Vol.2011-BIO-25 No.34
2011/6/24

文の接続関係を考慮した蛋白質構造解析文献からの相互作用記述文抽出方法

京極　陸[†1]　大川剛直[†1]

蛋白質の構造解析に関する研究の進展に伴い，解析の結果得られた蛋白質の機能や相互作用に関する情報が多数の文献に記載されている．これらの情報を有効活用するため，計算機による自動抽出が望まれている．そこで本研究では，文献から相互作用について記述された文を抽出する手法を提案する．このとき，ある相互作用について単独の文ではなく複数の文に跨って記述されることがある点に着目し，文同士のつながり方や接続関係などの文間の関係を考慮することで，より的確な文の抽出を実現する．

A method of extracting interaction sentence from protein structure literature considering connection between sentences

Riku Kyogoku[†1] and Takenao Ohkawa[†1]

With the development of protein structure analysis, the information about protein function and interaction is described in a lot of literatures. In this paper, we propose a method to extract interaction sentences from the literature using machine learning based approach. By focusing on the case that a piece of interaction information is described on multiple sentences, we achived the extraction of interaction sentences considering the relationship between sentences.

1. はじめに

蛋白質の構造解析に関する研究の進展に伴い，相互作用に関わっている蛋白質の部位や作用対象（他の蛋白質，DNA，化合物，イオンなど）ならびにその作用の形態が明らかにされている．これらの情報は多数の文献内に記述されており，その計算機による自動抽出が求められている．相互作用記述の抽出に関しては，蛋白質間相互作用 (PPI) の抽出[1),2)] や，局所的相互作用情報の抽出[3),4)] に関する研究が行われている．しかしこれらの研究で用いている手法は比較的単純であり，文中に存在する蛋白質等の記述や相互作用を表す動詞のパターンを用いて抽出していたり，各文を正例・負例と見なし，それぞれの文から単独で得られる特徴をもとに機械学習・分類によって相互作用記述文を抽出している．一方で，相互作用についての記述が 1 つの文だけでは明確ではなく，近接の複数の文にわたって記述されていることがあり，それらの情報は既存の手法では抽出できていない．

そこで本研究では，単文では判断できないような事例に対応するために，連続する文から得られる特徴を用いることで，文の接続関係を考慮した相互作用記述文の抽出方法を提案する．具体的には従来手法と同様に，文献中の文を機能に関する単語や頻出するパターンなどに基づいて，いくつかの特徴の集合として捉え，それを用いてその文が相互作用記述文かそうでないかを判断する分類器を訓練例から学習する．その際，単文では判断できないような事例に対応するために，連続する文から得られる特徴を考案し，接続関係を考慮した相互作用記述文の抽出を実現する．

2. 機械学習による蛋白質相互作用情報抽出

構造解析について書かれた文献は，蛋白質立体構造データが蓄積されているデータベースである PDB(Protein Data Bank)[5)] から参照できる．相互作用が記述されている文には，残基名，残基と相互作用する対象の物質名（他の蛋白質の残基や化合物など），相互作用名やその相互作用の効果などが見られる．以下に，複数の文にまたがって書かれている連続した相互作用記述文の例を挙げる[6)]．

“A zinc-binding site is formed from side-chains of AspH56 and LysH48 from one Fab and His L49 from a symmetry-related molecule. Binding of this zinc ion may, in part, be responsible for movement of the side-chain of His L49, which is located in the hapten-binding pocket, from its position in 17E8.”

機械学習によって相互作用記述文を抽出する際，既存の手法[3),4)] では，文献に含まれる 1 つの文を 1 事例として扱い，相互作用記述文を正例（positive），それ以外の文を負例（negative）とする 2 クラスの分類問題を考えている．本研究においてもこの手法に基づき、一般的な分類器である SVM[7)] を用いる．

†1 神戸大学大学院　システム情報学研究科　〒 657-8501 兵庫県神戸市灘区六甲台町 1-1
Graduate School of System Informatics, Kobe University

3. 文の属性

従来の研究[1)–4)]で用いられている属性としては，単文から得られる情報によって判断できるものが主であった．以下に例を挙げる．

(1) 相互作用記述文に頻出する単語　一般的に，相互作用記述文では相互作用名が記述されている場合が多い．そこで，相互作用についての情報を特徴付ける属性として，一般的に用いられる相互作用名が文中に含まれているかどうかを利用する．

(2) 相互作用記述文に頻出するパターン　相互作用記述文では，残基と相互作用の対象との関係を示す語として “between A and B” が用いられる場合が多く，この場合 A-B 間には相互作用があることが考えられる．また，相互作用記述文では，残基と相互作用の対象間に相互作用を意味する動詞が記述されている場合も多い．そこで，相互作用を記述する際によく用いられるこれらの文型をパターン化し，それにマッチングするかどうかを利用する．

(3) 文献中に頻出する単語　パラグラフ（文の集合，段落）ごとに抽出された重要な単語は，その文献中において重要である単語であると捉えることができる．したがって，文中にその単語が含まれているかどうかで，その文が文献中の重要な文かどうかが判断できると考える．そこで，文献中の文から自動で頻出単語を抽出し，パラグラフごとに最も重要度の高い単語を TF-IDF を用いて選出した後，それらをその文献の重要な単語と見なし，その出現を属性として利用する．

(4) 文献中に頻出するパターン　(2) において述べたパターンは，手動で定義しており形が固定されているため，それ以外の文型は検索できない．そこで，より多様なパターンに適応させるため，自動的かつ網羅的に生成したパターンを，文献中の全ての文において調査し，頻出なものをその文献の重要な特徴であると考える．したがって，自動で生成したパターンのうち頻出なものも属性として利用する．

(5) 残基間距離　相互作用記述文には，相互作用する物質の組が記述されることがある．ある残基が他の物質と相互作用するとき，残基中の原子と相互作用対象物質間の距離は近接することが知られており，この性質を属性として採用する[4)]．文中に記述されている相互作用対象物質間の距離を，三次元構造データを参照して計算し、その値がある閾値よりも小さいかどうかの判定結果を属性として利用する．

しかし，これらの属性を用いた抽出手法では 1 文を 1 事例として学習・分類するため近接する複数の文の総合的な解釈を通して相互作用情報が得られるような状況への対応が困難である．そこで本研究では，連続した文を考慮するための属性として，“接続詞”，“連続して出現する名詞または動詞”，“連続して出現する蛋白質または残基” の 3 属性を新規に採用し，前後の文との関連性をより詳細に学習する．本研究では，以下の属性を考案し手法に取り入れた．

(A) 接続詞　文と文の項目同士の関係を示す役割を担う “接続詞” は，文献中でも論理を組み立てる際に非常に重要である．本研究では，連続する文の関係性を考慮するために，文中に出現する接続詞を属性に採用する．また，前後の文がどのような関係を示すか判断するために，接続詞を “帰結”，“原因・理由”，“目的”，“追加情報”，“否定・逆接”，“強調”，“言い換え・強調” の 7 種類のカテゴリに分別し[8)]，それらも属性として利用する．

(B) 連続して出現する名詞または動詞　文献中では，接続詞の代わりに前文に出現した名詞や動詞を用いてより詳細な説明をしている文が頻出している．そこで，前文に出現した名詞や動詞が次の文にも出現しているかどうかということを属性として利用する．

(C) 連続して出現する蛋白質または残基　相互作用記述文では，前の文に出現した詳細な残基名，または蛋白質名が次の文にも出現する場合が見られ，それらの文は同じ物質に対して説明していることが多い[6),9),10)]．したがってこれらの文は同様の内容を説明していると解釈できるため，蛋白質名，または残基名が文を跨いで連続して出現するかどうかを判定し，その結果を属性として利用する．

4. 評価及び考察

実験には，PDB に登録されている立体構造データから参照される 14 種類の文献を使用した．また，使用する文献には人手により固有表現タグが付与されており，相互作用記述文もすでに特定されている．この 14 種類の文献から 1 つの文献をテスト用に相互作用記述がわからないものと想定し，他の 13 種類の文献を訓練例として学習した分類器をテスト用の文献で評価する．これを 14 種類の文献全てに行い，その平均を評価する．本研究ではこの評価実験を，単文から得られる属性 (1)∼(5) のみを用いた手法（手法 1）と，その属性に加え文の接続関係を考慮する属性 (A)∼(C) を取り入れた提案手法（手法 2）の 2 つの手法に対して行う．実験結果の評価には Precision, Recall, F 値を用いる．

接続関係を考慮した手法 2 は，接続関係を考慮していない手法 1 よりも Recall, F 値において高い精度を示した．以下に，手法 1 では抽出できなかったが手法 2 では抽出できた相互作用記述文の例を挙げる．

表 1 実験結果

手法	Precision(%)	Recall(%)	F 値 (%)
手法 1	87.55	68.72	77.00
手法 2	82.03	80.33	81.17

“In the normal course of activation, the peptide bond between Arg15 (or Lys 15) and Ile 16 (or Val16) [chymotrypsinogen numbering (13)] is cleaved to give rise to a new N-terminal segment NH3+-Ile/Val16-Xxx-Xxx-Gly19. This segment inserts into the body of the proteinase, allowing formation of a buried salt bridge between the free amino terminus and the carboxylate group of Asp194, which is relocated from a solvent-exposed position close to the active site to an internal position at the base of the Ile16 pocket.”

これは単文で得られる属性だけではなく，連続する文との関係性を考慮した属性を使用したことで，前後の文との接続関係を学習に用いることができ，数文で構成される相互作用情報を抽出できたためであると考えられる．Recall が高いことからも，これまで抽出できなかった相互作用記述文を抽出できたことが確認できる．

一方，手法 2 を用いてもうまく抽出できない場合があった．文の属性の中でも，とりわけ残基間距離は重要な役割を果たすが，以下に例示するような作用の対象が明確に記述されていない文に対しては残基間の距離が計算できない．

“In addition, a histidine residue interacts with the phosphonate oxygen atoms in each of the structures.”

この場合，a histidine residue がどの残基を指すのか，the phosphonate oxygen atoms がどの原子を指すのかについて前後の文を積極的に活用して特定することでこれまで計算されなかった残基間の距離が計算でき，精度の向上につながると考えられる．

5. おわりに

本論文では，蛋白質構造解析文献を対象として機能情報が記述された部分を文単体ではなく，いくつかの連なった文集合として捉えた相互作用記述文抽出方式について論じた．連続した文を考慮するための属性として，“接続詞”，“連続する文におけるパターン”，“連続して出現する蛋白質または残基” の 3 属性を新規に採用し，前後の文との関連性をの学習を可能とした．実験の結果，これらの 3 属性を用いて接続関係を考慮した手法は，接続関係を考慮しない手法よりも良い精度を示し，より多くの相互作用記述文を抽出できた．今後の課題として，作用対象となる残基が不明な場合にその残基を特定し，相互作用対象物質が明確に記述されていない文に対しても残基間距離を計算することで，精度の向上を図っていくことが挙げられる．

参 考 文 献

1) Q. C. Bui, S. Katrenko and P. M. A. Sloot, “A hybrid approach to extract protein-protein interactions”, *Bioinformatics*, Vol. 27, No. 2, pp. 259–265 (2011).
2) M. Miwa, R. Sætre, Y. Miyao and J. Tsujii, “Protein-protein interaction extraction by leveraging multiple kernels and parsers”, *International Journal of Medical Informatics*, Vol. 78, Issue 12, pp. e39-e46 (2009).
3) H. Liu, C. Blouin and V. Keselj, “Sentence identification of biological interactions using PATRICIA tree generated patterns and genetic algorithm optimized parameters”, *Data & Knowledge Engineering*, Vol. 69, Issue 1, pp. 137-152 (2010).
4) 兼田 佳和, Md. A. Munna, 大川 剛直, “蛋白質立体構造データを利用した文献からの蛋白質相互作用記述文抽出方式”, 電気学会誌 C 部門論文誌, Vol.125, No. 5, pp. 690-697 (2005).
5) 金久 實, “ポストゲノム情報への招待”, 共立出版 (2001).
6) J. L. Buchbinder, R. C. Stephenson, T. S. Scanlan and R. J. Fletterick, “A Comparison of the Crystallographic Structures of Two Catalytic Antibodies with Esterase Activity”, Journal of Molecular Biology, Vol. 282, pp. 1033-1041 (1998).
7) 平尾 努，磯崎 秀樹，前田 英作，松本 裕治，“Support Vector Machine を用いた重要文抽出法”, 情報処理学会論文誌，Vol. 44, No. 8, pp. 2230-2243 (2003).
8) 大山 研司，“科学論文に役立つ英語”, (2001).
9) P. D. Martin, M. G. Malkowski, J. Box, C. T. Esmon, and B. FP. Edwards, “New insights into the regulation of the blood clotting cascade derived from the x-ray crystal structure of bovine meizothrombin des f1 in complex with ppack”, *Structure*, Vol. 5, pp. 1681-1693 (1997).
10) M. Renatus, M. T. Stubbs, R. Huber, P. Bringmann, P. Donner, W.-D. Schleuning, and W. Bode, “Catalytic Domain Structure of Vampire Bat Plasminogen Activator: A Molecular Paradigm for Proteolysis without Activation Cleavage”, Department of Structural Research, Max-Planck-Institute of Biochemistry D-82152 Martinsried, Germany, and Research Laboratories, Schering AG, Mullerstrasse pp. 170-178 (1997).