# Automated Extraction of Information from the Literature on Chemical-CYP3A4 Interactions

Chunlai Feng, Fumiyoshi Yamashita,* and Mitsuru Hashida

Department of Drug Delivery Research, Graduate School of Pharmaceutical Sciences, Kyoto University, 46-29 Yoshidashimoadachi-cho, Sakyo-ku, Kyoto 606-8501, Japan

Received March 8, 2007

A text mining system is presented for automatically extracting information from the literature on chemical-CYP3A4 interactions (i.e., substrate, induction, inhibition). The system identifies chemicals and CYP3A4 forms according to a combination of name dictionaries and context features. In addition, it transforms sentences into multiple simple clauses each containing a single event and extracts information on chemical-CYP3A4 interactions using a simple but effective pattern matching method based on the order of three keywords (chemicals, CYP3A4, key verbs). Using this system, 2990 relations including 2700 identified interactions with CYP3A4 for 600 chemicals were extracted from a corpus of 2900 PubMed abstracts. In an evaluation test using 100 randomly selected abstracts, it achieved 87.4% recall and 92.3% precision for identification of the chemical name and 85.2% recall and 92.0% precision for the extraction of chemical-CYP3A4 interactions, respectively. This system will be applicable to interactions of chemicals with any functional proteins, such as enzymes and transporters, simply by changing the list of key verbs.

## INTRODUCTION

The cytochrome P450s (CYPs) are a superfamily of heme-containing mixed function oxygenases that catalyze the regio- and stereoselective oxidation of a wide variety of xenobiotics, including drugs. Such broad substrate specificity of CYPs often leads to unexpected drug−drug interactions. Competitive or noncompetitive inhibition of CYPs by coadministered drugs retards the body clearance of the drugs and results in an unexpected rise in their blood concentrations. On the other hand, induced expression of CYPs reduces or shortens the duration of pharmacological activity of the drugs by accelerating the body clearance. Prediction of drug−drug interactions associated with CYPs is an important issue during drug discovery and development as well as in their clinical applications.

Information on chemical-CYP interactions is usually available in a wide variety of publications. Several reviews[1] and databases (http://medicine.iupui.edu/flockhart/table.htm) devoted to CYP-mediated metabolism have been published. However, it is very time-consuming and labor-intensive to collect and organize literature information in view of the rapid generation and accumulation of experimental data. Recently, several computational approaches to extracting meaningful relationships of interest from the biomedical literature have been developed. The term co-occurrences approach has been used to predict term connections based on their occurrence statistics.[2] The template-based approach that performs pattern matching with specific linguistic structures has also been proposed to handle relationships in complex sentences.[3−5] Most recently, much attention has been paid to a natural language processing (NLP) based method that performs a substantial amount of sentence parsing to extract relationships.[6−10] However, almost all of these studies focused on information extraction involving genes and proteins. Due to differences in linguistic expression, it is generally difficult to adopt the NLP systems directed to the analysis of genes and proteins to other fields, such as the interactions between exogenous chemicals and biomolecules.

The present study was carried out to develop a specialized system for extracting information on chemical-enzyme interactions from the literature. The system is NLP-based, implementing several dictionaries and context information. The feasibility of using this system to extract relationships between chemicals and CYP3A4 was investigated. CYP3A4 is the most abundant human hepatic CYP isoform and is responsible for the metabolism of almost 50% of known drugs by humans.[11] Severe QTc interval prolongation and Torsade de Pointes due to concurrent use of terfenadine and imidazole antifungals (ketoconazole and itraconazole) is one of the most well-known examples of CYP3A4-associated drug−drug interactions.[12] Thus, it is important in terms of drug safety and efficacy to collect information on chemical-CYP3A4 interactions.

## MATERIALS AND METHODS

To extract information on interactions between chemicals and CYP3A4, three main steps were taken into consideration. First, chemicals and CYP3A4 names were identified in the text. Second, sentences containing any names of chemicals and CYP3A4 were transformed into simple clauses, each of which contains a single event. Finally, information on chemical-CYP3A4 interactions was extracted from the clauses by using a pattern matching method based on the order of keywords. The details of each step are described below.

* Corresponding author phone: +81-75-753-4535; fax: +81-75-753-4575; e-mail: yama@pharm.kyoto-u.ac.jp.

10.1021/ci700091m CCC: $37.00 
Published on Web 09/27/2007

**Table 1.** Example of Chemical Names Identification[a]

| process | input text | output | adopted pattern |
|---|---|---|---|
| I | R-(+)-**propranolol** in transgenic Chinese hamster CHL cell lines | R-(+)-propranolol | If **(Dchemical)**(space) then [word][number][punctuation]**(Dchemical)** |
| II | **ketoconazole** (KT) is a worldwide used oral antifungal agent | KT | **(Dchemical)**[space]"("[NN][number]")" |
| | both **testosterone** and nifedipine are CYP3A4 substrates | nifedipine | "both"(space)**(Dchemical)**(space)"and"(space)(NN) |
| | ketoconazole is an **inhibitor of CYP3A4** | ketoconazole | (NN)(space)(VBE){[RB][JJ][space]}**(effectorTerm)**(space)"of"(space)**(Dprotein)** |
| III | flunitrazepam has a high affinity for the **benzodiazepine receptor** | benzodiazepine is filtered | **(Chem)**(space)**(filterTerm)** |

[a] **Bold font** indicates words matched with dictionary entries. Symbols in the patterns are listed in Table 2.

**Entities Name Identification.** Various methods for identifying names of interest in the text have been proposed, including dictionary-based approaches,[6,13−15] rule-based approaches,[16,17] and machine-learning approaches.[18,19] We developed a combination of dictionary-based and rule-based approaches. Although the dictionary-based approach, which has a good performance in terms of accurate identification, was considered to be suitable, it would not necessarily be appropriate for a wide variety of chemical names.[20,21] Dictionary-based approaches often lead to partial matches of the chemical names. In addition, to register a vast number of chemical names, including their spelling variants, in a dictionary is difficult and impractical. Therefore, we created a set of pattern matching rules that use several dictionaries for assistance but identify chemical names principally based on the context itself. On the other hand, identification of CYP3A4 names was carried out by a dictionary-based approach alone, because CYP3A4 does not exhibit many name variants.

**Construction of Dictionaries.** The CYP3A4 name dictionary had 39 name variants, and the chemical name dictionary and protein name dictionary comprised approximately 100 000 and 30 000 entries, respectively, which were created by extracting the names of chemicals and proteins from the MeSH of the U.S. National Library of Medicine (NLM). Specific term dictionaries having words that are related to chemical names (concentration, effect, mg/mL, etc.) were also developed.

**Pattern Matching for Chemical Names.** Pattern matching to identify chemical names consisted of the following three processes:

Process I: Searching for chemical names in the text using a chemical names dictionary, analyze according to the context-based rules, and corrected it if a partial match occurred.

Process II: Identifying chemical names based on the context-based rules. Context-based rules include the ones based on coordinate conjunctions, the ones based on verbs or verbal nouns relating to the action of chemicals, the ones based on specific terms relating to quantity, quality, and usage of chemicals, and the ones for detection of abbreviations based on brackets.

Process III: Removal of false positives from the hits obtained in Processes I and II.

Examples of each process are shown in Table 1. Each rule has a priority number. When more than one rule matched the same region, the rule with the highest priority number was adopted.

**Sentence Processing.** Using a name dictionary developed by pattern matching, the entire text was scanned. The sentences containing both chemical and CYP3A4 names were subjected to sentence processing. First, each sentence was divided into noun phrases and verb phrases based on "part of speech" tags.[8] This step consisted of the following three substeps:

Level I: Creation of noun chunks (NG), including some prepositions (i.e., to, of, with, in, on, for, both, between, about), coordinating conjunctions ("and" and "or"), and serial commas.

Level II: Grouping NG, neighboring determiner (DT), preposition (IN), coordinating conjunction (CC), and 'wh'-determiner (WDT) into a noun phrase chunk. Creating verbal chunks, including some adjectival and adverbial terms.

Level III: Combining coordinated noun phrase chunks with punctuations into a noun phrase (NP). Combining coordinated verbal chunks into a verb phrase (VP).

The next step was to reconstruct a simple clause expressing a single event involving chemicals and CYP3A4. The voice of the verb phrase in the sentence was examined. If the verb voice was active, noun phrases neighboring the verb phrase (NP2 and NP3) were taken to make up a simple clausal structure of syntactic roles (case 1). If the verb voice was passive, the noun phrase on the right (NP3) was split at the position of "by" into NP3′ and NP3″ and NP3″ was taken (case 2). If "by" was not found, NP3 was taken as it was (case 3).

Case 1: [NP1] NP2 VP NP3 [NP4]
Output 1: NP2 VP NP3
Case 2: [NP1] NP2 VBP (NP3′ (by) NP3″) [NP4]
Output 2: NP2 VBP (by) NP3″
Case 3: [NP1] NP2 VBP NP3 [NP4]
Output 3: NP2 VBP NP3

If the sentence was complex, the first verb phrase was treated with the method described above. The second or later verb phrase (VP2) was treated according to cases 4−7. If the voice of the verb phrase (VBP1) preceding VP2 was passive, the noun phrase on the left of VBP1 (NP1) and the noun phrase on the right of VP2 (NP3) were taken to make up a simple clause for VP2 (case 4). If the voice of VP1 was active, the noun phrase on the left of VP2 (NP2) was taken (case 5). If NP2 was terminated with CC, NP1 was taken instead of NP2 since NP1 is most likely to be a subject common to both verbs (VP1 and VP2) (case 6). If NP2 contained ",", ";", or "CC", it was split into subphrases at the position of ",", ";", and "CC" (NP2′ and NP2″) and noun phrases neighboring VP2 (NP2″ and NP3) were taken (case

**Table 2.** Definition of Symbols

| symbol | definition |
|---|---|
| VP | verb phrase |
| VBP | passive verb phrase |
| VN | nearest verb variants from chemical or CYP3A4 |
| VBN | verb past participle |
| VBD | verb past tense |
| RB | adverb |
| Dprotein | protein, such as CYP3A4, CYP2D6 |
| NP | noun phrase |
| NNS | noun plural |
| NNP | proper noun |
| NN | noun |
| Key Verb | interaction verbs in key verb list, such as inhibit |
| JJ | adjective |
| IN | preposition or subordinating conjunction |
| filterTerm | term used in filter rules, such as receptor, channel |
| effectorTerm | term describe compound effect, such as inhibitor, inducer |
| DT | determiner |
| Dchemical | chemicals registered in chemical dictionary |
| CYP | CYP3A4 |
| Chem | chemicals |
| CD | cardinal number |
| CC | coordinating conjunction |
| \| | or |
| [ ] | optional |
| * | space or any other nouns |
| () | essential |
| " " | string |

7). If there was more than one noun phrase between VP1 and VP2, VP2 was treated as case 7.

Case 4: NP1 VBP1 NP2 VP2 NP3

Output 4: NP1 VP2 NP3

Case 5: NP1 VP1 NP2 VP2 NP3

Output 5: NP2 VP2 NP3

Case 6: NP1 VP1 NP2 (CC) VP2 NP3

Output 6: NP1 VP2 NP3

Case 7: NP1 VP1 (NP2′ (CC) NP2″) VP2 NP3

Output 7: NP2″ VP2 NP3

**Interaction Extraction.** Manual coding of a set of simple word patterns for each verb is a popular approach to extracting interactions between entities of interest, based on the fact that the verbs closely relate to the interaction. However, it might not be practically feasible to cover different patterns intrinsic to every verb. We propose here a method of pattern matching simply involving consideration of the appearance order of three types of keywords (that is, chemical names, CYP3A4 names, and key verbs, respectively) within a clause. The key verb list focuses on verbs relating to chemical-enzyme interactions but includes their verbal nouns and verbal adjectives. To improve the precision of information extraction, several filter rules were also implemented in the system.

**Creation of a Key Verb List.** In a preliminary study, we listed all the verbs from the CYP3A4-related abstracts and selected key verbs based on the frequency of their appearance. Possible verbal nouns and adjectives derived from the verbs were added to the list. In addition, several chemical-denoting nouns (e.g., substrate) were also added. All the entries are listed in Table 3.

**Interaction Extraction Processes.** The clauses obtained by the sentence processing mentioned above were processed by different algorithms depending on the positions of the chemical and CYP3A4 in the clause. One (case A) is the appearance of the chemical and CYP3A4 in same phrase, and the other (case B) is in different phrases. When plural chemical names appear in the phrases, first, it is necessary to discriminate the chemicals interacting with CYP3A4.

**Table 3.** Key Verbs Describing Chemical-CYP3A4 Interactions

| | | | |
|---|---|---|---|
| activate | bioactivation | biotrasformation | catalyze |
| clearance | conversion | dealkylation | dechloroethylation |
| decrease | defluorination | dehalogenation | demethylate |
| demethylation | depropargylation | desalkylation | detoxication |
| detoxification | eliminate | elimination | enhance |
| enhancement | epoxidation | form | formation |
| hydrogenation | hydroxylate | hydroxylation | improve |
| inactivate | inactivation | inactivator | increase |
| induce | inducer | induction | inhibit |
| inhibition | inhibitor | inhibitory | interfere |
| metabolic | metabolism | metabolite | metabolize |
| nitroreduction | oxidation | production | reduce |
| reducer | stimulate | substrate | sulfoxidation |
| suppress | transform | transformation | |

In the case where chemicals and CYP3A4 appear in the same phrase (case A), the following 3 filter rules were applied:

Rule 1: Ignore the two nearest-neighboring chemicals (NNC) on both sides of CYP3A4. Reject other chemicals if any of 7 prepositions (i.e., by, with, in, on, to, into, but) is present between them and the NNC.

Rule 2: If chemicals that passed Rule 1 were to the left of "from" or to the right of "to" or "into", reject them.

Rule 3: For chemicals to the left of CYP3A4, filter all of them if ",", "and", or an unclosed bracket ")" is present between NNC and CYP3A4. For chemicals to the right, reject all of them if ",", "and", or an unclosed bracket "(" is present.

In the case where chemicals and CYP3A4 appear in different phrases (case B), the following filter rule was applied:

Rule 4: If any of the prepositions (by, with, in, on, to, into, but) is present between chemicals, reject all chemicals but the first.

In addition to the above-mentioned filter rules for chemical names, a key verb filter rule was applied:

Rule 1: Reject a key verb if "and" or "or" is present between the key verb and the nearest of the chemicals that passed all filter rules for the chemical name.

Examples of each rule are shown in Table 4.

Finally, pattern matching based on the order of the chemicals, CYP3A4, and key verbs in the clause was performed. The patterns used are summarized in Table 5. The highest-priority pattern was adopted when plural patterns could be matched to the target clause.

In addition, the appearance of the negative elements "not", "no", "n't", "unable", and "unlikely" was analyzed to check if the clause provides a negative context. For case A, the phrase was regarded as providing a negative context if any negative elements appear prior to all of the chemicals, verb, and CYP3A4 or between them. For case B, the clause was regarded as providing a negative context if the verb phrase includes a negative element or if any negative elements in the noun phrases appear prior to the chemicals and CYP3A4.

## RESULTS

**Implementation and Evaluation.** The overall architecture of our system is shown in Figure 1. It is implemented on a platform named General Architecture for Text Engineering

**Table 4.** Typical Examples of Filter Rules[a]

| rule | | input text | excluded entities |
|---|---|---|---|
| chemicals filter rule | rule 1 | inhibitory effect of **sulfaphenazole on zaltoprofen** metabolism by **CYP3A4** | sulfaphenazole |
| | rule 2 | metabolism of **midazolam to l′-OH MDZ** by CYP3A4 | l′-OH MDZ |
| | rule 3 | LOP N-demethylation was significantly inhibited when coincubated with ketoconazole (a **CYP3A4** inhibitor) **and quercetin** (a CYP2C8 inhibitor) by 40 and 90%, respectively. | quercetin |
| | rule 4 | CYP3A4 and CYP3A7 mRNA expression levels were markedly up-regulated. by **dexamethasone** (DEX) **but** not by **rifampicin** (RIF). | rifampicin |
| key verbs filter rule | rule 1 | Ritonavir (a **CYP3A4** substrate **and** prototype CYP2D6 **inhibitor**) may potentially affect plasma concentrations of escitalopram. | inhibitor |

[a] **Bold font** describes the matched parts by filter rules.

**Table 5.** Patterns of Interaction Extraction

| | priority order[a] | pattern form | example |
|---|---|---|---|
| case A: chemicals and CYP3A4 in the same phrase | 1 (high) | (Chem\|CYP) VN2 (CYP\|Chem) | inhibitory effect of sulfaphenazole on zaltoprofen metabolism by CYP3A4 ... |
| | 2 | VN1 (Chem\|CYP) (CYP\|Chem) | 1-hydroxylation of midazolam (CYP3A4),O-de-ethylation of phenacetin (CYP1A2) and O-demethylation of dextromethorphan (CYP2D6) ... |
| regulative expression: VN1*(Chem\|CYP)* VN2 *(CYP\|Chem)* VN3 | 3 (low) | (Chem\|CYP) (CYP\|Chem) VN3 | troleandomycin, a CYP3A4 inhibitor ... |

| | active | passive | pattern form | example |
|---|---|---|---|---|
| case B: chemicals and CYP3A4 in different phrases | 1 | 1 | (Chem\|CYP) keyVerb(CYP\|Chem) | azelastine and its desmethylazelastine formation inhibited CYP3A4 activity |
| | 2 | 4 | (Chem\|CYP) VN3 (CYP\|Chem) | haloperidol is a substrate of CYP3A4 and inhibitor of CYP2D6 |
| regulative expression: VN1* (Chem\|CYP)*VN2*keyVerb* VN3*(CYP \|Chem) *VN4 | 3 | 5 | (Chem\|CYP) (CYP\|Chem) VN4 | CYP2A6 or CYP3A4 inhibition can diminish halothane metabolism |
| | 4 | 2 | VN1 (Chem\|CYP) (CYP\|Chem) | metabolism of azelastine, a CYP2D6 inhibitor is catalyzed by CYP3A4 |
| | 5 | 3 | (Chem\|CYP) VN2 (CYP\|Chem) | clinically important CYP3A4 inhibitors include itraconazole, ketoconazole, ... |

[a] Priority orders were defined according to the voice of the verbs in case B.

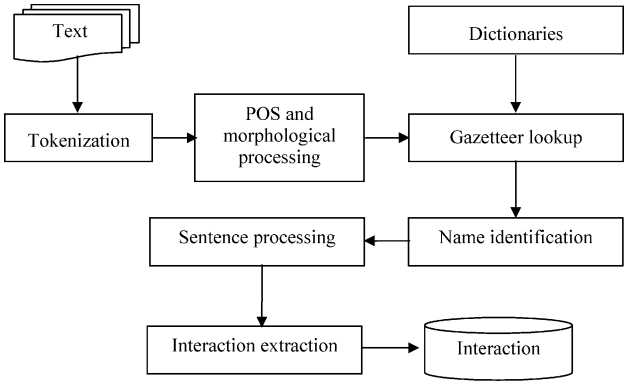



**Figure 1.** Component modules of application.

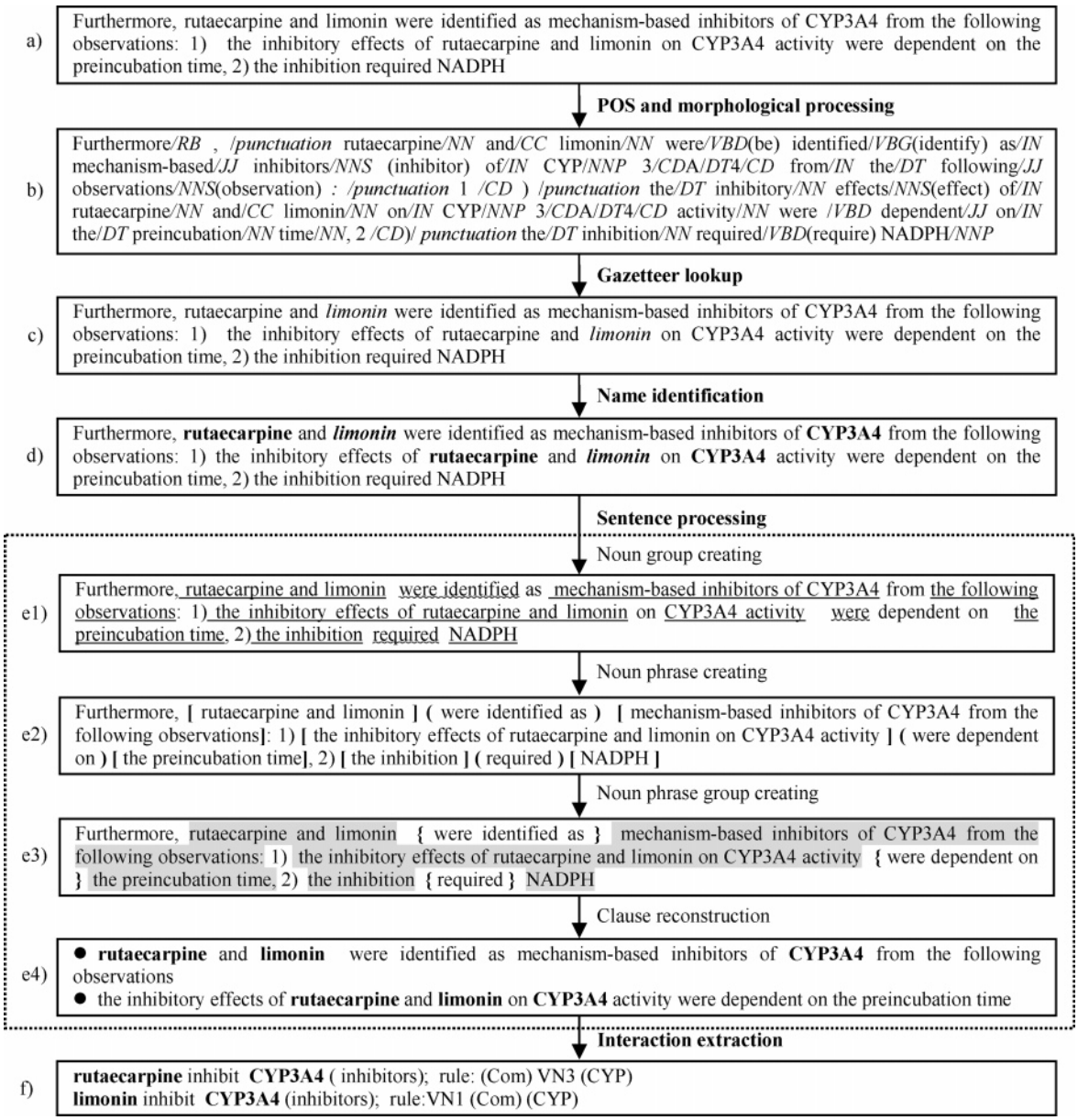
(GATE) developed at the University of Sheffield, U.K.[22] The "Tokenization" and "POS and morphological processing" processes used the tools provided by GATE without a change as preprocessing. Chart 1 shows a typical flow of interaction extraction from the text in our system. The performance of the information extraction system was checked with a corpus of 2900 MEDLINE abstracts involving interactions between chemicals and CYP3A4. The abstracts corpus was obtained from NLM PubMed using "CYP3A4" as a keyword. Following a series of information extraction processes, 2990 relationships including 2700 identified interactions with CYP3A4 for 600 chemicals were retrieved (Table 6). When repetitions of the records were removed, a total of 724 chemical-CYP3A4 interactions were obtained (Table 6).

In order to validate our system, 100 abstracts were randomly selected from the CYP3A4 abstract corpus, which were separated from the data examined when the system was constructed to evaluate the recall and precision of both chemical name identification and interaction extraction procedures. The recall was the ratio of the number of relevant items retrieved to the number of relevant items in collection, while the precision was the ratio of the number of relevant items retrieved to the number of items retrieved. Evaluation of chemical name identification was performed by using the GATE's measurement system, but the recall and precision of interaction extraction were evaluated manually due to difficulty in annotation. The test data set is available from http://dds.pharm.kyoto-u.ac.jp/downloads/tm-cyp.html. The results of chemical names identification and interaction extraction are summarized in Table 7. The recall for chemical names was not high enough with the dictionary-based approach alone, whereas it was improved by combining it with the context-based approach. The precision of information extraction was somewhat spoiled due to the higher extraction ability of the combination approach. When filter rules were introduced into the combination approach, the precision was drastically improved with a minimum loss of recall.

## DISCUSSION

The present study proposes a NLP-based text-mining system for interactions between chemicals and enzymes. This

**Chart 1.** Typical Example of Interaction Extraction Processes[a]




[a] a) Target sentence. b) The result of POS and morphology analysis. Italics indicate the POS, and parentheses () indicate the result of morphology analysis. c) The result of dictionaries Lookup. Italics indicate matched entities. d) Chemicals and CYP3A4 names were identified based on the rules. They are indicated by bold font. e-1) The result of a noun group creating step of sentence processing. Single-underlined items indicate noun groups (NG), and wave-underlined texts indicate verb groups (VG). e-2) The result of the noun phrase creating step of sentence processing. Square brackets [ ] indicate noun phrases (NP), and parentheses () indicate verb phrases (VP). e-3) The result of noun phrase groups creating step of sentence processing. Shaded texts indicate noun phrase groups (NPG), and curly brackets {} indicate verb phrase groups (VPG). e-4) Clauses involving relationships between chemicals and CYP3A4 were reconstructed. f) Interactions were extracted by pattern matching based on the order of keywords.

**Table 6.** Result of Chemical-CYP3A4 Interactions Extraction

| | substrate | inhibitor | inducer | total[a] |
|---|---|---|---|---|
| records with repetition[b] | 1435 | 849 | 416 | 2700 |
| records without repetition[c] | 398 | 210 | 116 | 724 |

[a] Chemicals that represent two or three interactions were counted in ease case repeatedly. [b] The number of records retrieved was indicated. [c] Records having the same meaning were removed.

system has two major features compared with previously reported methods. The first feature is that the dictionary- and context-based approaches were combined to identify chemical names effectively. Generally, the chemical names dictionary was used to address the first task of chemical name recognition. However, it is generally difficult to detect chemical names comprehensively with the dictionary alone. Partial matches often occur in simple text matching procedures, lowering the precision of name identification. In addition, the same words or phrases can be used with different meanings according to the context (e.g., APC can be a substance or a gene). Many chemicals have several different names. The recall of name identification is generally low due to the presence of too many chemical names or their variants. Although Narayanaswamy and Ravikumar[17] proposed a rule-based approach to recognize chemical names by using the chemical root forms (e.g., "meth" in "N-methylformamide"), various morphological features (e.g.,

**Table 7.** Evaluation of Chemical Names Identification and Interaction Extraction

| | | chemical names identification | | | | interaction extraction | |
|---|---|---|---|---|---|---|---|
| | | dictionary-based | context-based | combined (without filtering) | combined (with filtering) | keyword order-based (without filtering) | keyword order-based (with filtering) |
| number of records | true positive[a] | 851 | 558 | 1099 | 1084 | 142 | 140 |
| | false positive[b] | 76 | 61 | 221 | 67 | 19 | 12 |
| | false negative[c] | 393 | 666 | 145 | 160 | 22 | 24 |
| recall | | 0.670 | 0.497 | 0.874 | 0.874 | 0.866 | 0.852 |
| precision | | 0.886 | 0.906 | 0.750 | 0.923 | 0.882 | 0.92 |

[a] True positive indicates relevant items retrieved. [b] False positive indicates irrelevant items retrieved. [c] False negative indicates relevant items unretrieved.

caps only, combination of letters and digits), suffixes (e.g., "−ic" in "sulfonic"), and functional terms (e.g., steroid) as surface clues, functional chemicals including drugs are usually expressed as generic names and trade names which do not have such apparent surface clues. To improve the chemical names identification, we introduced rules for the detection of chemical names based on the context that were based on verbs or verbal nouns relating to the action of chemicals and specific terms relating to quantity, quality, and usage of chemicals. The combination of a dictionary- and a context-based approach was beneficial to avoid partial matches or mismatches of chemical names and identify names which are not listed in a dictionary. Practically, our method gave a high recall (0.874) and precision (0.923) for chemical names identification (Table 7).

The other feature was to perform pattern matching only based on the sequence of keywords (chemicals name, CYP3A4 name, and key verbs) in a clause when information on chemical-CYP3A4 interactions was extracted. Our system can give a higher recall than the one that considers individual patterns for each verb, since it requires only a small number of the rules to cover the patterns of clauses, as listed in Table 5. Although the precision tends to be worse without any filters, it can be overcome by implementing several rules to filter temporal hits based on prepositions, coordinate conjunctions, or brackets (Tables 4 and 7). Our method to extract relationships is simple and reasonably accurate. Although a method in which formulas for information extraction are automatically generated has been proposed by Huang et al.,[23] it appears that the quality of the formulas generated depend on a training set.

Rendic[1] has summarized reactions, substrates, inducers, and inhibitors of human cytochrome P450 enzymes. The information on CYP3A4 they collected from the literature included a total of 733 interactions (420 substrates, 64 inducers, and 249 inhibitors). Here, chemicals that represent two or three interactions were counted in each case repeatedly. The present analysis found 329 interactions (188 substrates, 30 inducers, and 111 inhibitors) out of them. In addition, we detected 395 interactions which have not been listed in Rendic's report.[1] Considering that only PubMed abstracts were subjected to the present analysis, it appears that our method performs well in extracting chemical-CYP3A4 interactions.

In this method, errors of information extraction occurred primarily due to errors in part of speech tagging that determine the boundaries between noun and verb phrases. More accurate speech tagging and rules for discriminating between noun and verb phrases remain to be improved in further studies.

This system will be applicable to interactions of chemicals with any functional proteins, such as enzymes and transporters, simply by changing the list of key verbs.

## REFERENCES AND NOTES


(1) Rendic, S. Summary of information on human CYP enzymes: human P450 metabolism data. *Drug Metab. Rev.* **2002**, *34* (1−2), 83−448.
(2) Stapley, B. J.; Benoit, G. Biobibliometrics: information retrieval and visualization from co-occurrences of gene names in Medline abstracts. *Pac. Symp. Biocomput.* **2000**, *5*, 529−540.
(3) Ng, S. K.; Wong, M. Toward Routine Automatic Pathway Discovery from On-line Scientific Text Abstracts. *Genome Inform. Ser. Workshop Genome Inform.* **1999**, *10*, 104−112.
(4) Humphreys, K.; Demetriou, G.; Gaizauskas, R. Two applications of information extraction to biological science journal articles: enzyme interactionns and protein structures. *Pac. Symp. Biocomput.* **2000**, *5*, 505−516.
(5) Thomas, J.; Milward, D.; Ouzounis, C.; Pulman, S.; Carroll, M. Automatic extraction of protein interactions from scientific abstracts. *Pac. Symp. Biocomput.* **2000**, *5*, 541−552.
(6) Ono, T.; Hishigaki, H.; Tanigami, A.; Takagi, T. Automated extraction of information on protein-protein interactions from the biological literature. *Bioinformatics* **2001**, *17* (2), 155−161.
(7) Ding, J.; Berleant, D.; Nettleton, D.; Wurtele, E. Mining MEDLINE: abstracts, sentences, or phrases? *Pac. Symp. Biocomput.* **2002**, *7*, 326−337.
(8) Pustejovsky, J.; Castano, J.; Zhang, J.; Kotecki, M.; Cochran, B. Robust relational parsing over biomedical literature: extracting inhibit relations. *Pac. Symp. Biocomput.* **2002**, *7*, 362−373.
(9) Horn, F.; Lau, A. L.; Cohen, F. E. Automated extraction of mutation data from the literature: application of MuteXt to G protein-coupled receptors and nuclear hormone receptors. *Bioinformatics* **2004**, *20*, 557−568.
(10) Hu, Z. Z.; Narayanaswamy, M.; Ravikumar, K. E.; Vijay-Shanker, K.; Wu, C. H. Literature mining and database annotation of protein phosphorylation using a rule-based system. *Bioinformatics* **2005**, *21*, 2759−2765.
(11) Guengerich, F. P. Role of cytochrome P450 enzymes in drug-drug interactions. *Adv. Pharmacol.* **1997**, *43*, 7−35.
(12) Thompson, D.; Oster, G. Use of terfenadine and contraindicated drugs. *JAMA* **1996**, *275*, 1339−1341.
(13) Egorov, S.; Yuryev, A.; Daraselia, N. A simple and practical dictionary-based approach for identification of proteins in Medline abstracts. *J. Am. Med. Inform. Assoc.* **2004**, *11* (3), 174−178.
(14) Ling, X.; Jiang, J.; He, X.; Mei, Q.; Zhai, C.; Schatz, B. Automatically generating gene summaries from biomedical literature. *Pac. Symp. Biocomput.* **2006**, *11*, 40−51.
(15) Chun, H. W.; Tsuruoka, Y.; Kim, J. D.; Shiba, R.; Nagata, N.; Hishiki, T.; Tsujii, J. Extraction of gene-disease relations from medline using domain dictionaries and machine learing. *Pac. Symp. Biocomput.* **2006**, *11*, 4−15.
(16) Fukuda, K.; Tamura, A.; Tsunoda, T.; Takagi, T. Toward information extraction: identifying protein names from biological papers. *Pac. Symp. Biocomput.* **1998**, *3*, 707−718.
(17) Narayanaswamy, M.; Ravikumar, K. E.; Vijay-Shanker, K. A biological named entity recognizer. *Pac. Symp. Biocomput.* **2003**, *8*, 427−438.
(18) Kazama, J.; Makino, T.; Ohta, Y.; Tsujii, J. In *Tuning Support Vector Machines for Biomedical Named Entity Recognition*. In the Proceedings of the Natural Language Processing in the Biomedical Domain (ACL 2002), Philadelphia, PA, U.S.A., 2002; pp 1−8.
(19) Zhou, G.; Shen, D.; Zhang, J.; Su, J.; Tan, S. Recognition of protein/gene names from text using an ensemble of classifiers. *BMC Bioinformatics* **2005**, *6* (Suppl 1), S7.

(20) Sirohi, E.; Peissig, P. Study of effect of drug lexicons on medication extraction from electronic medical records. *Pac. Symp. Biocomput.* **2005**, *10*, 308−318.
(21) Cohen, A. M.; Hersh, W. R. A survey of current work in biomedical text mining. *Brief Bioinform.* **2005**, *6* (1), 57−71.
(22) Cunningham, H.; Maynard, D.; Bontcheva, K.; Tablan, V. In *GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications;* Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02), Philadelphia, PA, 2002.
(23) Huang, M.; Zhu, X.; Hao, Y.; Payan, DG.; Qu, K.; Li, M. Discovering patterns to extract protein−protein interactions from full texts. *Bioinformatics* **2004**, *20*, 3604−3612.

CI700091M