自然言語処理　150－3
(2002. 7. 15)

免疫学文献からのたんぱく質相互作用情報抽出に向けて

保坂順子　梅津亮

理化学研究所　ゲノム情報科学研究グループ

〒230-0045 横浜市鶴見区末広町 1-7-22

{jhosaka, uryo}@gsc.riken.go.jp

生物学・医学文献、特に免疫学の文献から、たんぱく質間の相互作用情報を自動抽出するシステムの開発を進めている。その第一歩として、免疫学の専門家が PubMed から検索したアブストラクトから取り出した，

約１０００文について言語解析の専門家二人が評価した結果を分析したので、報告する。我々は、構文解析結果を基に、相互作用をあらわす単語を中心として、その相互作用要素を決定するという方式を取っている。そのため、評価者には情報が的確に抽出されているか、構文解析は正しいかという２点からの評価を依頼した。今回は動詞 "activate" に関係する相互作用要素を抽出した。本稿では、抽出に関する判断が一致したものの内、２２２文について分析を行った。また、そのうち構文解析が誤っているという判断が一致した５９文について、さらに詳しく検討した。その結果、専門用語の認定が困難であるということが確認された。また、並列句においては、同一主語に対して異なる態が使われるのがめだった。

Toward the extraction of protein-protein interaction information from immunology literature

Junko Hosaka　　Ryo Umetsu

Suehiro-cho 1-7-22, Tsurumi-ku, Yokohama, Kanagawa, 230-0045 Japan

{jhosaka, uryo}@gsc.riken.go.jp

Extracting protein-protein interaction from biological and medical literature is one of the main topics in Bioinformatics. We are developing such an IE system using a syntactic parser. Two language specialists evaluated our automatic extraction as well as the parsing result for 1000 sentences retrieved by an immunologist. Out of them we randomly selected 222 sentences on which the evaluators agreed for the extraction. Out of the sentences we analyzed 59 sentences on which the evaluators agreed for the parsing incorrectness. Our analysis confirms the difficulty of term recognition. It is also noticeable that the voices for a single subject vary.

1. はじめに

たんぱく質間相互作用抽出に代表される、生物学・医学文献からの情報抽出が近年盛んに行われている。単語の共起を使ったもの [1]、フルパーザを使ったもの [2]、抽出規則を人手で書き下したもの [3]、医学文献用に開発したパーザを、分子生物学用に変更を加えたもの [4] などがある。

我々は、生物学・医学文献、特に免疫学の文献からたんぱく質間の相互作用情報を自動抽出するシステムの開発を進めており、絞り込んだ抽出を目指している。この目的を達成するには、単語の共起だけでは情報が十分でない。深い解析をするには、時間と労力がかかりすぎ、抽出規則を網羅的に人手で記述するのは困難である。また、医学文献用に開発されたパーザは、入手困難である。

そこで我々は、構文解析パーザを使い、その結果を基に抽出を行っている。現在は、ニューヨーク大学で開発された ApplePie Parser ver.5.9[1]（再現率：77.5 パ

[1] http://www.cs.nyu.edu/cs/projects/proteus/app/

ーセント、適合率：75.58 パーセント）を使っている。これは、一般的なトピックを扱うように開発されたものであるので、我々は、このパーザに生物学・医学文献も扱えるように改良を加える予定である。

なお、本稿で取り上げた例文や、実験に使った文は、すべて免疫学の専門家が PubMed[2] から検索したアブストラクトを基にしている。また、例文は生物学者により、内容の正しさが確認されている。

2. システムの概要

開発中のシステムは、PubMed などの文献検索システムからの出力や、MAPI[3]などのデータベースに格納されている文を入力として想定している。PubMed は、米国立医学図書館 (NLM) が提供する実験的な生物学・医学の文献サービスで、MAPI は、理化学研究所で開発中のたんぱく質間相互作用に関する文を集めたデータベースである。

システムは、この入力を受け、文分割などの前処理を行い、構文解析パーザにかけ、その出力結果を基に、"activate" などの相互作用に関する単語を中心に、相互作用する要素を抽出する、という構成である。図 1 にシステム構成を示す：

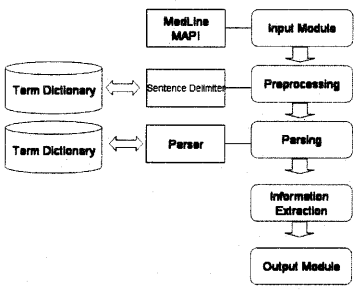

図 1：システム構成

2.1. 辞書の拡大

生物学・医学の文献では生化学、医化学、生命科学などに関する専門用語が多数使われている。専門用語には、"receptor" のようなシングルワードだけではなく、"allosteric binding sites" などのマルチワードも多数含まれ、用語の区切り判定が困難である。そこで、我々は、ApplePie Parser が持っている辞書に、オンライン専門用語辞典で扱っている用語を追加した。パーザ処理には、形態素情報が必要であるため、抜き出した用語には、その用語の定義を参照して、半自動的に形態素情報を付与した。以下、用語追加のために参照した辞典、およびその中から使用した用語数を示す：

biochemical glossary[4] (約 800 用語)
Cancernet Dictionary[5] (約 2,500 用語)
Medical Chemistry Dictionary[6] (約 100 用語)
Life Science Dictionary[7] (約 28,000 用語)
MeSH 用語[8] (約 120,000 用語)

MeSH (Medical Subject Headings) は、米国立医学図書館 (NLM) が提供する制限された語彙のシソーラスである。今回は、ここから物質名のみ使用している。

2.2. 処理速度

2.1. で示した辞典で扱われている用語をパーザの辞書に取り入れ、リナックスマシン（Memory: 1GB, CPU: Pentium3, 850MHz）で、処理速度を測定した。表 1 に処理速度、平均単語数を示す：

表 1　：パーザの処理速度

総文数	一文の平均単語数	処理時間	一文の平均処理時間
5083　文	22.6 単語	1 時間 2 秒	0.71　秒

2.3. 追加用語数とその使用頻度

表2 にApplePie Parser (APP) の辞書で扱っている単語数、MeSH 以外の追加辞書の用語数を示す：

表2：追加用語数

	単語・用語数			
APP	88707			
追加用語延べ数	31241			
追加用語ユニーク数	30566			
追加用語中APPにないもの	29808			
	マルチワード	13157 (44%)	シングルワード	16651 (56%)

表 2 から、シングルワード数は、マルチワード数に比べて 12 パーセントしか多くないことがわかる。用語の境界決定が、特にマルチワードで難しいということを考えると、表 2 の数字は、この問題に慎重に対処しなければならないということを示している。

[2] http://www.ncbi.nlm.nih.gov/entrez/query.fcgi

[3] http://big.gsc.riken.go.jp/BKDtext.htm

[4] http://www.fhsu.edu/chemistry/twiese/glossary/biochemglossary.htm

[5] http://www.cancer.gov/dictionary/

[6] http://www.chem.qmw.ac.uk/iupac/medchem/

[7] http://lsd.pharm.kyoto-u.ac.jp/index.html

[8] http://www.nlm.nih.gov/mesh/meshhome.html

追加用語のうちでパーザ辞書にはないものが、われわれが扱う分野の文献で、どのくらい使われているか調べた。表 3 に追加用語のソースと、免疫に関する 954 文中で使われているのユニークな用語数を示す：

表 3：追加用語ソースと使用数

追加用語ソース	用語数（ユニーク）
MeSH 以外	3395
MeSH	1193

MeSH 以外では、29808 用語中 3395 用語（11 パーセント）が使われている。また、MeSH では、パーザで扱っている単語を含まないと仮定すると、約 120,000 用語中、1193 用語（約 1 パーセント）が使用されている。使用頻度からは、MeSH 対象外にしてよいと考えられるが、これは、それぞれの用語の認定が文の解析にどの程度有効に働いているかを検討する必要がある。

3. 評価実験

動詞の activate (activates, activated, activating) を含む句を中心に、その動作主、被動作主の自動抽出を行った。約 1000 文を基に、抽出が正しいか、および構文解析が適切かの評価を外部の言語解析の専門家に依頼した。

3.1. 評価手順

作業は、評価者 A、B の二人が担当し、10 日間、一日 100 文のペースで、以下の手順で行った。担当者は、ApplePie Parser で使っている形態素情報に関する知識を有している：

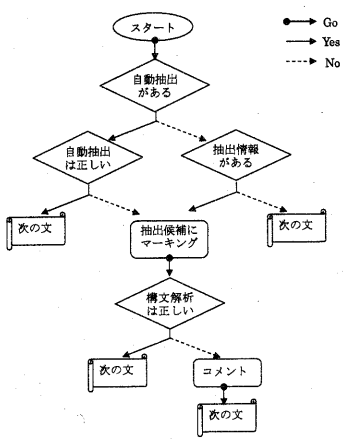

図 2：評価手順

構文解析の評価は、情報抽出が失敗したときのみ行っている。これは、抽出は構文解析結果を基に行っているので、抽出結果が正しければ解析結果も正しいと推測するからである。また、解析結果が正しいかどうかの評価は、抽出に直接関係する部分に限定した。これは、我々の目的が構文解析の精度向上にあるのではなく、情報抽出の精度をあげることにあるためである。我々は、抽出に関係する部分の精度が向上すれば、全体的な精度も上がると考える。構文解析が正しくないと評価者が判断した場合、その理由をコメントとして記述するように依頼した。

3.2. 評価用ツール

構文木と自動情報抽出を同時に検査できるように、以下の評価ツールを開発した。

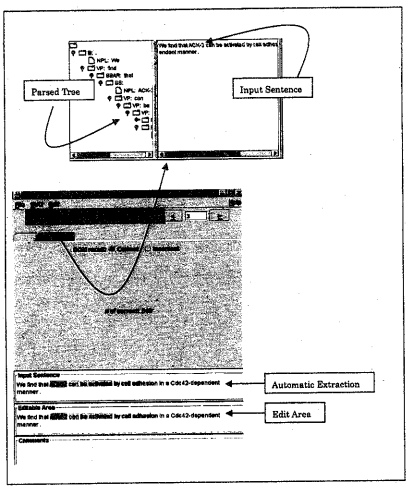

図 3：評価ツール

自動抽出結果は、動詞句、動作主、非動作主を色別に表示し、これに修正を加える場合は、エディット用ウィンドウで、同様にカラーリングできる。図 3 に示すように、評価者が抽出結果は誤りだと判断した場合は、構文木（Dom Tree 表示）をチェックし、さらに "correct/ incorrect" を選択した。

4. 評価結果の分析

情報抽出に対して、両評価者 A、B が同じ評価をしたものからランダムに選択した 222 文を対象に、その構文解析の分析を行う。評価者が、パーザの解析結果をどのくらいの割合で正解・不正解と判断したか、その内訳を表 4 に示す：

表4：構文解析の判定

対象文数	222
A が正解とした文数	127
B が正解とした文数	152
A、B ともに正解とした文数	114
A が不正解とした文数	95
B が不正解とした文数	70
A、B ともに不正解とした文数	59

4.1. データの信頼性とタスクの困難さ

データの信頼性を調べるため、カッパ値（K）を、以下の式により求める[5]。これは、偶然の一致を考慮した判定者間の一致性の指標で、主観が入る判定が複数の観察者の間で、または同一観察者では複数回の判定間で、どの程度一致するかを知るものである。我々は、この値は、タスクの難易度を測る目安にもなると考える：

$$K = \frac{P(A) - P(E)}{1 - P(E)}$$

P(A)　判定者間の判断が一致している割合
P(E)　判定者間の判断の一致が偶然起きる割合

今回実施した構文解析の正解・不正解の判定では、P(A)、P(E)は、以下の式で求めた：

$$P(A) = \frac{\text{正解の一致数} + \text{不正解の一致数}}{\text{文の総数}}$$

P(E) = 正解の偶然一致率＋不正解の偶然一致率

正解・不正解の偶然一致率は、以下の式で求めた：

$$\text{正解の偶然一致率} = \frac{A(T)}{\text{文の総数}} * \frac{B(T)}{\text{文の総数}}$$

$$\text{不正解の偶然一致率} = \frac{A(F)}{\text{文の総数}} * \frac{B(F)}{\text{文の総数}}$$

A(T) = Aが正解と判定した文数
B(T) = B が正解と判定した文数
A(F) = A が不正解と判定した文数
B(F) = B が不正解と判定した文数

表 4 から、P(A)、 P(E)、カッパ値 K は以下のように算出される：

P(A) = 0.78
P(E) = 0.52
K = 0.54

カッパ値は、K = 1 になった場合、完全な一致となり、以下、次のような一致とみなされる[6]：

0.81-1 "near perfect"、 0.61-0.80 "substantial"、 0.41-0.60 "moderate"、0.21-0.40 "fair"、0-0.20 "slight"、K < 0 "poor"。

今回実施した構文解析に関する判断の一致度は、この基準では、"moderate" の範囲にある。信頼性は中程度であるが、一方、我々は、今回のタスクは言語分析の専門家にも難しいのだと考える。

4.2. 構文解析誤りの分析

評価者 A, B ともに構文解析結果は不正解だと判断した文について、その原因を調べた。文は、ピリオドまでを一文とみなしており、ピリオドなしで２つ以上の文がつながっている場合は、これも一文とした。一文中に不正解の原因が２つ以上ある場合は、これらをすべて考慮した。表 5 に、構文解析の失敗の理由として考えられる原因と、その出現頻度を示す：

表 5：構文解析誤りの原因とその誤り頻度

不正解の原因候補		頻度
専門用語区切り		11
品詞付与		10
関係代名詞・先行詞		6
従属節の係り方		6
並列	動詞句	5
	名詞句	5
	前置詞句（by）	5
	文	1
名詞修飾の動詞の過去分詞		5
同格		1
その他		9

4.3. 構文解析誤りの例

表 5 で分類した不正解の原因を含む例を示す。ここで示す例文は、生物学者が内容の正確さを確認している。一文を読んだだけでは、内容が正しいかどうか判断できないこともあるが、その場合は、その文が使われているアブストラクトから判断している：

例文１は、専門用語区切り、品詞の付与に失敗している：

例文１：Herpes simplex virus type 1 (HSV-1) superinfection of CV-1 cells weakly transactivated a plasmid-borne metallothionein 1 (MT-1) promoter, but activated the expression of a marker gene controlled by an authentic HSV-1 promoter to a high level.

ここでは、"Herpes simplex virus type 1" がひとつの用語であるが、パーザはこれを誤認定している。また、"virus type" は名詞であるが、パーザは動詞として扱っている。図 4 に例文 1 の解析木を示す：

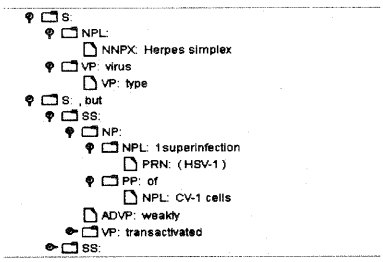

図 4：例文 1 の構文木

例文 2 は、関係代名詞の認定、前置詞句の並列の認定に失敗している：

例文 2 : However, the Pak1 (H83,86L) mutant that does not bind Rac or Cdc42 is activated in the absence of GTPase by alphaPix155-545 and by a mutant of alphaPix155-545 that no longer has exchange factor activity.

例文 2 は、関係代名詞の "that" を含むが、パーザはこれを "mutant" と共に形容詞句をつくると解析している。従って、"does not bind" の主語は "Pak1" になっている。また、活性化する 2 つの動作主 "alphaPix155-545" と "a mutant of alphaPix155-545" を、パーザは異なる文の要素として解析している。図 5 に例文 2 の解析木を示す：

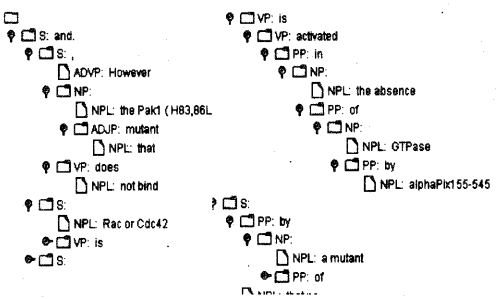

図 5：例文 2 の解析木

例文 3 は、従属節の係り先の同定に失敗している：

例文 3 : Infectious virus is not activated in the majority of substrains (Mov-4 to Mov-8 and Mov-10 to Mov-12), whereas the other mice develop viremia.

例文 3 では、"whereas" 以下は、主節全体にかかるが、パーザはこれを前置詞句で、"majority" を修飾するものとして解析している。図 6 に例文 3 の解析木を示す：

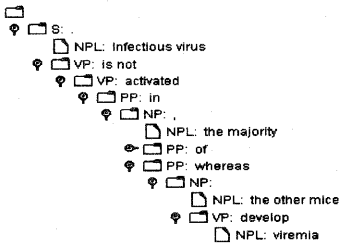

図 6：例文 3 の解析木

例文 4 は、動詞句の並列、名詞修飾の動詞の過去分詞の認定に失敗している：

例文 4 : We propose that the Pkc1 pathway promotes bud emergence and organized surface growth and is activated by Cdc28-Cln1/Cln2 at the G1-S transition and by Hcs77 upon heat shock.

例文 4 では、"that-clause" のなかで、"promotes" と "is activated" という 2 つの動詞句が使われているが、パーザは "is activated" は、"that-clause" の外にあると解析している。また、"organized" は、"surface growth" を修飾しているが、パーザは "is activated" と同レベルの動詞句として解析している。この例が示すように、生物学の文献では、一文中で、ある同一の主語に対する複数の動詞が能動態と受動態のように態を変えて使われることが目立つ。表 5 で示した動詞句の並列認定の失敗例では、5 文中 2 文で、異なる態が使われていた。図 7 に例文 4 の解析木を示す：

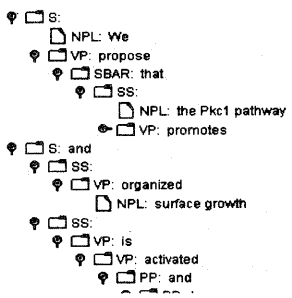

図 7：例文 4 の解析木

例文 5 は、同格の認定に失敗している：

例文５：ACK-2 could not be activated by addition of glutathione S-transferase (GST) -Cdc42 (Q61L), a GTPase-defective mutant, or by GTPgammaS-loaded GST-Cdc42 in in vitro kinase assays.

例文５では、活性化させる動作主は "addition of GST-Cdc42" と " GTPgammaS-loaded GST-Cdc42" の２種類で、"GTPase-defective mutant" は "addition of GST-Cdc42" と同格であるが、パーザは "GTPase-defective mutant" も動作主として解析している。図 8 に例文５の解析木を示す：

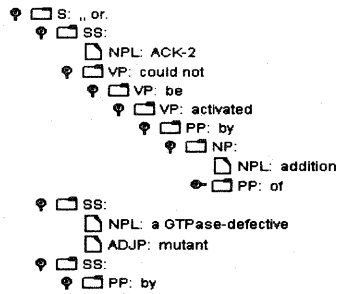

図 8：例文 5 の解析木

表 5 の「その他」は、To-不定詞の副詞的用法を名詞句と誤認定する、副詞的に使われている "both" を名詞とするなどを含む。

5. おわりに

たんぱく質間相互作用の自動抽出および構文解析の評価を言語解析の専門家二人に依頼し、その評価の分析を試みた。本稿では、評価者の抽出に対する判断が一致したもののうち、222 文をランダムに選び、そのうち評価者が一致して構文解析が不正解だとした 59 文について分析した。

例文４に示したように、同一主語に対する異なる態の使用がめだった。これが構文解析失敗の直接の要因かどうかの判定には、さらに評価者 A, B がともに正解とした文の分析を行う必要がある。

表 5 に示したように、品詞付与の誤りが多い。これは、生物学・医学の文献で、一般的な文章とは異なる構文が使われていることを示唆していると思われる。この改良には、より詳細に解析手順を調べる必要がある。

専門用語の区切りの判定は、分野の専門家でないものには困難である。また、すでに存在する専門用語辞典を参照するだけでは、日々増え続けている生物学・医学に関する専門用語には対応しきれない。我々は、この問題に対処するため、用語認識の研究も行っている。

文　献

[1] Jenssen, T-K., et al.: "A literature network of human genes for high-throughput analysis of gene expression", Nature Genetics, Vol.28, pp.21-28, 2001

[2] Yakushiji, A., et al.: "Event extraction from biomedical papers using a full parser", Proc. of PSB-2001, Vol.6, pp.408-419, 2001

[3] Blaschke, C. and Valencia, A.: "The potential use of SUISEKI as a protein interaction discovery tool", Genome Informatics, Vol.12, pp.123-134, 2001

[4] Friedman, C., et.al.: "GENIES: a natural –language processing system for the extraction of molecular pathways from journal articles", Proc. of ISMB-2001, Vol.17 Suppl.1, pp.S74-S82, 2001

[5] Cohen,J.: "A coefficient of agreement for nominal scales", Educational and Psychological Measurement, Vol.20, pp.37-46, 1960

[6] Carletta,J., et al.: "The reliability of a Dialogue Structure Coding Scheme", Computational Linguistics, Vol.23, No.1, pp.13-31, 1997