果から，両方を使うことによってより良い結果が得られるのではないかと考え，MEDLINE に超高頻度に現れる語と AP 通信に表れる語を除く実験を行った。その結果 E 尺度にはそれほどの変化が見られなかったが，再現率が 0.7%減少したものの，適合率で 3.1%上昇した。

4.2. GENIA コーパスでのタンパク質名抽出のエラーの調査

再現率を落としている原因を調べるために，AP 通信で最も良い結果を示した辞書による GENIA コーパスの抽出結果を分析しエラーを表 4 の 3 種類に分類した。文献[3]の評価基準における *strict* がこの分類での正しく抽出できた場合に相当する。

表 4:抽出結果の分類

分類	抽出数
正しく抽出できた場合	14,916
部分的にしか抽出できなかった場合	4,570
全く抽出できなかった場合	4,422

部分的にしか抽出できなかったものの例としては，GENIA コーパスでは“Oct1 protein”という部分にタグが付いていて，辞書には“Oct1”のエントリしかなかったために，部分的にしか抽出できなかったものや，表記のゆれ（例えば，“NF-kappa B”と“NF-kappaB”）によって正しく抽出できなかったものが含まれていた。表記ゆれによるものは，延べ 1,476 語含まれていた。

また，全く抽出できなかった場合には，公共タンパク質データベースにタンパク質名が存在しなかった場合と本稿で提案した要注意多義語除去によるものがあった。要注意多義語によるものは，延べ 638 語含まれていた。しかしながら，要注意多義語として除去した単語の 98.8%（1355 / 1371）は，GENIA コーパスのタンパク質名タグ部分には現れず，要注意多義語除去によって少しの再現率の低下で適合率を大幅に上げることができ，全体としての精度向上ができた。

適合率を落としている原因としては，タンパク質名辞書がタンパク質名として抽出した部分に，GENIA コーパスではタンパク質名以外のタグ（例えば，protein_family_or_group, DNA-domain_or_region）が付加されている場合があげられる。表 5 は，タンパク質名辞書が抽出した部分に対して，GENIA コーパスがどのようなタグを付加しているか内訳を示した表である。この表を見るとタンパク質名辞書が抽出した部分の約 85%は，タンパク質名部分を抽出しているが，残りの約 15%はタンパク質名以外としてタグ付けされた部分を抽出している。

表 5:タンパク質名辞書が抽出した部分の GENIA コーパスのタグの種類

タグの種類	抽出割合
G#protein_molecule	74.1%
G#protein_complex	11.3%
G#protein_family_or_group	5.0%
G#DNA_domain_or_region	1.9%
その他	7.6%

4.3. GENIA コーパスと Yapex コーパスの差異

GENIA コーパスと Yapex コーパスで辞書を用いてタンパク質名抽出を行うと再現率にはそれほど差が見られないが，適合率で約 25% Yapex の場合の精度が良かった。

このことから GENIA コーパスは Yapex コーパスに比べタンパク質名としてタグを付けている部分が少ないことがわかる。

実際，Yapex コーパスでタグが付けられている部分が GENIA コーパスではタグが付けられていない場合や，遺伝子名やタンパク質ファミリー名としてのタグが付けられている場合がある。専門家がどの語をタンパク質名として認識するかに相違があると考えられる。

タンパク質名抽出を実用的に用いるには，システム全体の目的に合うようなテストコーパスを選んでいく必要があると考えられる。

5. 結論

本稿ではタンパク質名辞書に含まれるタンパク質名のうち一般語と曖昧性を持つタンパク質名を自動的に検出する方法を提案した。提案手法では，公共データベースからタンパク質名を収集し初期辞書を作成し，次に，初期辞書に含まれるタンパク質名からリファレンスコーパスにおける頻度が閾値より高いものを一般語と曖昧性の高いタンパク質名として検出する。閾値を変化させながら GENIA コーパスでのタンパク質名抽出を行って最適な閾値を決定した。性質の異なる 3 つのリファレンスコーパスを比較したところ，一般語の判定には AP 通信のような専門用語を含まないコーパスが適していることがわかった。また MEDLINE における超高頻度語を（医学・生物学分野の）一般語として取り除くことでタンパク質名抽出の精度をさらに向上させられることがわかった。

提案手法を用いて作成したタンパク質名辞書によるタンパク質名抽出の精度を Yapex コーパスで評価した結果，適合率 71.0%，再現率 66.4%が得られた。提案手法のような簡便な方法で，既存のタンパク質名抽出の方法と同等の情報抽出精度を持つ辞書を自動的に構築することができた。

今後，既存のタンパク質名抽出の方法の利点を統合しさらに精度の向上を目指したい。

謝辞

本研究に関して貴重なご助言をいただいた（独）産業技術総合研究所生物情報解析センタの夏目徹先生に感謝いたします。

参考文献

[1] Ohta, Y., Natume, T., Nishikawa, T., Ohi, H., Hisamitsu, T. 2003. “ExMI: Extracting Molecular Interaction from Large Biomedical Literature,” *The 11th International Conference on Intelligent Systems for Molecular Biology*.

[2] Fukuda, K., Tsunoda, T., Tamura, A., Takagi, T. 1998. “Toward Information Extraction: Identifying protein names from biological papers,” In *Proceedings of Pacific Symposium on Biocomputing*, pp.705-716.

[3] Olsson, F., Eriksson, G., Franzen, K., Asker, L., and Liden, P. 2002. “Notions of Correctness when Evaluating Protein Name Taggers,” In *Proceedings of the 19th International Conference on Computational Linguistics (COLING 2002)*, pp.765-771.

[4] Yamamoto, K., Kudo, T., Konagaya, A., and Matsumoto, Y. 2003. “Protein Name Tagging for Biomedical Annotation in Text,” In *Proceedings of the ACL 2003 Workshop on Natural Language Processing in Biomedicine,* pp.65-72.

[5] Tsuruoka, Y., and Tsujii, J. 2003. “Boosting Precision and Recall of Dictionary-Based Protein Name Recognition,” In *Proceedings of the ACL 2003 Workshop on Natural Language Processing in Biomedicine,* pp.41-48

[6] Hanisch, D., Fluck, J., Mevissen, H. T., and Zimmer, R. 2003. “Playing Biology’s Game: Identifying Protein Names in Scientific Text,” In *Proceedings of Pacific Symposium on Biocomputing*, pp.403-414.

[7] Ohta, T., Tateishi, Y., Mima, H., and Tsujii, J. 2002. “GENIA corpus: an annotated research abstract corpus in molecular biology domain,” In *Proceedings of the Human Language Technology Conference,* pp.73-77.

[8] Rijsbergen, V. C. J. 1979. Information Retrieval （2ed.） *Butterworths*.