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

2005－NL－165 (6)
2005/1/11

蛋白質立体構造データに基づく原子間距離情報を利用した文献からの蛋白質相互作用情報抽出方式

兼田 佳和, Md. Ahaduzzaman Munna, 大川 剛直
大阪大学大学院情報科学研究科

蛋白質は他分子との相互作用により機能を発現することから、相互作用情報のデータベース化が望まれている。そこで、蛋白質構造解析に関わる文献から相互作用について記述された文 (相互作用記述文) を特定し、その文から必要な情報を抽出する方式を提案する。抽出すべき文を精度良く特定するため、テキスト処理に加え、文献に対応する立体構造の実データをもとに原子間距離の利用を図る。また、抽出対象となる固有表現が蛋白質・作用対象のどちらのフィールドに所属するかを判別するため、文の構造と立体構造データを併用する。所属先特定を文献 3 編、文特定を文献 11 編に適用した結果、精度はそれぞれ 96.4%、74.3%となった。

A Method of Extracting Protein Interaction Information from Literature using Distance between Atoms in Structure Data

Yoshikazu Kaneta Md. Ahaduzzaman Munna, Takenao Ohkawa
Graduate School of Information Science and Technology, Osaka University

As protein expresses its function through interaction with other substrates, it is required to create database of protein interaction information. We propose a method of identifing the sentences that describe protein interaction and extracting significant information from them. To identify the sentence, not only the text processing but the distance between atoms in the structure data is used. The structure data is also used to determine the field of the extracted named entities. The proposed method was applied to 3 documents for the field determination and 11 documents for sentence extraction, obtaining the precision 96.4% and 74.3% respectively.

1　はじめに

蛋白質は、局所的な相互作用部位に様々な物質が結合することでその機能を発現する [1]。このことから、相互作用に関する情報のデータベース化が望まれている [2] が、相互作用情報は数万編に上る蛋白質構造解析実験に関する論文に内在しているため、手作業での抽出は困難である。

文献から情報を抽出する研究例として、テンプレートを用いる手法 [3] が提案されているが、相互作用について記述された文 (相互作用記述文) は様々な表現が用いられており、そのテンプレート化には限界がある。そこで、テンプレートに依存せずに相互作用記述文を特定し、特定した文から定型項目を抽出する方式を提案する。

相互作用記述文では、相互作用に関わっている蛋白質の部位と作用対象 (化合物、ペプチド、DNA、イオンなど)、その作用の形態について記述されるが、文中に出現する蛋白質の部位と作用対象は立体構造上での位置が近接している。そこで、蛋白質の立体構造データを利用し、立体構造上の蛋白質－作用対象間の距離を考慮することで相互作用記述文を特定する。

相互作用記述文に含まれる原子や官能基などを定型項目として抽出する際には、それが蛋白質あるいは作用対象のどちらに属するかを判別する必要がある。原子・官能基は蛋白質の部位や作用対象とともに記述されることが多いため、原子・官能基の近辺の記述内容がその手がかりとなる。また、立体構造データ中には蛋白質の部位に関する情報が含まれているため、文献中に出現する部位が蛋白質の一部かどうかを判断するために立体構造データが有用な情報源となる。そこで、文の構造と立体構造データを併用することで蛋白質・作用対象のどちらに属するかを特定する。

2　蛋白質の立体構造と相互作用情報

2.1　蛋白質の構造と相互作用

蛋白質は複数のアミノ酸残基が一本の鎖状に繋がって構成される。各残基には鎖の一方の末端から識別番

号 (残基番号) が与えられ、この番号とアミノ酸の種類 (3 文字で略記) とを用いて残基が特定される。例えば、末端から 100 番目のアラニンは “Ala-100”、“Ala^{100}” と文献内で表記される。

蛋白質は原子間に生ずる相互作用によって他の物質と結合し、その機能を発現する。このような部位は相互作用部位と呼ばれる。文献内に記述される相互作用には、“hydrogen bond”、“van der Waals contact” などがある。また、相互作用する相手 (作用対象) としては化合物、ペプチド、DNA、イオンなどがあり、それぞれ “hapten”、“PPACK”、“(5’-D(Tcc)-3’)4”、“Zn^{2+}” などの例が挙げられる。

2.2 蛋白質立体構造データ

蛋白質の立体構造データは PDB(Protein Data Bank) として公開されており、2004 年 12 月までに 2 万 8 千件以上のデータが登録されている。立体構造データの例を図 1 に示す。各行の先頭にはレコード名があり、その後にはレコード名に応じた情報が記述される。例えば、‘JRNL’ には構造解析実験に関する文献の情報、‘ATOM’ には蛋白質原子の三次元座標やその原子が属する残基、残基番号などが記載される。

蛋白質は、その立体構造データの内容により複合体と単体とに区別される。複合体は、その蛋白質を除くポリペプチド鎖 (チェイン) が存在するか、あるいは化合物の座標が登録されているものである。単体は、蛋白質の座標のみが登録されているものである。

PDB に登録されている蛋白質は、4 文字の英数字からなる PDB-ID が付加されており、これによって蛋白質を識別できる。例えば、図 1 に示す蛋白質は ‘1eft’ という PDB-ID を持つ。本稿においても、蛋白質を識別するためにこの PDB-ID を使用する。

2.3 蛋白質の文献

本研究では、PDB 立体構造データの ‘JRNL’ レコードで参照される文献を対象とする。この文献は蛋白質の立体構造解析に関する実験の詳細について論じたものであり、立体構造決定に関する測定方法や実験条件の詳細、相互作用部位の正確な位置、他分子との相互作用の種類といった情報が記述される。これらの情報のうち、本研究では相互作用部位に関する情報の抽出を狙いとする。

```
HEADER    ELONGATION FACTOR                       24-AUG-93   1EFT
                                  ⋮
AUTHOR    M.KJELDGAARD,P.NISSEN,S.THIRUP,J.NYBORG
REVDAT   1   31-AUG-94 1EFT    0
JRNL        AUTH    M.KJELDGAARD,P.NISSEN,S.THIRUP,J.NYBORG
JRNL        TITL    THE CRYSTAL STRUCTURE OF ELONGATION FACTOR EF-TU
JRNL        TITL 2 FROM THERMUS AQUATICUS IN THE GTP CONFORMATION
                                  ⋮
HET    GNP    406      32     SEE REMARK 7.
HET     MG    407       1     MAGNESIUM ++
                                  ⋮
ATOM      1  N   ALA     1      75.082  -7.178  43.255  1.00 27.28
ATOM      2  CA  ALA     1      74.276  -6.678  42.092  1.00 34.44
ATOM      3  C   ALA     1      75.143  -5.790  41.184  1.00 33.12
                                  ⋮
```

図 1: 蛋白質立体構造データの例

2.4 相互作用情報

相互作用情報とは、蛋白質と他分子との相互作用に関する情報である。相互作用情報が記述された文中には、蛋白質の相互作用部位を示す語句 (原子や残基など) が現れている。また、複合体の場合には相互作用する相手分子に関する記述も見られる。以下に複合体に関する文献中に見られる相互作用記述文の例を示す。

“The methyl group of the inhibitor is hydrogen-bonded to the oxygen atom of Ile 60.”

この文は、“inhibitor” の “methyl group” と蛋白質残基 “Ile 60” との相互作用について述べたものである。これに対し、単体の場合には作用部位に関する記述のみが多く見られる。以下に単体に関する文献中に見られる相互作用記述文の例を示す。

“The active site triad consisting of Asp 64, Asp 121 and Glu 157 plays an important role in the catalytic function.”

この文では、相互作用部位に属する三つの残基 (Asp 64、Asp 121、Glu 157) について述べられている。

相互作用記述文から得られる相互作用情報は、最終的に蛋白質側の情報、作用対象側の情報、およびそのどちらにも属さない中立情報とに分けて定型化される。蛋白質側および作用対象側の情報には、蛋白質や化合物などの名称とその部位 (原子、官能基、残基など) に関する項目が含まれる。また、中立情報には、蛋白質－作用対象間の相互作用の名称が含まれる。例えば、上記の複合体に関する例文からは、蛋白質側の情報として残基 “Ile 60” と原子 “oxygen atom”、作用対象側の情報として官能基 “methyl group” と化合

物 “inhibitor”、中立情報として相互作用 “hydrogen-bond” が得られる。

3 立体構造データを利用した相互作用情報抽出方式

3.1 相互作用情報抽出方式の概要

複合体においては、残基内の原子と作用対象の原子は立体構造上で近接している [4]。従って、文中に現れる残基とその作用対象を特定し、その間の立体構造上での原子間距離を計算すれば、相互作用記述文特定の手がかりとなる。一方、単体の立体構造データには作用対象となる原子の座標が登録されていないが、このような蛋白質でも、それに類似する蛋白質が複合体として PDB に登録されていれば、その相互作用部位と類似する部分が単体の相互作用部位と推定でき、その推定部位に関する文が候補となる。そこで本研究では、入力文献中に記述された蛋白質が複合体、単体のどちらであるかをふまえて相互作用情報を抽出することを考える。

提案方式の概要を図 2 に示す。入力される蛋白質構造解析に関する文献は事前に固有表現タグ[1]が付加されているものとする。まず、文中の各固有表現が蛋白質と作用対象 (化合物、ペプチド、DNA、イオン) のどちらに所属するかを決定する。そして、立体構造上の距離と単語ベクトルによるフィルタリング手法を用いて、その文が相互作用記述文かどうかを判定し、相互作用記述文と判定されれば定型項目を抽出する。このとき、固有表現が蛋白質に所属していれば蛋白質側のフィールドへ、作用対象に所属していれば作用対象側のフィールドへ割り当てる。なお、文献からタグ付き文書を生成する方法については文献 [5] を参照されたい。

3.2 固有表現の所属先の決定

固有表現をテーブルに割り当てる際、蛋白質側の項目か、あるいは作用対象側の項目かを判断するため、固有表現の所属先を決定する。固有表現のうち、蛋白質、化合物、ペプチド、DNA、イオンといった物質の名前を表すものに関しては、その固有表現そのものが所属する物質となるが、これらの物質の一部である

[1]分野固有の名詞。本対象では蛋白質名や化合物名、残基名などが含まれる。

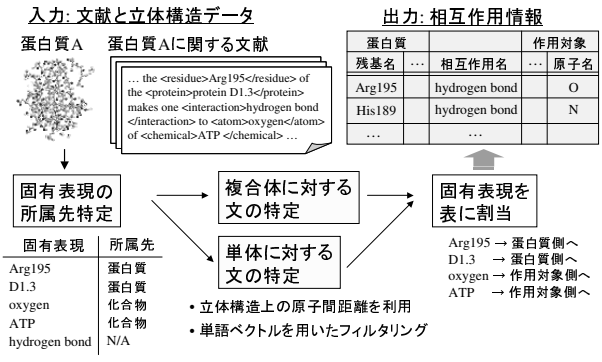

図 2: 相互作用情報抽出方式の概要

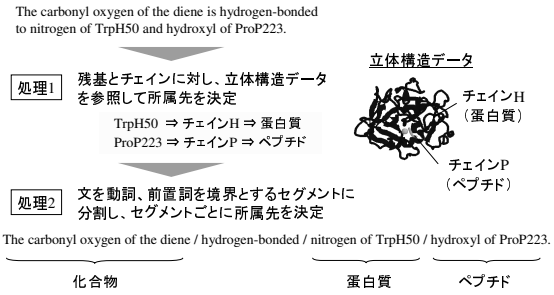

図 3: 固有表現の所属先決定方法

チェイン、残基、官能基、原子に関しては、それらがどの物質に所属するかを決定する必要がある。

固有表現の所属先決定方法の概要を図 3 に示す。チェインと残基の所属先としては、蛋白質あるいはペプチドが候補となるが、そのどちらであるかは固有表現の記述内容と立体構造データとを併用することで判断する。次に、原子と官能基は、それが蛋白質に所属する場合には文献内で残基と併記される (例えば “oxygen of Tyr35” など) ことに注目し、文を動詞・前置詞 (ただし “of” を除く) を境界とするセグメント単位に分割し、セグメント内に含まれる固有表現の種類によって所属する物質を決定する。もし同一セグメント内にチェインまたは残基が見つかった場合、原子と官能基の所属先をチェイン・残基の所属先と同一のものとする。また、化合物・ペプチド・DNA のいずれかが見つかった場合、見つかった物質を原子と官能基の所属先とする。いずれの固有表現も見つからなかった場合は、所属する物質が省略されていると考えられるので、立体構造データに含まれる作用対象すべてを所属先候補とする。

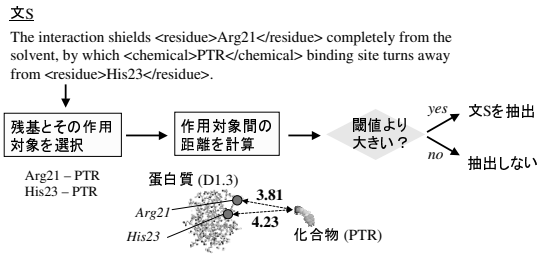

図 4: 複合体に対する相互作用記述文特定手法

3.3 複合体の相互作用記述文特定手法

3.3.1 複合体に対する特定手法の概要

複合体に対する相互作用記述文特定手法の概要を図 4 に示す。まず対象文から残基とその作用対象の候補を選択する。そして、残基とその作用対象の所属先となっている物質との間の原子間距離を計算し、ある閾値 T_{Dc} よりも小さければ相互作用記述文とする。ここで、残基や作用対象には複数の原子が含まれているが、距離の近い原子が一つでもあれば相互作用していると考えられるため、最も近接している原子間距離を残基－作用対象間の距離とする。

3.3.2 作用対象判定ルール

残基とその作用対象の候補を選び出す際に、その所属先が異なるもの全てを作用対象として選択すると、必要のない候補まで選択され、それが原因となって誤って相互作用記述文として判定されることがある。この問題を解決するため、相互作用記述文の文構造上のパターンに基づいて作用対象判定ルールを導入し、誤判定を回避する。相互作用記述文に見られるパターンと作用対象判定ルールを以下に挙げる。ただし、以下のルールにおいて [A] は固有表現タグ <A> で囲まれる文字列、[any-pn] は [residue]、[chemical]、[peptide]、[dna]、[ion]、[group]、[atom] のいずれでもよいものとする。また、*は任意個数の任意単語、~... は "..." 以外の単語を表す。

(1) グループ化に関するパターン

並列に記された複数残基と作用対象との相互作用について論じられている文では、複数残基は全て作用対象との相互作用に関わっていると考えられる。従って、並列に記され、かつ所属先が同一である残基あるいは作用対象はグループ化し、両グループを組とする。ただし、"between A and B" パターンは A-B 間に相互作用があると考えられるので、A と B を組とする。図 5 にグループ化の例を示す。

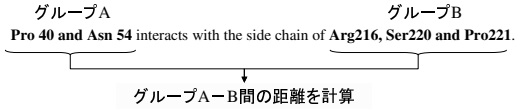

図 5: グループ化の例

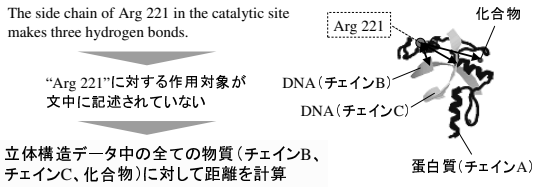

図 6: 作用対象が省略されている例

【ルール】 "between [any-pn1] and [any-pn2]" に照合した場合、[any-pn1] と [any-pn2] を組とする。"[any-pn1], [any-pn2]" または "~between [any-pn1] and [any-pn2]" に照合し、かつ [any-pn1] と [any-pn2] の固有表現と所属先が同一であった場合、[any-pn1] と [any-pn2] をグループ化する。

(2) 作用対象省略に関するパターン

作用対象が省略され、残基名のみが記されている文では、残基に対する作用対象を特定できない。従って、立体構造データに存在する全ての作用対象を組とする。作用対象の省略例を図 6 に示す。

【ルール】 " * [residue] * " が一文全体に照合し、かつ [residue] が他のルールと照合していない場合、[residue] と立体構造データ上の全ての作用対象を組とする。

3.4 単体の相互作用記述文特定手法

3.4.1 単体に対する特定手法の概要

単体に関する文献では作用対象が記述されていない文が中心となる。また、単体の立体構造データには作用対象が含まれていない。そこで、単体の相互作用部位に含まれる残基を推定することによって相互作用記述文の判定を行う。

単体に対する相互作用記述文特定手法の概要を図 7 に示す。まず単体に相同でかつ複合体として登録されている蛋白質を検索する。次に、相同蛋白質側で相互作用に関わっていると考えられる残基を決定し、それ

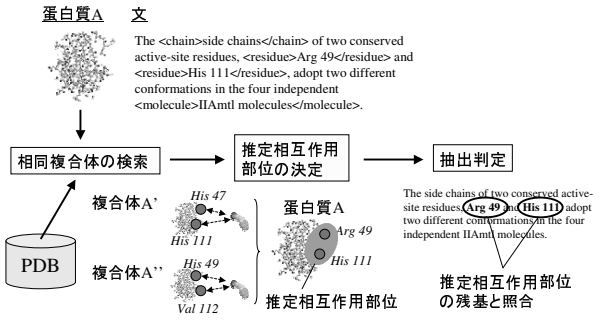

図 7: 単体に対する相互作用記述文特定手法

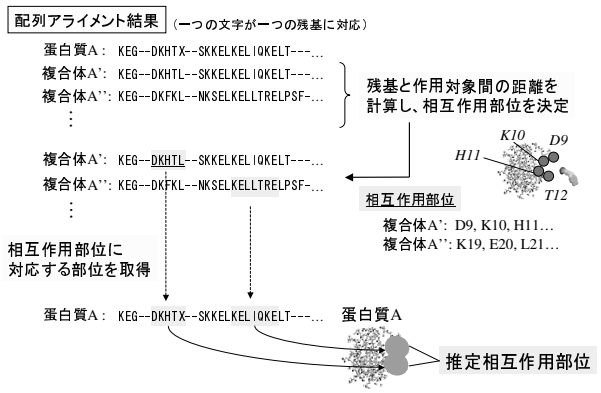

図 8: 推定相互作用部位の決定

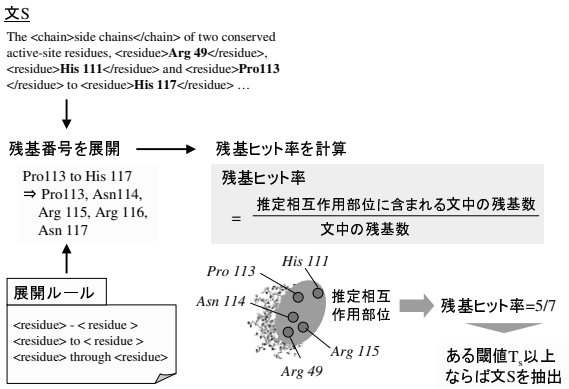

図 9: 相互作用記述文の判定方法

に対応する単体側の残基を特定することで推定相互作用部位を決定する。そして、推定相互作用部位における残基が入力文中にも記述されているかどうかを判定することで相互作用記述文かどうかを特定する。

なお、相同蛋白質の検索には “BLAST”[7] を利用する。“BLAST” は蛋白質の残基配列もとにして相同性を判定するプログラムである。

3.4.2 推定相互作用部位の決定

推定相互作用部位決定の概要を図 8 に示す。BLAST で検索した結果には複数個の配列相同な蛋白質が含まれるが、相同性が低い蛋白質から推定する相互作用部位は信頼性が低いと考えられる。従って、30%以上の相同性を持ち、その中で上位 T_N 個に含まれる蛋白質のみを相同蛋白質として利用する。相同蛋白質の相互作用部位を特定するために、その立体構造データに含まれる作用対象からの距離がある閾値 T_{Df} 以下となる残基を全て選出する。そして、配列の位置合わせ結果を利用してこれらの残基が対応付けられている単体側の残基を特定し、これを単体蛋白質側の推定相互作用部位とする。

3.4.3 相互作用記述文の判定方法

相互作用記述文判定の概要を図 9 に示す。単体の蛋白質に関する文献では作用対象が記述されないため、作用対象の判定によって不要文の抽出を回避することが難しい。従って、文中の残基のみに着目する。相互作用記述文に現れる残基は、ほとんどが相互作用に関わっていると考えられるため、推定相互作用部位の残基と文中の残基との一致率 (残基ヒット率) を計算することで相互作用記述文かどうかを判断する。残基ヒット率 R_{hit} の定義を以下に示す。

$$R_{hit} = \frac{n_i}{n_s}$$

ただし、n_s は文中の残基数、n_i は文中の残基でかつ推定相互作用部位に含まれる数である。この値がある閾値 T_S 以上であれば相互作用記述文とみなす。

3.5 不要文のフィルタリング

相互作用に関わる残基は、相互作用に関する文だけでなく、機能に関する文や文献中の図の説明に関わる文にも登場する。従って、残基による判定のみでは不要な文が数多く相互作用記述文として特定される。そこで、立体構造データを用いて相互作用記述文の候補を特定した後、単語ベースによる不要文のフィルタリングを行う。

不要文のうち、機能に関する文では蛋白質の機能や化学反応に関する用語 (例えば “catalysis”、“hydrolysis” など)、図の説明に関する文では図中における残基

の表示色や蛋白質の表示方法 (例えば “cyan”、“wire-frame” など) が多く現れる。そこで、これらをフィルタリングワードとして単語ベクトル [9] を作成し、文から作成した単語ベクトルとの間の類似の度合いを計算することで不要文のフィルタリングを行う。

フィルタリングワードから作成した単語ベクトルを v_f、文から作成した単語ベクトルを v_s とすると、下記の式によって両者の類似性 S を評価し、この値がある閾値 T_F よりも小さい場合、対象文をフィルタリングする。ただし、フィルタリングワードの個々の単語にはその重要度に応じて重みを設定し、その重みを乗じた上で類似性の計算を行う。

$$S = \frac{v_f \cdot v_s}{||v_f||||v_s||}$$

3.6 固有表現のテーブルへの割当

相互作用記述文を特定した後、その文から固有表現を抽出し、テーブルの各フィールドへ割り当る。このとき、各固有表現にはその所属先が割り当てられているので、その所属先情報に従い、蛋白質であれば蛋白質側のフィールドへ、化合物・ペプチド・DNA・イオンのいずれかであれば作用対象側のフィールドへ割り当てる。この処理を特定された相互作用記述文中に含まれる全ての固有表現に適用することで最終的な相互作用情報を得る。

4 評価実験

提案手法の有効性を確認するため、PDB から参照されている複合体および単体の蛋白質に関する文献を用いて評価実験を行う。評価は、下記の式で定義される適合率、再現率、F 値 [8] を用いる。

$$\text{適合率} = P = \frac{COR}{SYS}, \quad \text{再現率} = R = \frac{COR}{GLD},$$

$$F \text{ 値} = \frac{2 \times P \times R}{P + R}$$

ここで、GLD は抽出が望まれる正解データ数、SYS はシステムが抽出した総データ数、COR はシステムが抽出した情報で、かつ正解データに含まれるデータ数を表す。なお、提案手法そのものの精度を評価するため、固有表現タグが正確に付加された文献を入力とする。また、品詞の判定には品詞解析ツール “Brill’s Tagger”[6] を利用した。

表 1: 複合体に関する文献データ

PDB-ID	単語数	文数	ページ数	正解文数
1a0h	9534	359	13	11
1a0q	7569	389	10	18
1a26	5308	359	9	12
1a3l	3025	340	7	16
1a4k	5498	190	5	10
1a5i	8903	324	11	22
1a5y	7502	302	9	10
平均	6763	323	9.14	14.14

4.1 所属先特定の評価実験

実験対象としたのは表 1 中の “1a0h”、“1a26”、“1a5y” の蛋白質について記述された 3 編の文献であり、相互作用記述文の部分のみを評価の対象とした。実験の結果、3 編の文献中にある合計 126 個の残基・チェイン・官能基・原子の固有表現のうち、121 個 (96.5%) の所属先が正しく特定できた。失敗例としては、“1a0h” に関する文献では “O3 in NAG-1” という記述があり、原子と化合物の間に前置詞 “in” があったために所属先が正しく特定できていない。従って、包含関係を表す “in” の場合には、これをセグメントの境界と見なさない等の工夫が必要である。

4.2 複合体に対する文特定の評価実験

実験対象とした蛋白質の PDB-ID とその文献内の単語数、文数、ページ数、正解文数を表 1 に示す。また、実験結果を表 2 に示す。ただし、パラメータの設定は $T_{Dc} = 5.0$、$T_F = 1.52$ とした。また、フィルタリングには 222 個の単語を使用し、その重みは図に関する単語を 10、機能に関する単語を 1 とした。なお、表 2 の中で () 付きで示された値はフィルタリングしない場合の結果である。実験結果より、フィルタリングすることで再現率の劣化がほとんどなく適合率が向上したことから、その有効性を確認できる。

次に、蛋白質 “1a3l” に関する文献から正しく特定できなかった相互作用記述文の例を以下に示す。

“Tyrosine-L36 acts as a Lewis acid activating the dienophile for nucleophilic attack, and asparagine-L91 and aspartic acid-H50 form hydrogen bonds to the carboxylate side chain that substitutes for the carbamate diene substrate.”

この文は複文になっており、前半では “Tyrosine-

表 2: 複合体に対する提案手法の適用結果

PDB-ID	適合率	再現率	F 値
1a0h	0.625(0.556)	0.909(0.909)	0.741(0.690)
1a0q	0.750(0.692)	1.000(1.000)	0.857(0.818)
1a26	0.800(0.667)	1.000(1.000)	0.889(0.800)
1a3l	0.867(0.727)	0.813(1.000)	0.839(0.842)
1a4k	0.769(0.625)	1.000(1.000)	0.870(0.769)
1a5i	0.870(0.800)	0.909(0.909)	0.889(0.851)
1a5y	0.471(0.229)	0.800(0.800)	0.593(0.356)
平均	0.736(0.614)	0.919(0.945)	0.811(0.732)

L36” が関与する機能について、後半では “asparagine-L91” と “aspartic acid-H50” の間の相互作用について説明されている。この結果、前半部分で機能に関する単語が多く出現したために単語ベクトルが類似し、フィルタリングされたと考えられる。従って、複文の場合には文の構造情報を加味してフィルタリングの判定を行う必要があると考えられる。

また、蛋白質 “1a5y” に関する文献では、以下のような文が相互作用文として誤判定されている。

“The phenyl ring of Phe-182, immediately C-terminal to the WPD motif, is positioned exactly 7 Å above the S γ-P bond of the phosphocysteine intermediate.”

この文では、“Phe-181” の立体構造上の位置に関する情報が記述されているが、相互作用に関する情報とは直接関係がない。従って、このような記述に関してもフィルタリングするべきであると考えられる。

4.3 単体に対する文特定の評価実験

実験対象とした蛋白質の PDB-ID と文献内の単語数、文数、ページ数、正解文数を表 3 に示す。また、提案手法を適用したときの実験結果を表 4 に示す。ただし、パラメータの設定は $T_{Df} = 5.0$、$T_N = 4$、$T_S = 0.65$、$T_F = 1.52$ とした。なお、表 4 の中で () 内に示された値はフィルタリングしない場合の結果である。実験結果より、再現率の値が 0.943 となったことから、取りこぼした相互作用記述文が少なく、相互作用情報のデータベース化に対する信頼性が高いといえる。

次に、蛋白質 “1a3a” に関する文献で抽出できなかった相互作用記述文の例を以下に示す。

“The IIAmtl-binding site of HPr involves the loop comprising residues 13-21, the helix composed of residues 16-27 and the helix containing residues 48-56 (Figure 7).”

この文では、相互作用部位 “IIAmtl-binding site” が残基番号 13 から 21、16 から 27、48 から 56 で構成される部分に含まれていることが示唆されているが、この部分が実際の相互作用部位よりも広範囲であったために残基のヒット率が低下し、抽出できていない。

また、同文献では以下のような文が誤抽出されている。

“... Arg 49 is turned away and can not form hydrogen bonds with the phosphoryl group.”

この文では、残基 “Arg 49” が水素結合しないことが述べられているが、蛋白質 “1a3a” に相同な蛋白質 “1j6t” ではこれに対応する残基が相互作用に関わっていたため、誤抽出してしまったものと考えられる。このような文に対しては、相互作用を否定する記述があるかどうかを判別することで対応できると考え、今後の課題の一つである。

表 3: 単体に関する文献データ

PDB-ID	単語数	文数	ページ数	正解文数
1a03	7335	498	9	2
1a3a	8535	545	12	14
1a4l	9550	365	11	25
1a58	4338	199	6	6
平均	7440	402	9.50	12.25

表 4: 単体に対する提案手法の適用結果

PDB-ID	適合率	再現率	F 値
1a03	0.667(0.400)	1.000(1.000)	0.800(0.571)
1a3a	0.400(0.316)	0.857(0.857)	0.546(0.462)
1a4l	0.512(0.500)	0.913(1.000)	0.656(0.667)
1a58	0.857(0.667)	1.000(1.000)	0.923(0.800)
平均	0.609(0.470)	0.943(0.964)	0.731(0.625)

4.4 相互作用情報の抽出例

蛋白質 “1a0q” に関する文献における相互作用情報抽出例を以下に示す。

- 文: Arg96 from the light chain CDR L3 forms three hydrogen bonds with the hapten.
- 蛋白質側 … 残基: Arg96、チェイン: light chain、ドメイン: CDR L3
- 作用対象側 … 化合物: hapten

- 中立情報 … 相互作用: three hydrogen bonds

これは、相互作用情報が正しく抽出できた例である。また、同文献において、下記のような抽出例も見られた。

- 文: A second zinc ion binds to side-chains of AspH31 and HisH32 of CDR H1 and a third zinc ion binds to side-chains of GluL185 and HisL189.
- 蛋白質側 … 残基: AspH31、HisH32、GluL185、HisL189、ドメイン: CDR H1、チェイン: side-chains
- 作用対象側 … イオン: second zinc ion、third zinc ion

この文では、“AspH31 and HisH32” と “GluL185 and HisL189” がそれぞれ “second zinc ion” と “third zinc ion” に相互作用していることが述べられているが、テーブル化した際にその対応関係が欠落している。従って、このような場合にはそれぞれの相互作用情報を別々にテーブルに割り当てる処理が必要になると考えられる。

5 結論

本稿では、文献から相互作用情報を抽出するため、蛋白質の立体構造データにおける残基と作用対象との間の距離を利用した抽出方式を提案した。文から定型項目を抽出するために、文の記述と立体構造データを併用することで各固有表現の所属先を特定した。複合体では、文中に含まれる作用対象の全組について距離を求めると誤って特定されることがあるため、組を絞り込むために作用対象判定ルールを定義した。単体では、作用対象の三次元座標が立体構造データ中に含まれていないことから、単体と配列相同な複合体を利用することで、単体中の相互作用残基を推定した。また、不要文からの抽出を回避するため、機能や図の説明に関する単語に基づいてフィルタリングを行った。今後の課題を以下に挙げる。

- 相互作用記述文では、一文中に相互作用に関する情報と機能に関する情報が共に含まれる場合がある。このとき、単純な単語ベクトルによるフィルタリングでは必要な文が不要と見なされる。このため、複文などの文構造を加味したフィルタリングなどが必要となる。
- 文献には、「相互作用しない」などの否定形式で記述されてる文があるため、対象文が否定形式かどうかを判断する処理を追加する必要がある。

謝辞

日頃より御指導頂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] S. Goto, T. Nishioka, and M. Kanehisa: “LIGAND: Chemical Database for Enzyme Reactions,” *Bioinformatics*, Vol.14, pp. 591-599 (1998).

[2] N. Ito, H. Sakamoto, K. Kobayashi and H. Nakamura: “Development of PDBj-ML,” *Genome Infomatics*, vol.12, pp.508-509 (2001).

[3] C. Blaschke and A. Valencia: “The Frame-Based Module of the SUISEKI Information Extraction System,” *IEEE Intelligent Systems*, Vol.17, No.2, pp.14-20 (2002).

[4] M. Chanda: Atomic Structure and Chemical Bond including Molecular Spectroscopy, Tata McGraw-Hill (1972).

[5] M. Numa, Y. Kaneta, and T. Ohkawa: “Automatic Classification of Proper Names in Protein-related Literatures Using Database Retrieval on WWW,” *in Proc. of the 5th Conference on Computational Biology and Genome Informatics (CBGI’03)*, pp.903-906 (2003).

[6] E. Brill: “Some Advances in Transformation-Based Part of Speed Tagging,” *the 20th National Conference on Artificial Intelligence* (1994).

[7] S. F. Altschul, W. Gish, W. Miller, E. W. Myers and D. J. Lipman: “Basic Local Alignment Search Tool,” *J. Mol. Biol.* Vol.215, pp.403-410 (1990).

[8] M. T. Pazienza: “Information Extraction,” Springer-Verlag (1997).

[9] 徳永健伸, 辻井潤一編: “情報検索と言語処理,” 東京大学出版会 (1999).