# 博士論文の要旨及び審査結果の要旨

| | |
|---|---|
| 氏　　名 | 木下　直彦 |
| 学　　位 | 博士（医学） |
| 学位記番号 | 新大院博（医）第631号 |
| 学位授与の日付 | 平成27年3月23日 |
| 学位授与の要件 | 学位規則第4条第1項該当 |
| 博士論文名 | 英語論文から抗体情報を取得する検索モデル |
| 論文審査委員 | 主査　教授　赤澤　宏平 |
| | 副査　教授　山本　格 |
| | 副査　教授　河内　裕 |

博士論文の要旨

背景と目的：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。しかし，一つの抗原に対して複数の抗体が作製，販売されており，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，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、そ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，その性能を適応率と再現率を用いて詮索の精度について検証した。

材料と方法：初めに英語論文の中から，抗原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，作成したモデルを基にして、UniProt1)のタンパク質名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。検証に使用する論文については，NCBI(National Center for Biotechnology Information)が提供しているPMC(旧PubMed Central)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：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56.9%と著しい信頼度の低下が見られた。そこで，その原因を調査したところ，抗体情報未取得原因の75.8%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。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。

考察：抗体情報検索モデルの検証においては，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，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いきたいと考えている。

審査結果の要旨

本研究では、英文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考案し、その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評価した。

研究方法としては次の手法を採用した。最初に英文論文中の抗体名（抗−タンパク質名）と製造会社の名称の出現規則をモデル化した。次に、既存のタンパク質名のデータベースと自ら構築した抗体製造会社名のデータベースを用いた抗体情報検索ツールをＣ言語により開発した。最後に、モデルの妥当性を検証するために３つの指標（適合率、再現率、Ｆ値）を定義して、特定の抗体に関する英文論文を抽出し当該検索ツールの性能を評価した。

103 抗体に関する英文論文 21 編に基づき、３つの指標の真値を求めると、適合度 1.00、再現率 0.99、Ｆ値 0.995 であるのに対して、動物種ごとに層別するという改良を加えた検索ツールを用いることにより、適合度 1.00、再現率 0.796、Ｆ値 0.886 という結果が得られた。

結論として、本研究で考案された抗体検索モデルおよび抗体情報検索ツールは、動物種を特定することで、良い精度で抗体情報を抽出できることを示した。抗体情報の取得をこれまで以上に正確に取得するロジックと自動抽出ツールを開発した点に学位としての価値を認める。