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

2004－NL－162 (5)
2004／7／15

文献からの薬物相互作用情報の抽出に向けて

保坂順子 吉川澄美 松村和美 小長谷明彦

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, sumi, kmkazumi, konagaya}@gsc.riken.jp

我々は、生物・医学分野の学術文献から薬物相互作用情報を自動的に抽出することをめざしている。たんぱく質間相互作用抽出に代表される同分野での情報抽出は、作用関係を表す動詞を中心に行われることが多い。しかし、関係をあらわす表現は多様であり、一品詞を扱うだけでは不十分だと考える。今回、薬物相互作用に関する英語の参考書の要約部分から薬学の専門家が選択した３０文を使い、薬物の代謝という観点から抽出を試みたので、これを報告する。抽出は、構文解析木に xPath および構文木上の位置、レベル情報を属性値として指定して行い、その規則作成には構文情報に基づく情報抽出システムの開発ツール PBIE2 を使った。

Towards Extraction of Drug Interaction from Scientific Texts

Junko Hosaka, Sumi Yoshikawa, Kazumi Matsumura, Akihiko Konagaya

Riken Bioinformatics Group

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

{jhosaka, , sumi, kmkazumi, konagaya}@gsc.riken.jp

We are extracting drug interaction information from scientific biomedical literature. In the domain information extraction on protein-protein interaction is the best studied. Most methods use the verbs to be regarded as expressing interaction relationship. However, such relationship cannot be expressed with single part of speech only. Focusing on drug and its metabolism, we conducted a preliminary extraction study with 30 sentences selected from abstracts in the reference book on drug interactions by a domain specialist. Our extraction uses xPath expressions and some attribute-value pairs designating for, example, level and position on the parse tree. For the rule construction we used our tool PBIE2 that supports building a parse-based information extraction system.

1. はじめに

生物学・医学の分野では、増大する文献を有効利用するため、たんぱく質間相互作用の情報抽出に代表される自動処理化がすすめられている。パーザを使ったもの、テンプレートを人手で作成したものなどがあるが、いずれも作用関係を表すと思われる動詞を中心に抽出が行われている[1][2][3][4]。しかし、関係をあらわす表現は多様であり、一品詞を扱うだけでは不十分だという報告もある[5]。このような状況を把握し、自然言語処理のための分子生物学分野に関する共通コーパスを作成している GENIA プロジェクトでは[6]、今後名詞化した関係表現もアノテートする予定だとしている[7]。一方、我々は薬物の代謝という観点から薬物相互作用に関する情報の自動抽出を試みている。

薬物相互作用とは、薬の飲み合わせによ

って引き起こされる反応のことを言う。この反応は、薬物と生体物質との相互作用を介して引き起こされることが知られている。従って、薬物が体内でどのように代謝されるかというのは、有害な飲み合わせを防ぐためにも、また創薬にも重要な情報である。

薬物に関する情報抽出は、癌の治療にかかわるものが遺伝子および細胞との関係で報告されている[8]。これは、形態素解析と構文解析を行ったあと、複数の辞書を使い薬物などの認定を行い、さらに意味情報を使い相互作用を抽出するシステムである。

我々の抽出は、構文解析を行い、その構文木に xPath[1]と属性値指定する抽出規則を適用して行っている。情報抽出規則の作成には PBIE2 (Parsing-based Information Extraction) を使った[9]。このツールは、構文情報に基づく情報抽出システム開発のために、生物学者と言語学の専門家の要求を反映して作成したものである。なお、PBIE2 ではパーザは外部のものをプラグインして使用できるようになっており、今回の実験には、ニューヨーク大学で開発された Apple Pie Parser[2] （再現率：77.5 パーセント、適合率：75.58 パーセント）を使用した。これは、Penn Treebank を学習に使っている。

本稿では、まず、薬物相互作用に関する英語の参考書から薬物相互作用が表現されている文集合を選択した。次に、その抽出箇所にマーキングしたものを正解として、構文木の一部を人手で修正しながら抽出規則を作成した。さらに、専門用語を追加した辞書(拡張辞書)を用意し、パーザ付属の辞書(基本辞書)を使った場合との抽出比較実験を行った。

2. 文集合の選択

情報抽出規則の作成には、Stockley's Drug Interactions[10]の要約を使った。これは、現在最もよく使われている薬物相互作用に関する英語の参考書のひとつである。漢方薬も含め、一般的に薬と呼ばれるもの全般、食物、飲み物、農薬なども扱っており、2,500 あまりの相互作用をカバーしている。

文の選択は、以下の手順で分野の専門家が行った：

1. Stockley's Drug Interactions の要約を半自動的に文分割した。
2. 1.の文集合から、文字列として M/metabolism を含む文を選択した。
3. 2.の文集合から相互作用が表現されているものを選択した。

このようにして選択した文は 30 あった。この文集合からの例を示す。イタリックの部分は、薬物相互作用を表す抽出候補である。表 2 に抽出カテゴリーを示すが、ここではこの分類は省略する：

(1) Limited evidence suggests *erythromycin* may *inhibit* the *metabolism* of *propafenone*.

(2) *Grapefruit juice* delays the absorption of *quinidine* and *reduces its metabolism* to some extent, but no clinically relevant adverse interaction seems to occur.

(3) A case is reported of sedation, confusion and respiratory depression attributed to the *inhibition* of *methadone metabolism* by *ciprofloxacin*.

(4) The *metabolism* of *phenytoin* is *unchanged* by *zileuton*.

(5) Imatinib increases serum simvastatin levels and is predicted to interact with *other drugs* whose *metabolism* is *affected* by *CYP3A4 inhibition*.

表 1に文集合を構成する単語数を示す。単語は、スペースで区切られたものとした。異なり語は、大文字小文字は同一とみなした。活用語の場合は、原形が同じ場合は同一とみなし、品詞が異なるものは、別単語とした。また、 “can, could” など、モダリティーにかかわるものは、別単語とした：

文数	総単語数	異なり語数
30	486	204

表 1：文集合の単語数

表 1から、1 文は平均 16 単語からなっていることが分かる。

[1] http://www.w3.org/TR/1999/REC-xpath-19991116

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

3. 薬物相互作用表現の抽出

Stockley's Drug Interactions から選択した30 文のうち、ランダムに選んだ 10 文を使い、薬物の代謝を表す文字列にドメインの専門家がマーキングをした。このマーキングは、専門家があらかじめ準備した 7 つのカテゴリーに沿って行われた。これらのカテゴリーと、使われた頻度を
表 2に示す。頻度は、専門家がマーキングした 10 文と、最終的に処理した 30 文での使用である：

抽出カテゴリー	頻度	
	10 文	30 文
Metabolism	10	30
InteractionWord	10	30
Biomolecule	0	0
Drug	11	31
Enhancer	3	3
Inhibitor	7	27
Product	0	0

表 2：抽出カテゴリーとその使用頻度

表 2から、"Metabolism, InteractionWord, Drug, Inhibitor" の関係を表している文が多数であることがわかる。

3.1. 情報抽出規則作成ツール

抽出規則の作成には PBIE2 を使った。このツールでは、2 種類の構文解析結果と情報抽出結果が比較評価でき、さらに抽出規則の作成ができる。これらはダイナミックに連動していて、特に、作成した抽出規則を文に適用すると、マーキングウィンドウの該当箇所と、構文木の対応箇所が同じ色で自動的にマーキングされる機能は、構文解析結果を利用した情報抽出規則を作成するのに有用である。

図 1に抽出規則の編集ウィンドウで、編集メニューを呼び出したところを示す：

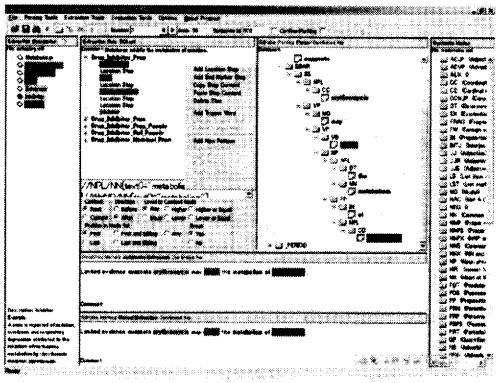

図 1：PBIE2 抽出規則編集ツール

編集ウィンドウは 3 つの部分から構成されている。上部では、規則の構成を定義する。規則はひとつ以上のパターンの集合からなっていて、それぞれのパターンはステップからなっている。ステップは、End Marker Step や Location Step のような直接抽出には使わない特殊なものは、編集用のメニューから選択し、それ以外は、編集ウィンドウの左側にある Extraction Categories のカテゴリーをドラッグして作成する。コピー、貼り付け、削除などは、編集用のメニューをマウスボタンのクリックで呼び出して行う。編集ウィンドウの中央部では、xPath で抽出候補部分を指定する。下部では、構文木上の方向、階層などを指定する。

3.2. 抽出規則の作成

PBIE2 の抽出規則編集ツールを使い、専門家により正解抽出文字列がマーキングされた 10 文をもとに、構文解析木からの抽出規則を作成した。このとき作成したのは、2 種類であった。これらは、薬物が前置詞を伴って使われているもの、および作用語が名詞化されているものであった。この 2 つの規則を残りの 20 文に適用し、抽出を行った。その結果を、専門家に評価してもらい、さらに規則を追加した。追加した規則は、受動態の規則と、薬物が代名詞を伴って使われているものであった。なお、構文木は、必要な箇所には人手で修正を加えた。このようにして、最終的に 5 種類の規則を作成

した。 表 3に抽出規則名と、これらの規則が適用される文数を示す。適用文数の後のカッコつき数字は、2節で引用した例文の番号であり、各規則の例文となっている：

抽出規則名	適用文数
Drug_Inhibitor_Prep	23(1)
Drug_Inhibitor_Pron	3(2)
Drug_Inhibitor_Nominal_Pron	1(3)
Drug_Inhibitor_Prep_Passiv	2(4)
Drug_Inhibitor_Rel_Passiv	1(5)

表 3: 抽出規則とその適用文数

PBIE2 には、抽出規則の選択方法は 2 種類ある。すべての規則を適用して、自動的に最適解を選択するものと、ユーザが文ごとに抽出規則を選択するものである。表 1では、後者の規則適用文数を示している。

3.3. 抽出規則の例

表 3から、Drug_Inhibitor_Prep が最もよく使われている規則であることが分かる。この規則の一部を、例文（1）を使い示す。

例文（1）では、 “propafenone” の “metabolism” が前置詞 “of” を使って表わされている。 この構文木を図 2 に示す：

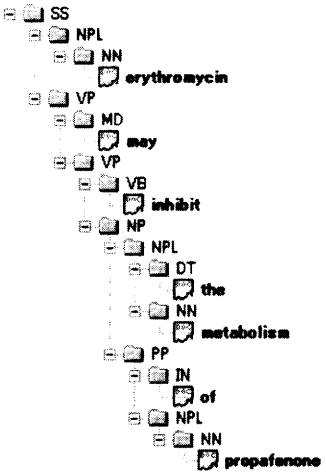

図 2: 前置詞”of”の後に Drug がある構文木例

例文（1）では、 “of” は抽出しないので、位置を指定し、 次に Drug の抽出を行う。

図 3に、“of” を指定する Location Step の規則を示す。Drug_Inhibitor_Prep は、8 つのステップから構成されていて、表 2 のカテゴリーの他に、この“of” 指定のような 4 つの特殊ステップを含んでいる。ここでは、2番目のステップで“of” を指定している：

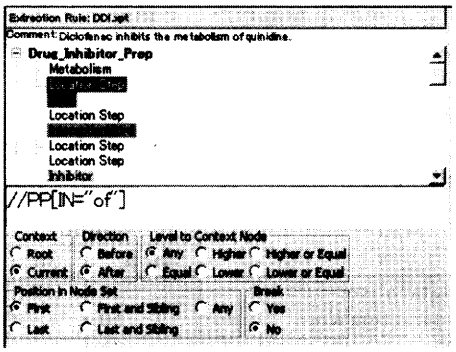

図 3：前置詞”of” を指定するステップ規則

XPath で、前置詞句の中の前置詞を指定し、さらに一つ前のステップから見た (Context=Current) 方向 (Direction=After)、レベル (Level to Context Node=Any) などを属性選択して絞り込んでいる。これは、XPath だけでは候補が複数あがる可能性があるからである。次に、PBIE2 で作成した Drug を抽出する規則を図 4に示す：

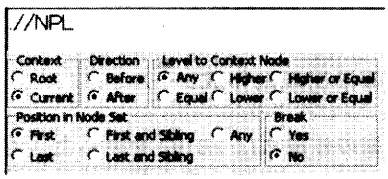

図 4：Drug を抽出するステップ規則

図 3の規則同様、XPath で名詞句を指定し、さらに一つ前のステップから見た (Context=Current) 方向 (Direction=After)、レベル (Level to Context Node=Any) などを属性選択して絞り込んでいる。

この二つのステップのデータ構造を示す：

```
<step name="Location Step" category="location_step" context="current" direction="after" level="any" position="first" break="no">
 <stepxpath>//PP[IN="of"]</stepxpath>
 </step>
```

```
<step name="Drug" category="Drug" context="current" direction="after" level="any" position="first" break="no">
 <stepxpath>.//NPL</stepxpath>
 </step>
```

このようにして記述した抽出規則を構文解析木に適用した結果を、図 1に示している。

4. 異なる辞書使用による情報抽出比較

構文木を基に抽出を行う場合、抽出精度は構文解析の精度に依存する。今回 5 種類の抽出規則を作成した際には、構文木に若干手を加えた。しかし、大量のデータを扱う場合は、すべての構文木を確認することは困難である。そこで、5 種類の抽出規則を使い、構文木を操作しない場合の精度を調べた。その際、基本辞書を使った解析と、すべて既知語になるように用語を追加した拡張辞書を使った。

追加した用語は 71 語である。表 1から、30 文の異なり語数は 204 単語であるから、約 3 分の 1 が未知語である。71 用語の品詞付けには、以下の辞書を参考にした：

1. 基本辞書（パーザに付属の辞書）
2. Life Science Dictionary[11]

表 4に追加用語の品詞と、その品詞を付与した用語数を示す。用語の中には、複数の品詞が付与されたものもある：

品詞	用語数
NN（普通名詞、単数）	61
NNS（普通名詞、複数）	6
JJ（形容詞）	4
VBN（動詞の過去分詞）	1
VBD（動詞の過去形）	1

表 4：追加用語の品詞とその頻度

表 5に正解と異なる抽出を行った場合の回数を示す。カテゴリーのあとのカッコ内の数字は、正解文字列の数を示す。表の “n, i, w” は、次の略である：

n: 抽出なし
i: 不完全な抽出
w: 誤った抽出

このうち、不完全な抽出というのは、正解文字列と比較して、一部のみの抽出、または正解文字列以外の部分も抽出している場合である。今回は Enhancer の規則は作成しなかったため、Inhibitor として扱った：

抽出カテゴリー	使用辞書					
	基本辞書			拡張辞書		
	n	i	w	n	i	w
Metabolism (30)	1	0	0	1	0	0
InteractionW (30)	2	0	1	2	0	0
Drug (31)	2	4	1	1	4	1
Inhibitor (30)	2	3	2	1	3	1
Sum (121)	7	7	4	5	7	2

表 5：異なる辞書を使った際の抽出精度

辞書を拡張することにより、精度が上昇した。不完全な抽出も含めると、基本辞書を使った場合には 15 パーセント（18/121）、拡張辞書を使った場合には 12 パーセント (14/121) の誤り率である。3 種類の不正解のうち、誤った抽出がもっとも重大だと考えるが、拡張辞書を使った場合は、全体的な割合は 2 パーセント（2/121）と低い。

一方、表 5には表れていないが、辞書拡張により抽出精度が落ちたものが一箇所あった。例文 (6) の図 5に示す部分である：

(6) *Rifampicin* markedly *increases* the *metabolism* and clearance of *zaleplon* and is expected to reduce its hypnotic effects.

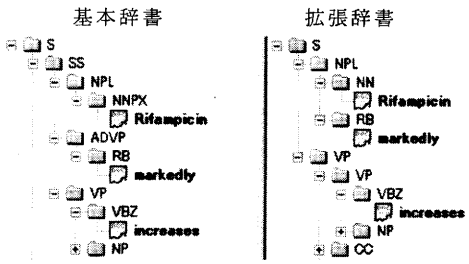

図 5：辞書の拡張により精度が下がった例

この文の Inhibitor は、“Rifampicini”である。基本辞書では、これは未知語のためパーザは独自の品詞として解析し、正しい抽出が行われているが、拡張辞書では普通名詞が付与されているため、パーザは普通名詞として解析し、さらに、副詞の“markedly” を同じ名詞句内のものとして解析した。その

ため現行の抽出規則では、"markedly"も Inhibitor として抽出されている。品詞を固有名詞にすることも考えられるが、普通名詞でも、並列句の前までに文を短くしたところ、正しい解が得られた。

5. 今後の課題

構文木を基に行う情報抽出では処理できないものに、照応、関係代名詞の先行詞の同定、物質の働きのあいまい性の解消がある。

例文(2)で、専門家が抽出したい Drug は、"quinidine and reduces its metabolism" の "quinidine" と "its" の両方である。この要求を満たすには、照応の導入が必須である。

次に、専門知識が必要な例を挙げる：

(7) *Zafirlukast* inhibits cytochromes CYP2C9 and CYP3A4, which may possibly affect the metabolism of a number of drugs but formal interaction studies are lacking.

学校文法によると、which の先行詞は "Zafirlukast inhibits cytochromes CYP2C9 and CYP3A4"である。仮に、"which" の前のカンマをとると、"cytochromes CYP2C9 and CYP3A4" が先行詞となり、構文的にはこれが抽出対象である。しかし、専門家が Inhibitor として抽出したいのは、意味的に "Zafirlukast" だということである。

さらに、曖昧性の解消にも専門知識が必要だと考える。今回は、Enhancer に関する抽出規則は作成しなかったが、Inhibitor に関する規則で抽出された。これは、両者を区別する必要がない場合は有効である。しかし、区別が必要な場合は問題である。どちらか一方の作用に限定されている物質については、辞書を利用することが考えられる。どちらの作用を与えることも可能な場合には、何らかの知識の導入が必須である。

6. おわりに

本稿では、文献からの薬物相互作用情報の抽出に向けて、テスト文の選択、抽出規則の作成、基本辞書と拡張辞書を使った情報抽出比較実験について報告した。抽出規則の作成には、PBIE2 を使った。比較実験では、拡張辞書を使った場合、抽出精度の向上がみられた。特に作用語に関しては誤った抽出はなかった。作用語の自動収集は、Hatzivassiloglou らにより行われているが、これは動詞に限っている[12]。一方、我々の手法では、品詞を限定していない。今回は、少ないデータでの実験にとどまったが、今後は、作用語の抽出を中心に Medline からの関連文献などに応用していく予定である。

文　献

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

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

[3] Friedman, C., et al. (2001): "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.

[4] Pustejovsky, J., et al. (2002): "Robust Relational Parsing over Biomedical Literature: Extracting Inhibit Relations", Proc. of the PSB-2002, 7:362-373.

[5] 吉川澄美, et al. (2004): "薬物相互作用の固有表現分析", 言語処理学会第 10 回年次大会併設ワークショップ「固有表現と専門用語」, pp.17-20.

[6] GENIA Project http://www-tsujii.is.s.-u-tokyo.ac.jp/GENIA/

[7] Tateisi, Y. et al. (2004): "Annotation of Predicate-argument Structure on Molecular Biology Text", Proc. of the IJCNLP-04 Workshop "Beyond shallow analyses. Formalismus and statistical modeling for deep analyses".

[8] Rindflesch, T.C. et al. (2000): "EDGAR: Extraction of Drugs, Genes and Relations from the Biomedical Literature", Proc. of the PSB-2000, 517-528.

[9] Hosaka, J. et al. (2004): "PBIE Toolkit for Data Collection: Toward Parsing-based Information Extraction", Companion Volume to the Proc. of the IJCNLP-2004, pp.29-32.

[10] I. H. Stockley (ed.) (2002): Stockley's Drug Interactions (6^{th} edition), London: The Pharmaceutical Press.

[11] Online Life Science Dictionary http://lsd.bioscinet.org/WebLSD/

[12] Hatzivassiloglou, V. and Weng, W. (2002): "Learning Anchor Verbs for Biological Interaction Patterns from Published Text Articles", Proc. of the NLPBA.