テンプレートマッチングと照応解析を利用した文献からの蛋白質活性部位情報抽出方式

兼田佳和[†] 沼 雅之[†] 大川剛直[†]

蛋白質の機能は活性部位と呼ばれる局所的な部位によって決定されることから，詳細な活性部位情報をデータベース化することが望まれている．活性部位情報は蛋白質の構造解析実験に関する論文に記述されているが，この数は数万編に上るため，人手で抽出するのは時間的・労力的に困難である．そこで本研究では，文献から活性部位情報を自動抽出する方式を提案する．文献には複文を用いた表現が多いため，単純な文字列照合によるテンプレートマッチングは難しい．したがって，入力文の構文解析を行うことで複文と単文用のテンプレートとの照合を可能にする．テンプレートマッチングによる情報抽出の結果には照応詞が含まれる．蛋白質に関する文献では照応詞に数量表現が多用されることに注目し，照応詞–先行詞間の数量一致性に基づく照応解析手法を提案する．提案手法を 10 編の文献に適用して評価を行った結果，照応解析と情報抽出の F 値はそれぞれ 0.81，0.70 となった．

Automatic Extraction of Protein Active Site Information from Literature Using Template Matching and Anaphor Analysis

Yoshikazu Kaneta,[†] Masayuki Numa[†] and Takenao Ohkawa[†]

The function of a protein is determined by its local functional site, called an active site. For functional analysis of proteins, it is important to accumulate information about the location and the chemical properties of active sites. This paper presents a method of automatically extracting protein active site information from literature. Since complex expressions frequently occur in scientific documents, syntactic analysis is used to match complex sentences, while the extraction rule is employed for simple sentences. The anaphora is contained in the information extraction result. Because numerical phenetics and symbolic expressions are frequently used in the anaphora, we propose an anaphora analysis method based on the consistency of the numerical phenetics between an anaphor and an antecedent. We applied the proposed method to actual scientific literature, obtaining 0.81 anaphora analysis F-measure and 0.70 information extraction F-measure.

1. はじめに

蛋白質は，化学物質との相互作用によって触媒や免疫といった特定の機能を発現する高分子である．その機能や化学物質との相互作用は人間の生命活動を担う基盤であることから，これらの現象の理解を通じた医療や創薬への応用が期待されている[1]．蛋白質の機能や相互作用は，蛋白質全体ではなく局所的な部位（活性部位）によって決定される[2] ことが知られており，活性部位が蛋白質のどこに位置しているか，その部位がどのような物理的・化学的性質を持つかといった情報は，蛋白質の機能を理解するうえできわめて重要となる．このような活性部位情報は，立体構造解析に関わる文献内に記述されているが，いまだデータベース化されておらず，現状では有効に活用されていない．

活性部位情報を活用すれば，蛋白質の機能推定や機能分類の高度化が期待できる．それにともない，標的蛋白質に対する薬物設計の進展と創薬の効率化，蛋白質の構造–機能間のギャップ解消による蛋白質設計原理の解明といった効果もあげられる[3]．しかしながら，数万編に上る文献から手作業で活性部位情報を抽出するのは時間的にも労力的にも困難であるため，活性部位情報の自動抽出や抽出作業の効率的な支援が望まれている．

蛋白質関連の文献から情報抽出を行う研究例として，辞書を使わずに蛋白質名を同定する手法[4]，動詞・蛋白質名とテンプレートとのマッチングによる情報抽出[5)~7)] などが行われているが，いずれも作用を表す単語と蛋白質名を抽出することを目的としているため，

† 大阪大学大学院情報科学研究科
Graduate School of Information Science and Technology, Osaka University

活性残基や結合に関わる原子といった活性部位の詳細情報は得られない．また，主語と目的語，およびその間の関係を表す語句を抽出する研究[8),9)]も行われているが，これらの研究は抽出すべき情報が主語と目的語に集約されていることを前提としたものである．

筆者らは，活性部位情報を有効活用するためには，単に対象となる蛋白質の名称だけでなく，活性部位の位置を示す残基名や活性の種類なども重要であるという観点から，これらの情報を自動抽出するシステムExAct（EXtraction system of ACTive site information）を構築している．本システムでは，入力文献内の固有表現の特定，テンプレートマッチングの活用，および照応関係[10),11)]の解決により，全自動で正確な情報抽出を試みている．本論文では，後段のプロセスに焦点をあて，テンプレートマッチングと照応解析を実現する手法について論じる．

テンプレートマッチングにおいては，文献内には複文を用いた複雑な文があることから，単文用テンプレートとの照合が難しい．そこで，複文と単文用テンプレートとの照合を可能にするため，構文解析[11)]を利用したマッチングを行う．

照応関係を解決する方法としては，センタリング理論[12)]，名詞の省略形を考慮した照応解析[13)]などの研究例があるが，これらは照応詞とその参照先である先行詞が1対1の関係にあることを前提としている．これに対し，蛋白質の文献ではいくつかの残基情報などをひとまとめにしてその性質を論じることが多い．したがって，照応詞では数詞や数字を用いた数量表現が，またその先行詞では括弧などの記号付き表現によっていくつかの単語がまとめられている表現が多く見られる．そこで，照応詞と先行詞の数量的な一致性に基づいて照応解析を行う手法を提案する．特に先行詞側ではコンマなどによる列挙表現が多いことに注目し，表層情報に基づいて同類語をグループ化することで先行詞との数量一致度を計算する．

2. 蛋白質と活性部位情報

2.1 蛋白質の構造

蛋白質は，アミノ酸を構成しているアミノ基とカルボキシル基とがペプチド結合を行うことによって1本のペプチド鎖になったものである．自然界には20種類のアミノ酸が存在しており，その名前はアルファベット3文字を用いて表される．ペプチド結合により脱水したアミノ酸はアミノ酸残基と呼ばれ，各残基にはペプチド鎖の末端から数えて何番目にあたるかを表す番号（残基番号）が割り振られている．たとえば，末端から100番目にあるアルギニンは文献内で “Arg100”，“Arg^{100}” などと表記される．

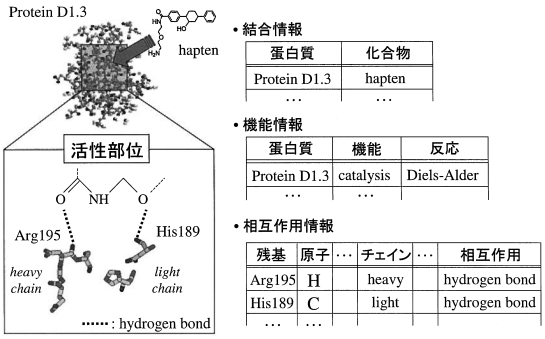


図 1 活性部位情報の一例
Fig. 1 An example of active site information.

2.2 活性部位の相互作用

活性部位では，原子間に生ずる相互作用によって特定の物質と結合・遊離し，様々な機能を発現する．個々の原子間の相互作用は非常に弱い結合力であるが，複数の相互作用が集まることで強い結合力が生じ，物質間の結合が保たれる．文献内に記述される相互作用には，“hydrogen bond”（水素結合），“van der Waals contact”（ファンデルワールス結合）などがあるが，相互作用の種類は明示されない場合もしばしばある．また，相互作用する相手物質としては別蛋白質，DNA，化合物，イオンなどがあり，それぞれ “PARP”，“(5′-D(Tcc)-3′)4”，“hapten”，“Zn^{2+}” などの例があげられる．

2.3 活性部位情報

蛋白質の立体構造解析に関する文献には様々な情報が記載されているが，構造の測定方法や実験条件などはすでにデータベース化されているため，抽出する必要はない．文献内に潜在している重要な情報としては，結合情報，機能情報，相互作用情報などがあり，これらを自動的に抽出することが望まれる．**図 1** に各情報の例を示す．

結合情報：文献内に記述されている蛋白質と他の物質との結合を表す情報である．蛋白質の名称，およびそれに結合する物質の名称の組として定義される．

機能情報：蛋白質がどのような機能を発現するかを表す情報である．蛋白質の名称と機能の名称，および化学反応の名称の組として定義される．蛋白質の機能には，触媒や免疫といった生化学的機能や，活性エネルギー上昇による特定物質の安定化といった分子レベルの機能が含まれる．

相互作用情報：蛋白質分子と他物質との作用相互に関する情報である．活性位置におけるアミノ酸残基の

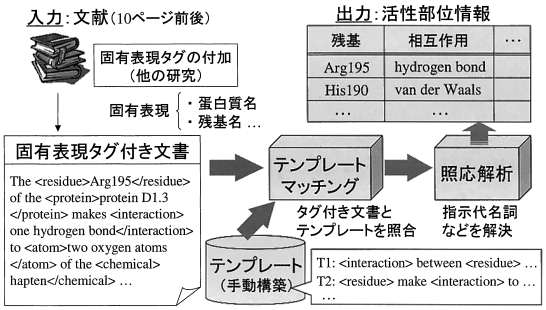


図 2 活性部位情報抽出システムの概要
Fig. 2 Extraction system of active site information.

名称と残基番号，その残基で特に結合に関わる官能基と原子の名称，その残基が属するチェイン情報，および相互作用の名称の組として定義される．チェイン情報は，結合に関わる残基が主鎖・側鎖のどちらであるかと，重鎖と軽鎖のどちらであるかを表す．

3. 活性部位情報抽出システム — ExAct

文献からの活性部位情報抽出システム ExAct の概要を**図 2** に示す．まず手作業によりテンプレートを作成し，データベースに格納しておく．テンプレートとは，抽出すべき情報が文献内でどのように記述されているかを表したものである．たとえば，ある 2 つの残基間に相互作用があるという情報を抽出したい場合，次のようなテンプレートを記述する．

- <interaction> of <residue> and <residue>
- <residue> make <interaction> to <residue>

テンプレートは以下の 2 つの要素から構成される．

- 通常単語：“make” などの通常の文で用いる単語
- タグ：<residue> のようなタグ付き文字

本システムでは，入力文献が与えられたとき，固有表現タグの付加とテンプレートマッチング，照応解析という 3 段階の処理を行う．固有表現とは一般に人名や組織名などを指す[10]が，本論文の対象では蛋白質名や残基名，化合物名などである．固有表現タグとは，これらの固有名詞にそれが属するカテゴリを与えるものである．固有表現を特定した後，それと動詞，前置詞などをキーとしてテンプレートマッチングを行い，照合した場合にはその記述に従って情報抽出する．抽出した情報に指示代名詞といった照応詞が含まれる場合には照応解析を行うことでその先行詞を特定する．なお，一般的な情報抽出システムでは，複数文からの抽出結果をまとめるイベント結合処理[10]が含まれるが，本研究で抽出対象となる多くの情報は 1 文中に記述されているため，この処理を省略している．

表 1 固有表現タグ一覧
Table 1 List of proper noun tag.

固有表現タグ	内容	例
<protein>	蛋白質	antibody D1.3
<chemical>	化合物	esterase
<peptide>	ペプチド	chloromethyl ketone
<atom>	原子	oxygen atom
<ion>	イオン	zinc ion
<molecule>	分子	water molecule
<group>	官能基	hydroxyle group
<residue>	残基	Tyr100
<chain>	チェイン	side chain, main chain
<sec_structure>	二次構造	α-helix, β-sheet
<tert_structure>	三次構造	oxyanion hole, pocket
<domain>	ドメイン	CDR L1

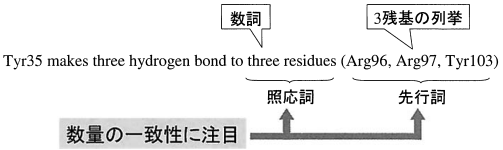


図 3 数量表現が現れる照応関係
Fig. 3 Anaphora with quantifier.

本方式における主な研究課題としては，固有表現タグの付加をいかに高精度に行うか，それに続く情報抽出の処理をいかに高精度に行うかの 2 つがあげられる．本論文では後者の課題，すなわち情報抽出をいかに高精度に行うかを対象とし，固有表現タグ付加手法[14]は稿を改めて論じる．なお，固有表現タグ付加手法の精度は現状で 77.8%であり，**表 1** の固有表現タグが付加される．

4. テンプレートマッチングと照応解析を利用した活性部位情報抽出方式

4.1 アプローチ

文献には複文を用いた表現が多用されるため，単文用のテンプレートと複文とをいかに照合するかが重要となる．これらを正確に照合するためには，文の構造を把握し，単語間のつながりを認識する必要がある．そこで本論文では，構文解析を利用して単語間のつながりを求めることにより，テンプレートマッチングを可能とする．

一方，照合した文に照応関係がある場合，照応詞に対応する先行詞を正確に見つける必要がある．蛋白質構造解析の文献を分析すると，ある残基集合をひとまとめにしてその機能や相互作用について論じるといった記述が多い．このため，照応詞には数量表現が頻繁に現れている．このような照応関係では，**図 3** に示すように，照応詞と先行詞に含まれる数量を特定するこ

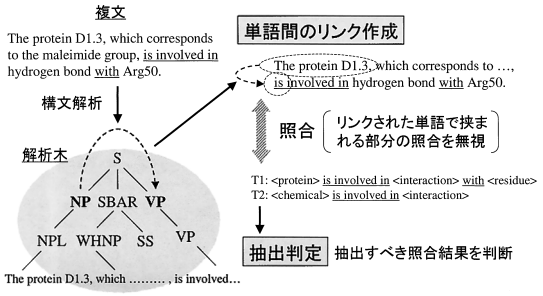


図 4 テンプレートマッチングの概要
Fig. 4 Outline of the template matching method.

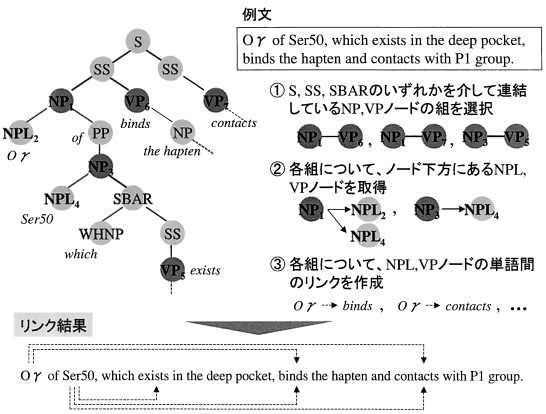


図 5 構文解析を利用した単語間のリンク方法
Fig. 5 Link recognition between two words.

とで照応関係を正しく解析できると考えられる．

数量関係の一致性を導入するためには，先行詞中に含まれる単語の数を調べる必要がある．先行詞では図 3 のようにコンマなどによる列挙表現が頻出することから，表層情報を利用した同類語のグループ化により，これを実現する．

4.2 構文解析を利用したテンプレートマッチング

4.2.1 テンプレートマッチングの概要

テンプレートマッチングの概要を**図 4** に示す．はじめに入力文に対して構文解析を行い，解析木を作成する．そして，接続詞や関係詞のある複文を単文と同様に扱うために，単語間の構文的なつながり（リンク）を作成する．このリンクをたどることによって，テンプレートと照合不可能な節を読み飛ばすマッチングを実現する．最後に，照合したテンプレートについて情報を抽出するかどうかの判断を行う．なお，構文解析には “Apple Pie Parser” [15] を利用する．

4.2.2 構文解析を利用した単語間のリンク

構文解析を利用した単語間のリンク作成方法を**図 5** に示す．複文では主に接続詞と関係詞を用いた表現が多く見られる．接続詞を用いた表現の場合，図 5 の例文に示すように，接続詞の後方の語（“contacts”）とその主体となる語（“Oγ of Ser50”）とをリンクする必要がある．また，関係詞を用いた表現の場合，その直前の語（“Ser50”）と関係詞の直後の語（“exists”），および関係詞節終了直後の語（“binds”）とをリンクする必要がある．これらの語は解析木において文ノードや関係詞ノードで結ばれていることから，この 2 つのノードで連結される名詞，動詞ノードをリンクの対象として選択する．選択したノードの下にはそれと同種のノードが複数個存在する場合があるが，可能な限り多くの照合を試行するためにすべてのノードをリンクの対象とする．ただし，下方にある同種のノードを探索するとき，文ノードや関係詞ノードをまたがる位置にあるノードは別のノードとリンクすると考えられるので，これらを超えないように探索する．単語間のリンク方法の具体的な手順を以下に示す．

（1） 名詞ノード（NP, NPL）と動詞ノード（VP）の任意の 2 個の組合せにおいて，2 ノード間を結ぶパス上に存在するノードが文ノード（S, SS）または関係詞ノード（SBAR）のみである組合せを探索する．見つけた組合せを $L = \{(NP_1, VP_1), (NP_2, VP_2), \cdots, (NP_n, VP_n)\}$ とする．L の各要素に対して以下の処理を行う．

（2） NP_i 以下の子ノードに含まれる NPL ノードを探索する．ただし，探索途中に SS, SBAR ノードのいずれかがある場合は探索を打ち切る．見つけた NPL ノードを $L_{NPL_i} = \{n_{i1}, n_{i2}, \cdots, n_{im_{n_i}}\}$ とする．同様の処理を VP_i について VP ノードを発見するために行う．発見したノードを $L_{VP_i} = \{v_{i1}, v_{i2}, \cdots, v_{im_{v_i}}\}$ とする．

（3） L_{NPL_i} の要素と L_{VP_i} の要素の全組合せ $\{(n_{i1}, v_{i1}), (n_{i1}, v_{i2}), \cdots, (n_{im_{n_i}}, v_{im_{v_i}})\}$ に対し，そのノードに含まれる単語をリンクする．

4.2.3 文とテンプレートの照合判定

文とテンプレートの照合は単語単位での前方一致優先照合で行う．ただし，リンクする 2 単語で挟まれる部分は読み飛ばし可能とする．また，抽出対象となる情報では形容詞・副詞は特に重要でないため，これらの単語は照合を行わない．なお，品詞判定には “Brill’s Tagger” [16] を利用する．

テンプレートタグ部分の照合については，たとえば “The contact is made by <atom> of <protein> . . . ” のようなテンプレートがあった場合，<protein> の部分が <chemical> に置き換わっても文脈が成立することを考慮し，テンプレートタグと固有表現タグの種類が異なっていても照合したと見なす．

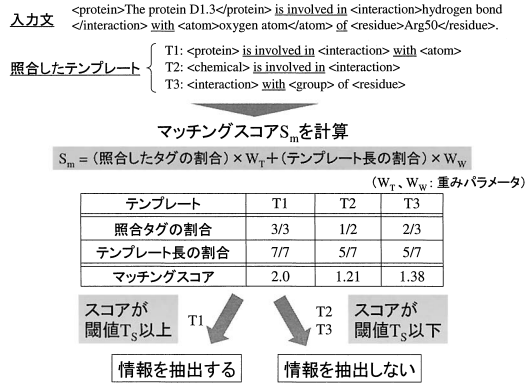


図 6 マッチングスコアによる抽出判定
Fig. 6 Extraction judgement by matching score.

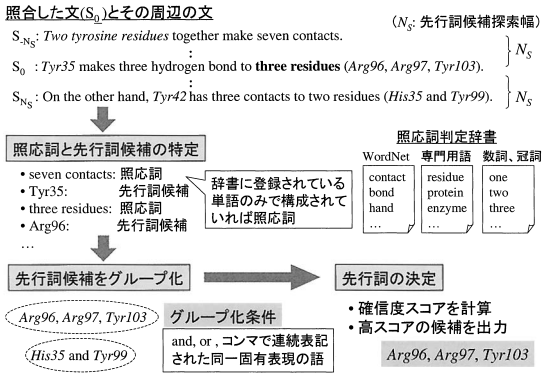


図 7 提案する照応解析手法の概要
Fig. 7 Outline of the anaphora analysis method.

4.2.4 マッチングスコアによる抽出判定

文とテンプレートが照合した場合，異なるテンプレートタグと固有表現タグの照合を許可しているため，照合結果をすべて抽出すると信頼性が低い出力になる可能性がある．これを回避するため，照合結果の信頼性を判断する指標値としてマッチングスコアを導入する．照合結果の信頼性は，テンプレートタグとの一致度が高く，かつテンプレートに含まれる単語数が多い（不適当な文と照合しにくい）ものほど高いと考えられることから，これらの加重和としてマッチングスコアを定義する．マッチングスコア S_m の計算式を以下に示す．

$$S_m = \frac{N_{tm}}{N_{tt}} \times W_T + \frac{N_{wt}}{N_{max}} \times W_W$$

ここで，N_{tm} はテンプレートタグと照合した文中の固有表現タグが同一であるタグの個数，N_{tt} はテンプレートのタグ数，N_{wt} はテンプレートの通常単語数とタグ数の和，N_{max} は全テンプレート中における通常単語数とタグ数の和の最大値を表す．また，W_T，W_W はそれぞれの評価項目に対する重みパラメータである．S_m がマッチングスコア閾値 T_S 以上となれば照合した文からの情報抽出を行う．逆に S_m が T_S を下回る場合には，照合した文からは情報抽出を行わない．マッチングスコアによる抽出判定の例を**図 6** に示す．なお，照合した文からの情報抽出はルールを用いて行う．ルールとは，テンプレート中の固有表現タグと合致した部分をテーブルのどの項目にあてはめるかを記述したものである．

4.3 数量表現の一致性に着目した照応解析手法

4.3.1 照応解析手法の概要

提案する照応解析手法の概要を**図 7** に示す．蛋白質の文献における照応詞には数量表現が多く見られることを考慮し，数詞，指示代名詞，冠詞，一般名詞，および物質の大まかな分類を表す専門用語（“protein”，“enzyme” など）のいずれかに含まれる単語のみで構成されている名詞を照応詞と判定する．ここで，一般名詞かどうかの判断には WordNet ☆の名詞辞書を利用する．照応詞に数量表現が存在した場合，その照応詞は複数の先行詞を参照していると考えられる．そこで，照応詞–先行詞間の数量関係の一致性を調べるため，先行詞候補となる名詞に対してグループ化を行う．グループ化の対象となる名詞は，照応詞と同じ固有表現タグが付加されており，かつ連続して表記されていると考えられる．したがって，照応詞と同じ固有表現を持ち，かつ等位接続詞（“and”，“or”）またはコンマ記号（“,”）で連続表記されているものをグループ化する．以下に提案する照応解析手法の概略手順を示す．

（1） テンプレートとマッチした文を S_0 とし，S_0 を中心とする前後文への先行詞探索幅を N_S とする．

（2） S_0 内の名詞のうち照応詞と判定されたものを照応詞リストに登録する．登録された照応詞を $\{r_1, r_2, \cdots r_n\}$ とする．

（3） 文 $S_{-N_S} \cdots S_{N_S}$ に含まれ，かつ照応詞と判定されない名詞のうち，照応詞 r_i と同一固有表現のものを先行詞候補 a_{ij} とし，先行詞候補リストに登録する．登録された候補を $\{a_{11}, a_{12}, \cdots a_{nn_i}\}$ とする．

（4） 文 $S_{-N_S} \cdots S_{N_S}$ に含まれる先行詞候補のうち，“and”，“or”，“,” のいずれかで連続表記される同一固有表現をグループ化する．

（5） 照応詞 r_i に対する先行詞候補 a_{ij} の確信度スコア C_{ij} を計算する．

（6） $C_i = \max_j C_{ij}$ となる先行詞候補 a_{ij} を照応詞 r_i に対応する先行詞と決定する．

☆ http://www.cogsci.princeton.edu/~wn/

4.3.2 確信度スコアの計算

先行詞の候補となるのはテンプレートと照合した文の前後文に含まれる名詞であるため，1 つの照応詞に対して複数の先行詞候補が存在する．したがって，どれが正しい先行詞であるかを決定する必要がある．そこで，先行詞候補の確からしさを表す指標値として確信度スコアを導入する．本論文では，確信度スコアを計算するために以下の要素を考慮する．

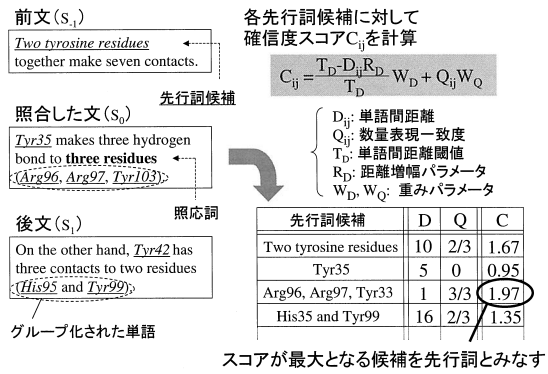
- 照応詞と先行詞候補の出現位置の近さ
  一般的に，照応詞と先行詞は比較的近い位置に現れる．このため，照応詞に対する単語間の距離が小さい先行詞ほど確からしいと判断できる．また，後方照応の場合には，前方照応に比べて先行詞と照応詞とが非常に近接して現れると考えられる．
- 数量表現の一致性
  照応詞に数量表現が用いられた場合，基本的にはその数量と同じ数だけの先行詞を参照していると考えられる．ただし，文脈によっては照応詞の数量と先行詞の数が完全に一致するとは限らない．たとえば，“Oxygen atoms contact to two residues: His35-C, His35-N and Tyr99-N.” のような文では，文脈上では “His35” と “Tyr99” の 2 つの残基が先行詞となるが，表層的には 3 つの残基が列挙されていると判断される．

これらの要素を考慮し，先行詞候補の出現位置が照応詞に近く，かつ数量表現の一致性が良いものほどスコアが高くなるようにするため，照応詞–先行詞間の単語間距離と数量表現一致度との加重和により確信度スコアを定義する．また，後方照応では前方照応に比べて先行詞が比較的近い位置に現れる特性に注目し，後方照応における単語間距離は前方照応よりも増幅させることにする．以上の考察に基づき，照応詞 r_i と先行詞候補 a_{ij} の間の確信度スコア C_{ij} を以下のように定義する．

$$C_{ij} = \frac{T_D - D_{ij} \times R_D}{T_D} \times W_D + Q_{ij} \times W_Q$$

ここで，T_D は単語間距離閾値，D_{ij} は r_i と a_{ij} との間の単語間距離，Q_{ij} は数量表現一致度，W_D と W_Q はそれぞれの変数に対する重みパラメータである．また，R_D は距離増幅パラメータであり，前方照応の場合には 1，後方照応の場合には 1 以上の値をとるものとする．以下に D_{ij}，Q_{ij} の計算方法を示す．また，**図 8** に確信度スコアの計算例を示す．

- 単語間距離 D_{ij}
  ある単語 a に対する文献の頭からの単語数を $X(a)$ とすると，



先行詞候補	D	Q	C
Two tyrosine residues	10	2/3	1.67
Tyr35	5	0	0.95
Arg96, Arg97, Tyr33	1	3/3	1.97
His35 and Tyr99	16	2/3	1.35

図 8 確信度スコアの計算
Fig. 8 Calculation of the certainty score.

表 2 評価用文献データ
Table 2 Literature data for the evaluation.

蛋白質	単語数	ページ数	正解項目数
1a0f	10034	13	198
1a0h	9534	13	95
1a0o	3677	5	120
1a0q	7335	9	152
1a3l	7209	7	110
1a4j	5496	5	71
1a5a	2844	3	62
1a5h	7444	7	140
1a5v	6039	6	97
1a5y	7668	9	170
平均	6728	7.7	118.6

$$D_{ij} = |X(r_i) - X(a_{ij})|$$

- 数量表現一致度 Q_{ij}
  r_i に数量表現が含まれる場合，以下の計算で求める．ただし，ある単語 a に含まれる数量表現が表す数を $Y(a)$ とする．

$$Q_{ij} = \frac{\min(Y(r_i), Y(a_{ij}))}{\max(Y(r_i), Y(a_{ij}))}$$

  r_i に数量表現が含まれない場合，r_i と a_{ij} の単数形・複数形が一致していれば $Q_{ij} = 0.5$，一致していなければ $Q_{ij} = 0$ とする．

5. 評価実験

提案手法の有効性を確認するために，実際の蛋白質構造解析に関する文献を使って評価実験を行う．本実験では，蛋白質立体構造データベースである PDB (Protein Data Bank)☆によって参照されている文献 10 編を対象とする．使用する文献の蛋白質コードとページ数，単語数，および正解項目数を**表 2** に示す．ここで正解項目とは，活性部位情報を図 1 に示すよう

☆ http://www.rcsb.org/pdb/

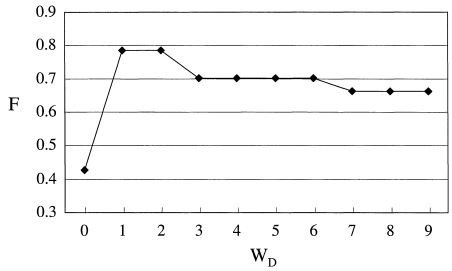


図 9 単語間距離の重み W_D の調整結果
Fig. 9 The parameter tuning result of W_D.

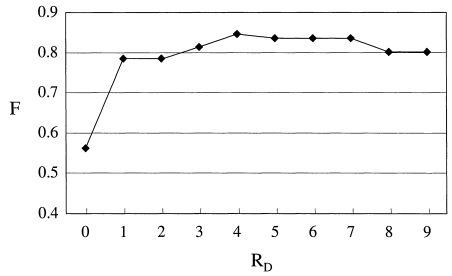


図 10 距離増幅パラメータ R_D の調整結果
Fig. 10 The parameter tuning result of R_D.

なテーブルにあてはめたとき，1 つのセルに対応するものであり，ある文から必要な情報すべてを抽出できていなくても，セル単位での抽出結果が合致していればその部分は正解と判定する．なお，以下の実験では本方式自身の有効性を評価するため，人手により正確に固有表現タグを付加した文献を入力とする．また，本実験で用いるテンプレートは，表 2 の文献とは異なる 15 編の文献から作成した 333 個である．

評価は，下記により定義される精度 P と再現率 R，F 値 F[17] を用いて実施する．

$$P = \frac{CR}{SS}, \quad R = \frac{CR}{GD}, \quad F = \frac{2 \times P \times R}{P + R}$$

ここで，GD は抽出が望まれる正解項目数，SS はシステムが抽出した項目の総数，CR はシステムが抽出し，かつ正解に含まれる項目数を表す．

5.1 照応解析手法の評価実験

5.1.1 パラメータ調整実験

提案手法におけるパラメータ W_D，W_Q，R_D を決定するための実験を行った．実験対象としたのは，表 2 に示すものとは異なる 6 個の蛋白質（“1a26”，“1a3r”，“1a5i”，“1a5z”，“2a2g”，“2a39”）に関する文献である．これらの文献に対し，$W_Q = 1$ とし，W_D と R_D のぞれぞれを変化させたときの照応解析の F 値を計算した．なお，W_D 調整時には $R_D = 1$，R_D 調整時には $W_D = 1$ とした．それぞれの実験結果を**図 9**，**図 10** に示す．これらの図より，高い F 値が得られた $W_D = 1.5$，$R_D = 4$ の値を採用する．また，N_S と T_D に関しては，6 編の文献中で最も照応詞から離れている先行詞に対して単語間距離を計算し，$N_S = 5$，$T_D = 152$ と決定した．

表 3 照応解析の実験結果
Table 3 Evaluation of the anaphora analysis.

蛋白質	GD	SS	CR	P	R	F
1a0f	23	22	21	0.95	0.91	0.93 (0.89)
1a0q	23	22	17	0.77	0.74	0.76 (0.67)
1a0o	11	11	9	0.82	0.82	0.82 (0.44)
1a3l	10	10	7	0.70	0.70	0.70 (0.44)
1a4j	8	8	7	0.88	0.88	0.88 (0.38)
1a5a	4	4	2	0.50	0.50	0.50 (0.50)
1a5h	4	4	4	1.00	1.00	1.00 (1.00)
1a5v	4	4	3	0.75	0.75	0.75 (0.75)
1a5y	3	3	3	1.00	1.00	1.00 (1.00)
平均	10.0	9.8	8.1	0.82	0.81	0.81 (0.67)

5.1.2 精度評価実験

提案手法を表 2 に示す文献に適用し，その精度と再現率を評価する．照応解析手法を適用した結果を**表 3** に示す．なお，表 3 中の括弧付きで示された値は，数量表現を考慮しない場合の結果である．また，表 2 中の蛋白質 “1a0h” については活性部位情報を含む文中に照応関係が存在しないため，本評価実験の対象外としている．

表 3 より，提案手法の F 値が数量表現を考慮しない場合と比べて上昇したことから，提案手法の有効性を確認することができた．照応解析の成功例としては，“1a3l” の文献で，照応詞 “three residues” に対応する先行詞として，それと同文中にある “Asn L91” ではなく，その前文にある “Tyr L36, Asp H50, and Asn L91” を正しく認識できていた．

一方，失敗例としては，“1a4j” の文献で，以下の文における照応詞 “nine haptens” に対応する先行詞が前文中の “BSA” と判定された．

- . . . conjugated to bovine serum albumin (*BSA*) (Table 2 and Fig. 3). Both antibodies bind *nine haptens* containing a broad range. . .

しかしながら，正解の先行詞は図中に存在する 9 個の hapten であり，本文中に先行詞は存在していない．これに対処するためには，本文テキストだけでなく，図のキャプションあるいは図中のテキストまで先行詞の探索範囲を広げる必要があり，今後の課題の 1 つである．

また，“1a3l” の文献では，以下の文における照応詞 “the antibody” に対する先行詞が “keyhole limpet hemocyanin” と判定された．

- Antibody 13G5 was raised to 7, which had

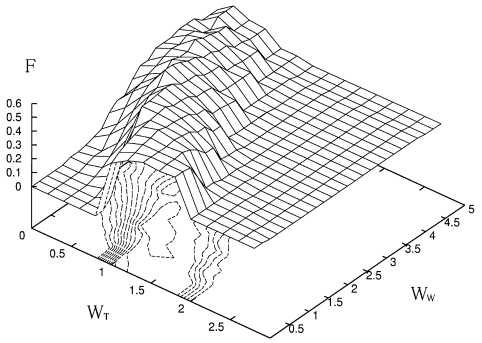


図 11 マッチングスコアの重みパラメータ W_T，W_W の調整結果
Fig. 11 The parameter tuning result of W_T and W_W.

表 4 情報抽出の実験結果
Table 4 Evaluation of the information extraction.

蛋白質	SS	CR	P	R	F
1a0f	173	132	0.76	0.67	0.71 (0.63/0.56)
1a0h	102	71	0.70	0.75	0.72 (0.72/0.70)
1a0o	53	49	0.92	0.41	0.57 (0.48/0.45)
1a0q	143	93	0.65	0.61	0.63 (0.51/0.50)
1a3l	100	79	0.79	0.72	0.75 (0.69/0.53)
1a4j	92	52	0.57	0.73	0.64 (0.59/0.54)
1a5a	46	39	0.85	0.63	0.72 (0.67/0.58)
1a5h	119	101	0.85	0.72	0.78 (0.75/0.57)
1a5v	97	66	0.68	0.68	0.68 (0.64/0.61)
1a5y	171	134	0.78	0.79	0.79 (0.77/0.71)
平均	109	79	0.75	0.67	0.70 (0.65/0.58)

been linked through its carboxylic acid group as an amide to *keyhole limpet hemocyanin. The antibody* catalyzes the disfavored exo Diels-Alder...

しかし，“keyhole limpet hemocyanin” は蛋白質の集団を表す蛋白質ファミリの名称であるため，不正解である（正解は “Antibody 13G5”）．すなわち，蛋白質ファミリにも蛋白質名と同様の <protein> タグを付加しているために失敗したと考えられることから，精度の向上に向けて，より厳密な固有表現の同定が望まれる．

5.2 情報抽出の評価実験

5.2.1 パラメータ調整実験

提案手法におけるパラメータ W_T，W_W，T_S を決定するための実験を行った．実験対象としたのは，照応解析のパラメータ調整で用いたものと同じ 6 個の蛋白質（“1a26”，“1a3r”，“1a5i”，“1a5z”，“2a2g”，“2a39”）に関する文献である．$T_S = 1.0$ とし，W_T，W_W をそれぞれ変化させたときの情報抽出の F 値を計算した結果を**図 11** に示す．この図より，高い F 値が得られた $W_T = 1.0$，$W_W = 1.4$ の値を採用する．

5.2.2 精度評価実験

提案手法を用いて文献から情報抽出を行い，精度と再現率を評価する．実験対象は表 2 に示した文献 10 編である．検証事項は，構文解析と照応解析を行わない場合，構文解析のみを行う場合，そして構文解析と照応解析の両方を行う場合のそれぞれについて，情報抽出の正確さがどの程度変化するかである．実験結果を**表 4** に示す．なお，表 4 中の括弧内に示された値は，右側が構文解析と照応解析を行わなかった場合，左側が構文解析のみを行った場合の結果である．

実験結果より，“1a3l” や “1a5h” の文献では主に構文解析による F 値の改善が見られ，“1a0q” の文献では主に照応解析による F 値の改善が見られる．また，全体的に構文解析と照応解析による精度の改善が見られることから，提案手法の有効性を確認することができる．情報抽出の成功例としては，“1a3l” の文献では以下の文とテンプレートが照合し，相互作用情報（相互作用：“additional hydrogen bond”，残基：“TyrL36”...）が得られた．

- Residue TyrL36, which is directed toward the dienophile portion of the inhibitor, is oriented by an additional hydrogen bond from...
- <residue> be oriented by <interaction>...

これは複文と単文用テンプレートが正しく照合できた例である．

一方，複文と単文用テンプレートが照合せず，情報抽出に失敗している例も見られた．“1a5y” の文献における例を以下にあげる．

- We now describe one such study of the antibody 39-A11 (6), which catalyzes a Diels-Alder reaction, a widely used...
- <protein> catalyze <reaction>

この例では，構文解析で “catalyzes” が名詞句の一部として誤認識されていたため，単語間のリンク作成の処理において “antibody 39-A11” と “catalyzes” をリンクすることができず，“which” をテンプレートと照合しようとして失敗している．本実験で用いた文献における構文解析結果を調査したところ，活性部位情報記述箇所における解析精度は 62.5%であったことから，構文解析精度向上によって，情報抽出全体の精度のさらなる改善が期待できる．

また，不要な情報が抽出されている例も見られた．“1a0q” の文献では，以下の文とテンプレートから相互作用情報が抽出されている．

- Ionic interactions in other antibodies occur between ArgH94 and AspH101...
- <interaction> in <protein> occur between

<residue> and <residue>

この文では同一蛋白質に属する 2 つの残基間の作用について記述されているが，この蛋白質では，内部の残基間相互作用よりも化合物との相互作用の方が重要であるため，不要な情報である．しかしながら，表層的な情報の照合のみでは残基間の作用が重要であるかどうかを判定することができないため，誤って抽出されている．このような例に対処するため，残基がどのような蛋白質に属しているかを判断する工夫が必要となる．

6. 結　　論

本論文では，文献から蛋白質活性部位情報を自動的に抽出するためのテンプレートマッチング手法と照応解析手法について論じた．蛋白質文献では複文が多く見られることに留意し，複文と単文用テンプレートを照合可能とするために構文解析を利用した．また，照応詞の数量表現に着目し，先行詞候補をグループ化して両者間の数量表現一致度を計算することで先行詞を特定する照応解析手法を提案した．今後の課題を以下に示す．

- 照応解析において，先行詞が図表部分に存在することがあるため，図表にどのような内容が記述されているかを認識するための工夫が必要となる．
- 本論文では活性部位情報として主に名詞を扱ったが，抽出すべき情報には文そのものや文意から推定される情報が含まれる場合もある．特に機能情報は，単語のみを部分的に抽出してもその機能の全体像を理解することが難しい．このような情報の抽出に対応することが必要である．

謝辞　日頃よりご指導いただく薦田憲久教授に深謝する．本研究の一部は科学技術振興機構 BIRD および文部科学省科学研究費補助金からの助成による．

参 考 文 献

1) 日本薬学会（編）：次世代ゲノム創薬，中山書店 (2003).
2) 中村春木：構造ゲノム科学—構造生物学によるゲノム情報解析へのアプローチ，蛋白質 核酸 酵素，Vol.44, No.3, pp.112–119 (1999).
3) Ito, N., Sakamoto, H., Kobayashi, K., and Nakamura, H.: Development of PDBj-ML, *Genome Infomatics*, Vol.12, pp.508–509 (2001).
4) 福田賢一郎，角田達彦，田村あゆち，高木利久：医学生物学文献からの専門用語の抽出に向けて：タンパク質名の自動抽出，情報処理学会論文誌，Vol.39, No.8, pp.2421–2429 (1998).
5) Ono, T., Hishigaki, H., Tanigami, A. and Takagi, T.: Automated extraction of information on protein-protein interactions from the biological literature, *Bioinformatics*, Vol.17, No.2, pp.155–161 (2001).
6) Blaschke, C. and Valencia, A.: The Frame-Based Module of the SUISEKI Information Extraction System, *IEEE Intelligent Systems in Biology*, Vol.17, No.2, pp.14–20 (2002).
7) Thomas, J., Milward, D., Ouzounis, C., Pulman, S., and Carroll, M.: Automatic Extraction of Protein Interactions from Scientific Abstracts, *Proc. Pacific Symp. Biocomputing*, pp.384–395 (2000).
8) Sekimizu, T., Park, H.S. and Tsujii, J.: Ideintifing the Interaction between Genes and Gene Products Based on Frequently Seen Verbs in Medline Abstracts, *Proc. 9th Workshop Genome Informatics*, pp.62–71 (1998).
9) Palakal, M., Stephens, M., Mukhopadhyay, S., Raje, R. and Rhodes, S.: A Multi-level Text Mining Method to Extract Biological Relationships, *IEEE Computer Society Bioinformatics Conference* (*CSB '02*), pp.97–108 (2002).
10) Pazienza, M.T.: *Information Extraction*, Springer (1997).
11) 長尾　真（編）：自然言語処理，岩波書店 (1996).
12) Groze, B.J., Joshi, A.K. and Weinstein, S.: Providing a Unified Account of Definite Noun Phrases in Discourse, *21st Annual Meeting of the Association for Computational Linguistic*, pp.44–50 (1983).
13) 佐竹正臣，白井清昭，奥村　学：照応関係を考慮した新聞記事の固有表現抽出，第 8 回言語処理学会年次大会 (2002).
14) Numa, M., Kaneta, Y. and Ohkawa, T.: Automatic Classification of Proper Names in Protein-related Literatures Using Database Retrieval on WWW, *5th Conference on Computational Biology and Genome Informatics* (*CBGI '03*), pp.903–906 (2003).
15) Sekine, S. and Grishman, R.: A Corpus-based Probabilistic Grammar with Only Two Non-terminals, *4th International Workshop on Parsing Technology* (1995).
16) Brill, E.: Some Advances in Transformation-based Part of Speed Tagging, *20th National Conference on Artificial Intelligence* (1994).
17) 徳永健伸：情報検索と言語処理，東京大学出版会 (1999).

(平成 16 年 8 月 13 日受付)
(平成 17 年 1 月 7 日採録)

兼田　佳和

昭和 53 年生．平成 15 年大阪大学大学院工学研究科情報システム工学専攻博士前期課程修了．同年同大学院情報科学研究科マルチメディア工学専攻博士後期課程入学．蛋白質情報処理に関する研究に従事．

沼　　雅之

昭和 55 年生．平成 15 年大阪大学工学部電子情報エネルギー工学科卒業．同年同大学大学院情報科学研究科マルチメディア工学専攻博士前期課程入学．文献からの情報抽出に関する研究に従事．

大川　剛直（正会員）

昭和 38 年生．昭和 63 年大阪大学大学院工学研究科通信工学専攻博士前期課程修了．現在同大学院情報科学研究科マルチメディア工学専攻助教授．工学博士．知識処理，蛋白質情報処理に関する研究に従事．IEEE 等の会員．