英語論文から抗体情報を取得する検索モデル

木 下 直 彦

新潟大学大学院医歯学総合研究科
生体機能調節医学専攻腎科学大講座構造病理学
(指導:山本格教授)

Search Model to extract antibody information from English articles

Naohiko Kinoshita

Niigata University Graduate School of Medical and Dental Science
Course for Biological Functions and Medical Control
Institute of Nephrology Structural Pathology
(Director: Prof. Tadashi Yamamoto)

要旨

【はじめに】抗体はバイオ分野の実験において重要な研究ツールのひとつである．しかし，一つの抗原に対して複数の抗体が作製，販売されていることも多く，研究者は抗体の選定や判断に困難をきたしている．そこで，論文内，特に英語論文からタンパク質に対する抗体情報を取得するための検索モデルを作成し，モデルに基づいた抗体情報自動検索ツールを作成し，その性能を検証した．

【材料と方法】初めに英語論文の中から，抗原としてのタンパク質名および，製造会社が記載されている語の出現位置をリスト化した．その後，製造会社の出現位置を限界点とする区間を設定し，その区間を終点限界点に位置する会社のメタデータ空間と定め，メタデータ空間内にあるタンパク質名を抗原とする抗体-製造会社の関係性を取得する抗体情報検索モデルを作成した．その後，UniProt[1)]のタンパク質データベースを用い，自動的に抗原となるタンパク質を検索/検出し，製造会社と関連付け，リポジトリ化する抗体情報自動検索ツールを作成し，その性能の検証を行った．

【結果】抗体情報検索モデルの検証を行ったところ，適合率 100%，再現率 99.0%，F 値 99.5%という結果であった．次に，抗体情報検索ツールの性能の検証を行ったところ，適合率は 100%であったが，再現率が 39.8%と大きく減少し，結果として F 値は 56.9%と著しい信頼度の低下が見られた．そこで，その原因

Reprint requests to：Naohiko Kinoshita　**別刷請求先：**〒951－8510　新潟市中央区旭町通 1－757
Structural Pathology Niigata University　新潟大学構造病理学教室
1－757Asahimachi－dori　Chuo－ku,　木下　直彦
Niigata　951－8510　Japan

を調査したところ，抗体情報が取得できなかった原因の 72．3%は，動物種が異なる同名タンパク質が UniProt データベースに登録されていたため，タンパク質を特定できなかったものであった。そのため、データベ。スを動物種別に再構築し，再度検証を行ったところ，再現率は 79．6%まで上昇し，F 値も 88．6%という高い信頼率となった．

【考察】今回の抗体情報検索モデルについては F 値が 99.5%という高い信頼性を示したため，このモデルは抗体情報取得に有効であると考えられた．また，抗体情報検索ツールでは，当初再現率が 39.8%と，著しい低下がみられたが，動物種ごとにタンパク質データセットを編成し，再検証した結果，再現率が 79．6%まで上昇し，F 値も 88．6%という信頼率まで改善された．また，再構築後の抗体情報検索ツールについて，論文別の信頼度調査を行ったところ，F 値が 0.8 以上の論文が全体の 80.9%を占めたことから，本研究で作成した抗体情報検索ツールは，研究者が抗体選定する際に有用であると考えられた．今後この仕組みを改良/発展させることで、さらに多くの研究者に有用な抗体情報検索ツールを構築してゆきたいと考えている．

キーワード：抗体，情報検索，バイオインフォマティクス

はじめに

抗体は免疫蛍光法などによるタンパク質の細胞，組織局在，ウエスタンブロッティングによる検出，ELISA 法による定量，フローサイトメトリーによる細胞の分離，分析など，バイオ分野の研究において最も重要な研究ツールのひとつである．

現在，抗体は様々な会社や研究者が製造しているため，一つの抗原に対して複数の抗体が存在していることも多い．先に述べたように，抗体の品質や性能は実験の根幹にかかわる重要項目であり，特に免疫蛍光染色においては抗原の検出の可否によってはやり直しを余儀なくされてしまうこともある．また，抗体は高価であるため，数種類を試すことは経済的にも負担が大きい．これらの理由から，研究者は時間をかけて抗体の選定を行うが，期待通りの結果が得られないことも多い．

研究者が抗体を選定する際の指標には以下のものがあげられる．

1．研究者自身の経験
2．熟練者のアドバイス
3．抗体精製会社が提示しているデータ（染色画像など）
4．過去の使用論文

指標１，２については，経験則に基づくものであり，自身にその経験がなく，また身近に熟練者が不在の場合，指標になり得ない．また指標３については研究者にとって有用であるが，会社が提供するデータ通りの結果が得られるかどうか分からない．そこで，本研究では指標４として挙げられた抗体を使用した過去の英語論文に着目し，実験等に使用された抗体ごとに製造会社名を自動的に抽出し，リポジトリ化することで，様々な研究者による抗体の使用実績が可視化され，それが抗体選定の指標となり得るのではないかと考えた．

英語論文内から抗体情報を検索する際、情報処理分野の研究において，パターンマッチング，ベクトル空間を用いた検索モデル，検数理モデル化など、様々な情報検索手法が提案されており，特に，計算言語学分野の研究においてはシソーラスに基づく単語間の意味関係の有無判定 [2)]，コーパスにおける共起性に基づく語彙のクラスタリング

[3] など客観的で再現性のある手法が提案され，有用と考えられてきた．本研究における抗体検索に着目した場合，検索空間が英語で記載された公開論文，検索対象は抗体情報という限定的な資源に特化した情報検索ツールとなるため，客観的なモデルに比べ，信頼度の高い検索モデルが要求される．

一般的に検索システムにおける信頼度の測定には二つの指標が用いられている．一つは適応率(precision)で正確性を表す指標として,以下の式で表現される．

$$precision = R/N$$

R：適合検索後数(正しく検索されたデータの個数)

N：検索結果数(検索されたすべてのデータの個数)

もう一つは，再現率(recall)で網羅性を表す指標として，以下の式で表現される．

$$recall = R/C$$

R：適合検索後数(正しく検索されたデータの個数)

C：正解検索数(正しく検索されるべきデータの個数)

適合率と再現率は一般的にトレードオフの関係にあり，この２つの値の相加平均を使用したＦ値(F-measure)を用いて最終的な信頼度を表す．

$$F\text{-}measure = \frac{2 \cdot precision \cdot recall}{precision + recall}$$

本研究においても，この３つの値を用いて抗体情報検索モデル，抗体情報検索ツールの信頼度の測定を行った．

方法

１．抗体検索モデルの策定

言語学者の Zelling Harris は，文章内の単語や形態素の違いはその分布は狭い範囲の時に相関関係が見られることに言及している [4]．このことは自然言語処理系における単語同士の繋がりの強さは出現位置の距離が近いほど関係性が高いことを示している．英語論文内における抗体の情報，特に抗原とその製造会社の位置関係においてもその傾向は顕著である(図 1)．

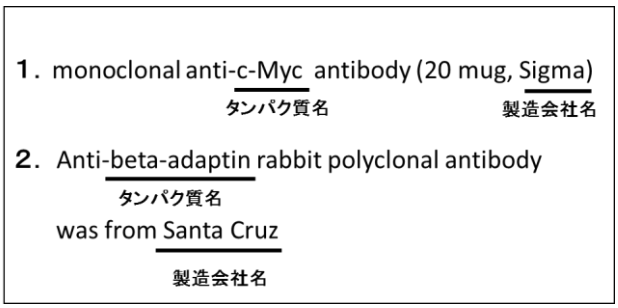


図１　タンパク質名と会社名の位置関係の例

1．参考文献 [5] より引用，2．参考文献 [6] より引用

さらに，英語構文の性質上，製造会社は抗体名より後ろに存在することが多いと推測されることから製造会社の出現する位置を限界点とする区間をメタデータ空間と定義し，終点限界点に位置する会社のメタデータがその区間内に存在していると仮定した（図 2)．

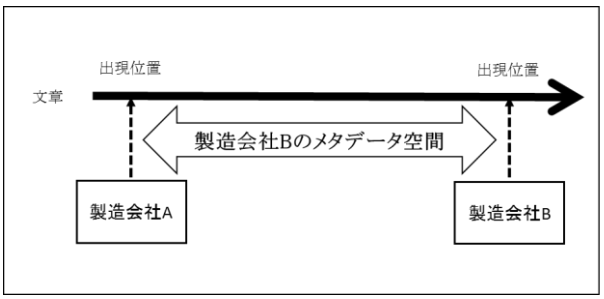
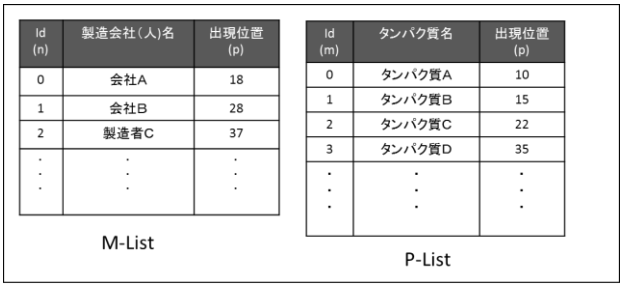


図２　抗体情報検索モデル

まず，論文内から製造会社(製造者)，タンパク質名の単語が出現する位置を取得し，リスト化した．これらをそれぞれM-List ，P-List とする．M-List[n]に着目した場合，関連する抗体はM-List[n-1]の出現位置[p]から M-List[n]の出現位置[p]に存在すると考え,この区間内にあるタンパク質名をM-List から抽出し,関連付けを行った.



Id (n)	製造会社(人)名	出現位置 (p)
0	会社A	18
1	会社B	28
2	製造者C	37
・ ・ ・	・ ・ ・	・ ・ ・

M-List

Id (m)	タンパク質名	出現位置 (p)
0	タンパク質A	10
1	タンパク質B	15
2	タンパク質C	22
3	タンパク質D	35
・ ・ ・	・ ・ ・	・ ・ ・

P-List

図３　M-List と P-List

考察１では，この抗体検索モデルの信頼度について適合率，再現率，Ｆ値の検証を行った.

２．抗体情報検索ツール

策定された抗体情報検索モデルを基にして,実際に複数の論文から自動的に抗体情報の抽出を行うツールを作成した．このツールは，UniProt Consortium が作成しているUniProtKB のタンパク質データベースから取得したタンパク質名データセットを用い，複数の論文から，自動的にタンパク質名と照合処理を行い，製造会社と関連付けを行うものである(図４).

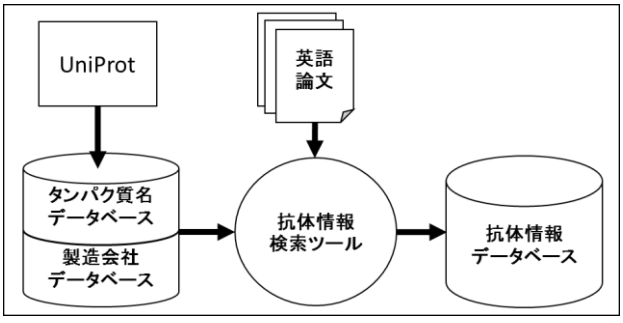


図４ 抗体情報検索ツール

検証２ではこの検索ツールの信頼性について調査を行った．また，本研究において検証に使用する論文については，BMC(National Center for Biotechnology Information)が提供しているPMC(旧 PubMed Central)[7]の Open Access Subset 内の論文の中から，抗体を使用している頻度が比較的高く，引用頻度が高い 21 論文を選定し，その中に記載されている 103 抗体の情報抽出について検証を行った.

結果

検証１：検索モデルの信頼性

今回策定した検索モデルの信頼性の結果は表１の通りであった.

適合率 (precision)	再現率 (recall)	F値 (F-measure)
1.00	0.99	0.995

表１　検索モデルの信頼性

また，再現率が 1.0(100%)にならなかった原因を調査したところ,実験プロトコル内の抗体情報に会社名の記載がない論文が１件あることが確認された.

検証２：抗体取得ツールの信頼性

抗体情報検索ツールの信頼性の検証結果を表 2 に示す.

適合率 (precision)	再現率 (recall)	F値 (F-measure)
1.00	0.398	0.569

表 2　抗体取得ツールの信頼性

適合率はモデルの信頼度と同様, 1.0(100%)であったが, 再現性については 103 抗体のうち 41 抗体(39.8%)が検出されたのみであった. その原因を調査するため, 論文単位で適合率, 再現率を調査した（表 3).

論文3	適合率 (precision)	再現率 (recall)	F値 (F-measure)
A	1.000	1.000	1.000
B	1.000	1.000	1.000
C	1.000	1.000	1.000
D	1.000	1.000	1.000
E	1.000	0.750	0.857
F	1.000	0.600	0.750
G	1.000	0.565	0.722
H	1.000	0.500	0.667
I	1.000	0.500	0.667
J	1.000	0.429	0.600
K	1.000	0.308	0.471
L	1.000	0.286	0.444
M	1.000	0.222	0.364
N	-	0	0
O	-	0	0
P	-	0	0
Q	-	0	0
R	-	0	0
S	-	0	0
T	-	0	0
U	-	0	0
全体値	**1.00**	**0.398**	**0.56.9**

表 3 論文別信頼度調査

その結果, 論文によって, 網羅率の指標となる再現率の最高値(100%)と最低値(0%)の論文が多くみられ、分散値が大きい結果となった. また, 再現率低下の直接原因となる未検出 62 抗体の理由を調査した結果, 図 5 の通りとなった.

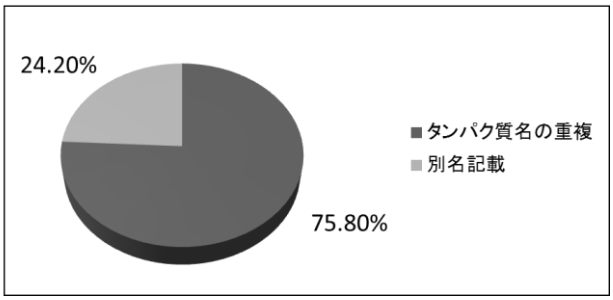


図 5　抗体の未検出原因

「タンパク質名の重複」とは, UniProt データベース内で該当するタンパク質名が複数存在していたため, その特定ができなかったものを指す. 例えば, Clathrin heavy chain という記載があった場合, UniProt で検索するとタンパク質の ID として, P25870 (CLH_DICDI), P22137 (CLH_YEAST) などが検出され, タンパク質の特定が出来なかった(図 6).

「別名記載」とは UniProt データベース内に記載されていたタンパク質名とは異なる文字列が論文内に記載されていたものである. 例えば, 論文内では AP50 となっているが, UniProt のタンパク質名としては AP-50 と記載されていた.

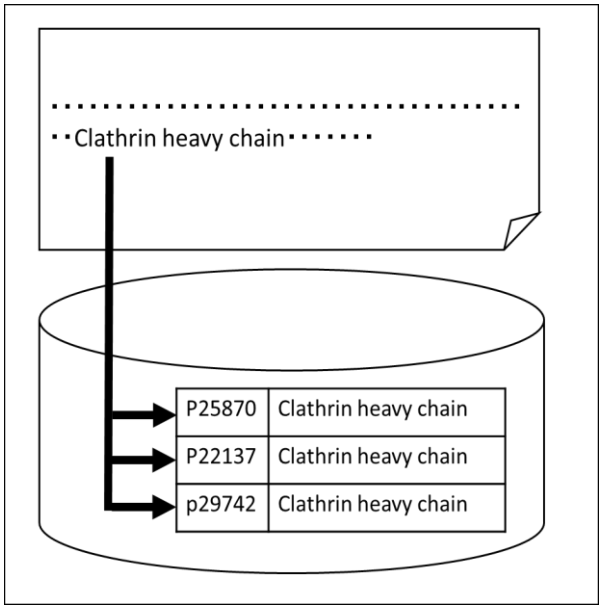


図 6　タンパク質名の重複

未検出原因の 75%を占めるタンパク質名の重複の問題をさらに詳細に分析したところ,動物種の違いによって引き起こされているものが多く存在しており，種を限定し，例えば，種をヒトだけなどに限定すると単一になることが分かった．そこで，UniProt データベースを動物種ごとに分解/再構成したデータセットを用いて，ツールの改善を行った後，再度検証を行ったところ，再現率が 79.6%まで改善され，それに伴いＦ値も 88.6%と高信頼性を示す値となった(表 4)．

適合率 (precision)	再現率 (recall)	F値 (F-measure)
1.00	0.796	0.886

表 4　改善後の信頼度

また,改善前と改善後の論文ごとのＦ値の変化を調べたところ，0.8 以上である論文が 66.6%から 80.9%へ増加した．(図 7)．

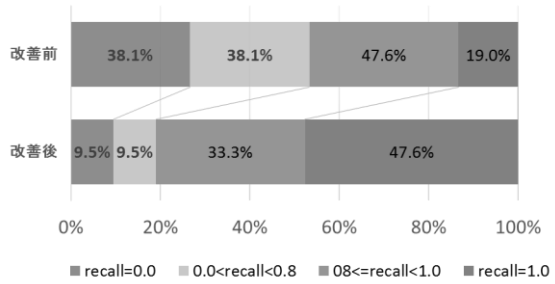


図 7 改良によるＦ値の比較

考察

検証 1 の結果，Ｆ値が 99．5%と高い信頼性が確認され，特に適合率については 100%というすべてが適合する結果となった．これは，製造会社メタデータ空間内に存在しているタンパク質名が，すべて抗原情報であることを示しており，高い精度で検索できることを示している．これらの結果から，本研究で採用した抗体情報検索モデルは信頼度の高いものであると考えられた．

また，検証 2 における抗体情報検索ツールの信頼度は，適合率は抗体情報モデルと同様に 100%となったが，再現率は 39.8%と著しい低下がみられ，結果的にＦ値も 56．9%となり，網羅性の低い結果が顕著に示された．

そこで,再現率の低下の直接原因となる未検出抗体の原因調査を行った結果，UniProt データ上での異種動物で同名のタンパク質の存在による特定不可能なものが 75.8%を示しており，動物種ごとに分解/ 再構成して,再度検証を行ったところＦ値が 88.6%まで改善することが出来た．

また，再構築した抗体情報検索ツールの論文別の信頼度調査を行ったところ，Ｆ値が 0.8 以上の論文が全体の 80.9%を占めたことから，本研究で作成した抗体情報検索ツールは，研究者が種を限定した抗体選定する際に有用であると考えられた．

このように，種を特定することで，例えばヒトのタンパク質に対する抗体と製造会社については、文献情報から検索できる仕組みを確立することが出来た．この仕組みを利用することで，研究者は容易に必要な抗体の情報を得ることが出来るようになるのではないかと考えている．

本研究では，抗体情報として，タンパク質と製造会社という情報源だけに特定したモデルの策定であるため,クローン番号やモノクローナル/ポリクローナルなど他の抗原情報については考慮されていない．また、検索ツールでは、製造会社を自動取得する仕組みについては考慮されていない．これらの問題について,今後さらなる研究を行い、この仕組みを改良，発展させることで，信頼度の高い自動取得モデルの策定を行い、多くの研究者に有用な抗体情報検索ツールを構築してゆきたいと考えている．

結論

本研究で提案した抗体検索モデルおよび研究ツールは、動物種を特定することで、高い信頼度をもって、抗体情報の抽出および検索ができることを示せた。この仕組みを利用することで、研究者は今まで以上に容易に抗体情報を取得することが出来、抗体選定に有用なツールになり得るのではないかと考えている。

今後は, より多くの論文から抗体情報を取得し, リポジトリ化することで, 様々な抗体情報の分析ができ, 研究者にとって必要な抗体情報を提供できると考えている.

謝辞

この研究を行うにあたり，研究の機会を与えていただきました新潟大学医歯学総合研究科生体機能調節医学専攻腎科学大講座構造病理学・山本格教授並びに研究室の諸先生方に心から御礼申し上げます．また，研究を支えていただきました構造病理学教室の皆様にも合わせて感謝いたします.

文献

1) UniProt (http://www. uniprot. org)
2) Morris, J. and Hirst, G.: Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational* vol.38 No.3,1991
3) Brown,P.F Della Pietra,V.J. deSouza,P. V. Lai,J.C.and Mercer,R.L.: Class-based n-gram Models of Natural Language, Computational Linguistics Vol.18 No.4 /pp.465-479,1992
4) Zelling S Harris /edited by Henry Hiz : "Papers on Syntax",1981
5) Ari J.Firestone, Joshua S :Small-molecule inhibitors of the AAA+ ATPase motor cytoplasmic dynein : Nature 2012 March 18; 484(7392): 125–129,2012
6) Mario P, Michael A.C, Nanako M, Antonio C, Laura-Jo A, Sebastian W, Dermot M. F. Cooper: Insights into the residence in lipid rafts of adenylyl cyclase AC8 and its regulation by capacitative calcium entry: Am J Physiol Cell Physiol. Mar 2009; 296(3): C607–C619, 2009
7) PMC(http://www.ncbi.nlm.nih.gov/pmc/)
8) Steven Kirsch : Infoseek's experiences searching the internet.SIGIR Forum,Vol.32, No.2,pp.3–7,1998.
9) Lawrence Page,Sergey Brin,Rajeev Motwani and Terry Winograd.The pagerank citation ranking : Bringing order to the web. Technical report, Stanford Digital Library Technologies Project, 1998.