日本語によるゲノム情報の検索と理解を助けるポータルの開発

●金子 周司
京都大学大学院薬学研究科

＜研究の目的と進め方＞

本研究では，インターネットで公表されるゲノム科学の成果を日本人が検索あるいは理解しやすくすることを目的として，過去に構築してきたライフサイエンス辞書のシソーラス資源を最大限に利用する日本語ポータルの研究開発を行う。

本研究のゴールは，第一にゲノム科学の研究成果論文やデータベースを日本人の初学者や一般人が検索する場合に，入力したキーワードと密接に関連する別のキーワードを同時に提示することによって情報検索を容易にする連想検索サーバを開発・公開することである。また第二に，検索結果として表示される英語ページにおいて，利用者が求める箇所でオンデマンドに専門用語の対訳および解説を表示して，利用者の理解を助ける汎用辞書ツールを開発することである。

この開発で用いる電子辞書とシソーラスはこれまでにライフサイエンス辞書プロジェクトで独自に構築してきたものを利用する。英語で書かれたゲノム科学情報のあらゆるWebページを日本語で検索して内容を理解できるサーバを無料で公開することは，ゲノム科学の研究成果を広く社会に提供する実用的なインターフェースとして幅広い利用と応用が見込める。

＜2008年度の研究の当初計画＞

1．同義語辞書と概念共起を用いた関連キーワード解析

ゲノム情報の理解に必要と考えられる解剖部位，生物学名，病名および症候名，生体分子および医薬品名，方法，現象などの専門用語について，ライフサイエンス辞書とMeSHのすり合わせによりシソーラスツリーおよびシノニム辞書を制作する。PubMed抄録を収集した文献コーパスに対して，シノニム辞書を適用して統制語によるタグ付けを行うPerlスクリプトを開発する。統制語タグが同一抄録中で共起する頻度を解析し，各用語について出現頻度，共起キーワードおよび共起頻度データを得る。得られたデータが専門的に見て妥当な関連性を表すかどうか，研究者による評価を行う。この評価に基づいて，検索キーワードの選別を行い，最適化したデータを得る。

2．関連キーワードを提示する情報検索エンジンの開発

シソーラスと共起解析データをライフサイエンス辞書に実装することによって，日本語および英語のいずれによっても表記のゆれを吸収して統制語による情報検索を可能にするポータルシステムを開発，公開する。このポータルの実用性や有用性については，大学院生，研究者，一般市民の協力による評価と最適化を行う。

3．日本語訳を表示する辞書ツールの開発

ウェッブブラウザで表示されるゲノム情報などの英語ページにおいて，可能な限り簡単な操作で専門用語を辞書引きできるツールを開発する。Mac OS X 10.5においては辞書.appがシステム標準としてありSafariからショートカットで複合語レベルでの辞書検索が実現できるので，これに最適化した専用辞書を制作する。さらに理想的にはシステムやブラウザの種類に依らず，かつ複合語を自動認識するマウスオーバー辞書の制作が望まれる。

＜2008年度の成果＞

1．同義語辞書と概念共起を用いた関連キーワード解析

これまでに，ライフサイエンス辞書に収録された約18万語の英日対訳の専門用語をMeSH DescriptorおよびSupplemental Conceptsに準じた2.5万語の統制語に集約することで対訳シソーラスとシノニム辞書を制作し，すべての統制語について日本語化を完了した。続いて，PubMedより代表的な学術誌に過去10年間に掲載された論文抄録（600 Mバイト）を収集し，シノニム辞書によってテキスト中に最長一致で統制語のタグを施した。次に，同一抄録中で共起する統制語のペアを収集することによって合計143万対の共起頻度を求め，出現した2万語の統制語ごとに上位30対までの共起概念データを得た。このリストを目視によって検討し，曖昧性の排除と統制語の最適化を行い，2万語の統制語に対して約30万語の共起概念を整理した。

2．関連キーワードを提示する情報検索エンジンの開発

このようにして得た共起概念をシソーラスに関連づけることによって，検索された日本語あるいは英語を自動的に統制語に直して表示するとともに関連性の高い共起概念を表示し，外部リンクにより既存ポータルに検索語を渡せるデータを制作した。

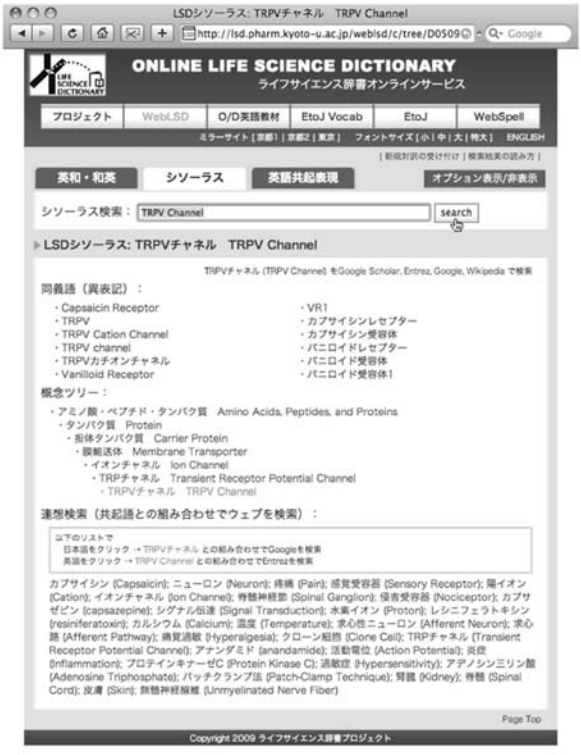

図1　公開したシソーラスと連想検索

これらはオンライン版ライフサイエンス辞書WebLSDのサブセットとして，ある統制語に対して付与される同義語，概念ツリー，共起概念を一覧できるシソーラスとして公開した（図1）。

その結果，従来より公開している英和・和英およびKWIC形式での英語共起表現辞書へ新たにシソーラスが加わり，その間を相互に行き来しながら，専門用語の階層性や同義語を調べることができる我が国はじめてのオンライン専門用語辞書が完成した。また，共起概念を用いる連想検索によって，特定の遺伝子や生体分子に対して関連する現象，疾患，医薬品，手法，並列概念などと併せてEntrezなどの外部データベースを絞り込み検索できるゲノム科学ポータルを構築した。

3．日本語訳を表示する辞書ツールの開発

辞書ツールについては，Mac OS X 10.5 Leopardに付属する辞書アプリケーションが優れた仕様をもっていることが調査の結果わかったので，まずは対訳辞書，シソーラス，連想検索をall-in-oneで含むスタンドアロン辞書の制作を試みた。その結果，WebLSDと同等のコンテンツを含みながら，Mac OS X Leopard標準のWebブラウザSafariやメールソフトMailの中において，ショートカットキーに触れるだけでカーソル位置の英文テキストから複合語も含めた専門用語を認識し，さらにその訳語をポップアップ表示することができる辞書が完成し，2009年1月より公開した（図2）。この辞書においては「詳細」リンクをクリックすることによってシソーラスや連想検索が可能となっており，単に対訳を調べるだけでなく，そのまま用語の解説や関連情報をインターネットから収集することができる。

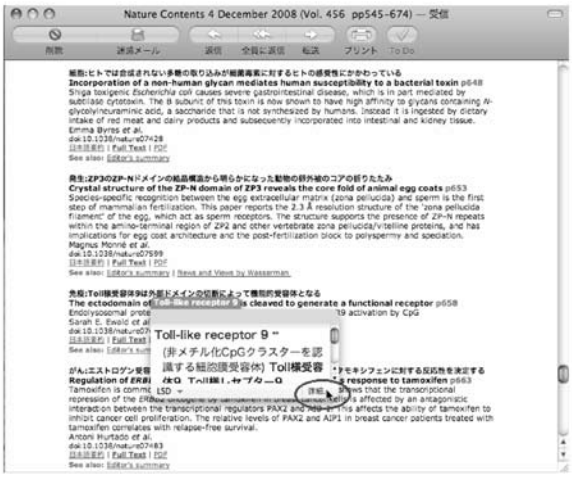

図2　公開したLeopard LSD (Mac OS X 10.5専用)

＜国内外での成果の位置づけ＞

本研究で開発したシソーラス・連想検索ポータルおよびポップアップ専門用語辞書は，我が国で他に例のない大規模なゲノム科学の専門用語辞書であるのみならず，きわめて有用性と拡張性の高いツールに位置づけることができる。Webベースで公開したシソーラス・連想検索ポータルは，誰もが無償で24時間使うことができるように運用されており，ゲノム科学の教育や一般社会への研究成果の還元など，様々な局面において活用していただくことができる。

本研究で解析に用いた英文コーパスはPubMedより収集したため，広領域にわたる専門用語を含み，連想検索で提示される用語も多岐にわたっている。しかしコーパスを変えることによって例えば臨床医学とゲノム医学との関連性を強調したり，環境生物学とゲノム医学との関連を調べるなど，より専門領域に特化したポータルを構築することが容易にできる拡張性にも富んでいる。

また今回，Leopard LSDとして公開したスタンドアロン辞書は，我が国で初めてWebやHTMLメール上での複合語にも対応できるポップアップ形式の対訳辞書であるのみならず，シソーラスや連想検索リンクも内包しており，Mac OS Xに限定されるとはいえ，まったく新しい高機能な教育研究ツールとして活用される可能性を秘めている。

＜達成できなかったこと、予想外の困難、その理由＞

本研究における最大の困難は，ひとえに用語の収集と整理といった地道な辞書の構築作業である。テキスト解析に基づく用語の抽出に関しては英語，日本語とも自然言語処理による半自動的なテータ収集が可能であるが，そうして生まれる何万語もの用語についての対訳やシソーラスの定義に対して有用なツールはあまりなく，最終的には人間の知識と作業が必要となる。

また，シソーラスツリーの構築にあたり，既存のシソーラスに合致しない概念の用語が意外に多いことも作業を困難にしている。本研究で2.5万統制語にアサインできた18万同義語以外にも，同じカテゴリーに属する英語あるいは日本語は1万語以上残されており，それらを体系化する独自の方針や方策を考える必要が生まれている。

＜今後の課題＞

今後の課題として，上述したようにシソーラスの拡張を行いながら体系化されていない専門用語をまとめていく必要がある。本年度は医学・生命科学を網羅する既存のシソーラスとしてMeSHに準拠することでツリー構築を行ったが，次年度は各種の標準病名，生物学名，酵素分類などの体系を参考にしながら，複数軸のシソーラスを実現させて行きたい。情報検索に最適な粒度の統制語の最適化をはかる等，データの妥当性も再検討する必要があり，これらによって公開ポータルとしての有用性を高めたい。

また，辞書ツールについては，多数派を占めるWindowsパソコン内で利活用できる新たなアプリケーションの開発が必要である。これについてはFirefoxにアドオンできるマウスオーバー辞書という形で，支援班の協力を得ながら開発に着手したところであり，次年度には公開できると思われる。もし可能であれば，提示する共起概念の視覚化などについても検討したい。

さらに本研究で構築しているツールは，本当に学生や初学者にとってただ便利なだけでなく，ゲノム情報に対する興味を高め，得られる情報の質を高めることに寄与できるかどうかも関心あるところである。これについては教育実践の場はすでにあるので，良い方策を考えて実践してみたい。

なお最後に，本研究で得られた辞書テキストなどの情報資源については自前サーバ構築やツール開発だけでなく，積極的に外部のデータベース提供者に対しても無償によるデータ提供を行っている。今後，さらに利用していただけるよう広報も行っていきたい。

＜成果公表リスト＞

1. 0901161321
金子周司，鵜川義弘，大武博，河本健，竹内浩昭，竹腰正隆，天野博夫，藤田信之：医学用語シソーラスに基づく効率的医療情報検索システムの開発，医療情報学，28 (Suppl) 639-642 (2008)
2. 0602211137
ライフサイエンス辞書WebLSD
http://lsd.pharm.kyoto-u.ac.jp/