オントロジー化した科学技術用語シソーラスによる生物学的機能、ロールの推論

Reasoning of Biological Functions and Roles using the Refined Japan Science and Technology Thesaurus

櫛田達矢 [1*]　古崎晃司 [2]　増田壮志 [2]　建石由佳 [1]　渡邊勝太郎 [3]　松邑勝治 [3]　川村隆浩 [3]　高木利久 [1,4]

Tatsuya Kushida[1], Kouji Kozaki[2], Takeshi Masuda[2], Yuka Tateisi[1], Katsutaro Watanabe[3], Katsuji Matsumura[3], Takahiro Kawamura[3], and Toshihisa Takagi[1,4]

[1] 国立研究開発法人 科学技術振興機構バイオサイエンスデータベースセンター
[1]National Bioscience Database Center, Japan Science and Technology Agency
[2] 大阪大学産業科学研究所
[2]The Institute of Scientific and Industrial Research, Osaka Univ.
[3] 国立研究開発法人 科学技術振興機構情報企画部
[3]Dept. of Information Planning, Japan Science and Technology Agency
[4] 東京大学大学院理学系研究科
[4]Dept. Biological Sciences, Grad. School of Science, The Univ. of Tokyo

Abstract: We attempted to predict biological functions for gene products and discover disease-related gene products using their concepts and relationships within the refined Japan Science and Technology thesaurus to investigate and evaluate the effectiveness and practicality of such a method. Thus, we inferred that 1600 or more concepts imply 100 or more biological functions, roles, and qualities using the inheritance approaches of the ontological structure, such as “is-a inheritance” and “part-of composition”. Moreover, we demonstrated that we could more efficiently and precisely discover disease-related gene products, such as thromboembolism-related gene products, using the knowledge graph that was constructed from the refined thesaurus than from the graphs constructed from the original thesaurus.

1 科学技術用語シソーラスとは

JST 科学技術用語シソーラス（科学技術用語シソーラス）は、科学技術分野の文献の索引付けに用いられる語彙を階層化したものである。化学工学、電気工学、土木工学、コンピューター科学、材料科学、環境科学、数学、物理学、天文学など各分野の約 24.5 万概念を収録している。そのうち、生命科学分野の概念は約 9 万語（同義語、異表記語除く）を収録し、この分野だけでも約 70 のサブカテゴリーを持ち、米国国立医学図書館 (NLM) が提供する生命科学用語集(MeSH) (https://www.ncbi.nlm.nih.gov/mesh)に匹敵する大規模で網羅な言語資源である。

各概念は上位語 (BT)、下位語 (BT)および関連語(RT)の関係を用いて構造化される。中でも RT は、異なる種類、レベルの概念間を関係づけにも用いられている（例、生命科学カテゴリーの生命現象と遺伝子産物間）。このシソーラスにおける RT の適用範囲は広く、異なる科学技術分野の概念の関係付けにも用いられている（例、生命科学と化学工学）。このように分野が異なる科学技術概念間の直接的な関係を取り扱うオントロジー、シソーラスは他にはほとんどなく、科学技術用語シソーラスの特徴の 1 つである。

2 これまでの研究成果

2.1 科学技術用語シソーラスのオントロジー化

しかし、科学技術用語シソーラスでは概念間の関係が BT、NT、RT の単純な 3 種類しかなく、多様で

* kushida@biosciencedbc.jp

複雑な概念間の関係を厳密に記述するには限界がある。そこで生命科学分野を対象に、複数の生命科学専門家およびオントロジストの共同作業で RT の細分類化と標準化を効率的かつ高精度で行う手法を開発し、複雑な概念間の関係の記述を可能にする科学技術用語シソーラスのオントロジー化を進めた[1]。この作業は、オントロジーエディタ「法造」(http://www.hozo.jp/) を使って行った。このオントロジー化したシソーラスを検索システムやデータベースに組み入れることで、上位下位関係を用いた緩和検索や RT を使った単純な関連検索だけではなく、細分類化した関係を用いた知的探索が可能になることが期待される。

2.2 外部データを使った科学技術用語シソーラスの拡張

オントロジー化した科学技術用語シソーラスではあるが、概念の収集を広範囲にかつ浅く行っているため、複数の領域において概念の収集と整理が不十分であることが分かっている[2]。例えば、生命現象とそれを制御する遺伝子産物の情報が不足している。そこで、これらの情報を Gene Ontology などのオープンデータを使って補うことが適当であると考えた。BioPortal (https://bioportal.bioontology.org/) が提供する Web ツールである Recommender および Annotator を使い、BioPortal に登録されているオントロジーを対象に、概念間のラベルおよびシノニムのマッピングを行い、その結果を生命科学の専門家が確認し、同一概念と判断した場合、両概念を skos:exactMatch を使って対応付け、科学技術用語シソーラスに統合した。

本研究は、オントロジー化および概念を拡張した生命科学分野における科学技術用語シソーラスを用いて、(1) オントロジー階層を使った機能、ロール、性質の推論および、(2) 科学技術用語シソーラスから構築したナレッジグラフを使った新規の疾患関連遺伝子産物の発見に取り組み、科学技術用語シソーラスの有効性と推論へ活用の可能性の検討を行った。

表 1. 科学技術用語シソーラスのオントロジー階層を使った推論結果の例

推論のタイプ	例
is-a 階層を利用した（下位継承による）機能推論	上位概念 “ABC 輸送体”の機能 “生体輸送” を下位概念 “P 糖タンパク質”に継承推論
is-a 階層を利用した（下位継承による）ロール推論	上位概念 “コリン作動薬”のロール “Alzheimer 病治療薬” を下位概念 “ニコチン作動薬”に継承推論
is-a 階層を利用した（下位継承による）性質推論	上位概念 “アルカリ性ホスファターゼ”の性質 “アルカリ性ホスファターゼ活性” を下位概念 “骨型アルカリ性ホスファターゼ”に継承推論
全体部分構造を利用した機能推論	部分構造 “スプライシング因子”の機能 “RNA スプライシング” を全体構造 “スプライセオソーム”に継承推論

3 科学技術用語シソーラスを用いた推論

3.1 オントロジー階層を用いた推論

オントロジー階層を使った生命科学分野における遺伝子産物や細胞構造に対する機能 (function) や制御関係の推論は Gene Ontology の取り組みがよく知られる[3]。今回、科学技術用語シソーラスの関係の細分類化によって、is-a 関係 (subClassOf) および全体-部分関係 (has part) を使った機能、ロール(role) および性質 (quality) の継承による推論が可能になった。例えば、耐性菌の性質である薬剤耐性は、耐性菌の下位概念である MRSA に継承されるなど、これまでに約 1600 件の概念に対して 100 種類以上の機能やロール、性質の推論が可能になった（表 1)。

3.2 ナレッジグラフを用いた推論

科学技術用語シソーラスからナレッジグラフを作成する手法を図 1 に示す。(I) 法造を使い、法造形式の科学技術用語シソーラスを RDF（Turtle 形式）へ変換。(II) RDF をトリプルストアに格納し、SPARQL の実行環境を整備。(III) SPARQL を実行し、概念 “線維素溶解 (Fibrinolysis) ”から 3 ステップ以内で繋がる概念の集合のデータを獲得、このデータをネットワーク可視化ツール Cytoscape (http://www.cytoscape.org/) を使ってグラフとして可視化し、これを “線維素溶解ナレッジグラフ” と呼ぶことにした（図 2)。なお線維素溶解は様々な心血管疾患との関連が指摘され、その患者や医療費が持続的に増加しており大きな社会問題になっている。そのためこれらの情報を適切に収集、整理することは重要と考えられる。

図 2 左上では、遺伝子産物 CLEC2（赤・菱形ノード）が、Platelet Aggregation（血小板凝集）に対して制御の機能 (function) （波線）を持つ一方、血小板凝集の進行の後、引き続いて (precedes) Thromboembolism （血栓塞栓症）が発症することを示している。これは CLEC2 が血小板凝集を通して間接的に血栓塞栓症を制御する可能性があることを意味している。同様に、図 2 中央下の blood coagulation factor （血液凝固因子）（ピンク・菱形ノード）は、血小板凝集を通して間接的に、血栓塞栓症を制御する可能性があることを意味する。また、図 2 右真ん中にある酵素 nattokinase（ナットキナーゼ）や plasmin（プラスミン）（緑・菱形ノード）は、Fibrinolysis（線維素溶解）を通して間接的に Fibrinolytic purpura（線維素溶解紫斑病）や fibrinolysis increased（線溶亢進）を制御する可能性があること意味している。すなわち、概念（ノード）間を直接繋ぐ関係（エッジ）がない場合でも、間接的に繋がれる関係の意味を解釈することで、その概念間の関係を推論することが可能になる。

一方、図 2 左中央では遺伝子 PRKCH gene （ピンク菱形ノード）が cerebral infarction（脳卒中）を制御する機能を持つことがわかるが、脳卒中は血栓塞栓症と RT の関係にあり、これは両者に関係はあるが、詳細な関わり方が不明であることを意味する。そのため、PRKCH gene は脳卒中を制御するが、その先で RT を介して繋がる血栓塞栓症にどのような影響を及ぼすかはよくわからないことを示している。

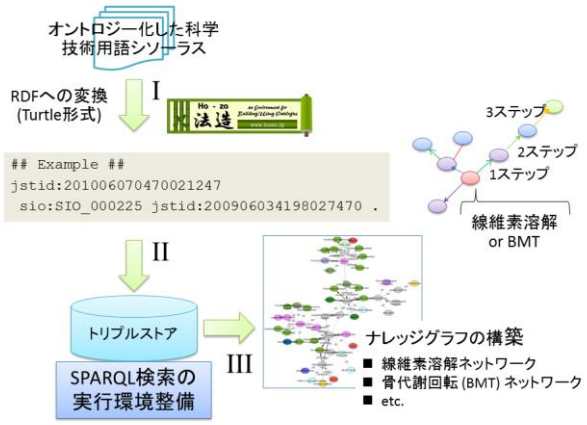

図 1. 科学技術用語シソーラスからナレッジグラフを構築するプロセス

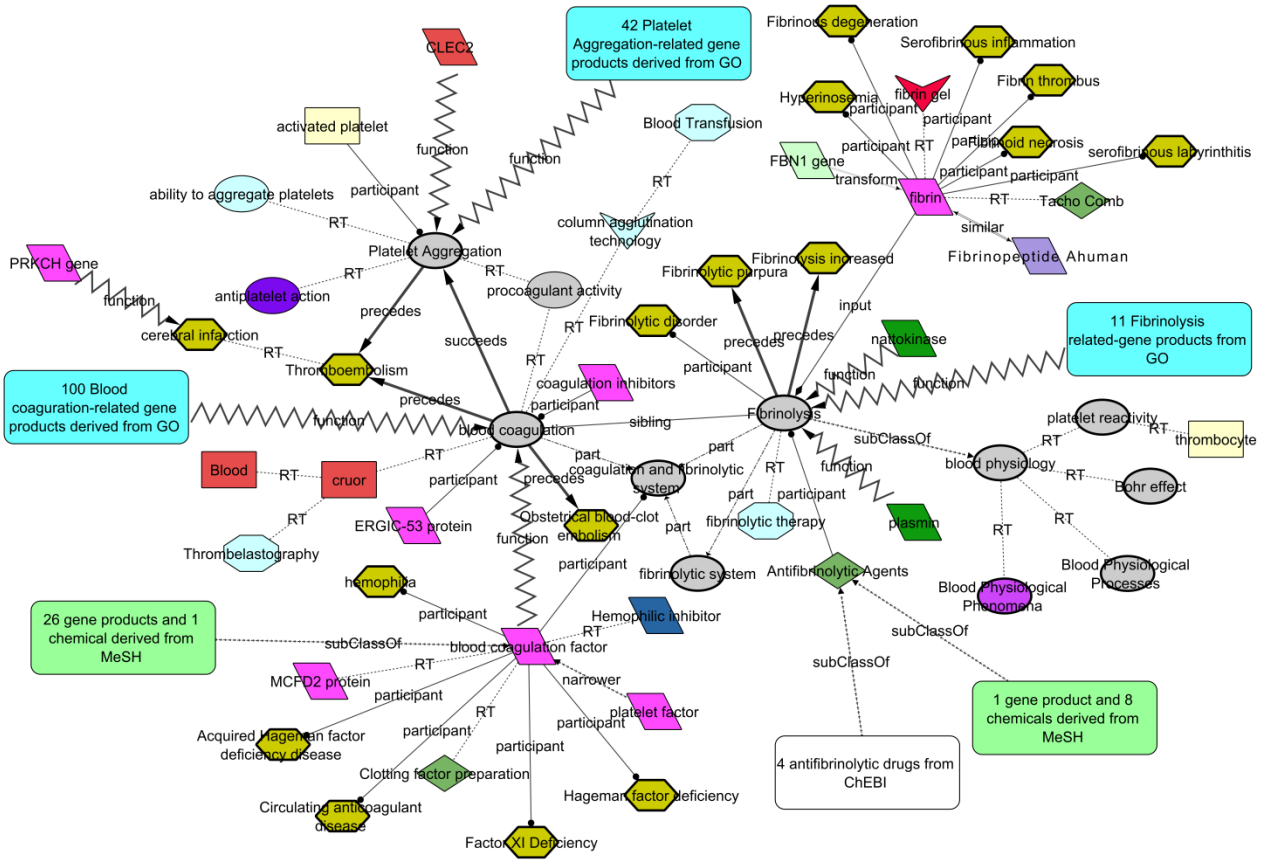

図 2. Gene Ontology, MeSH, ChEBI を使って拡張した線維素溶解ナレッジグラフ
図中では skos:exactMatch のエッジを省略している。

すなわち、ある物質（遺伝子産物、化合物）と疾患間の到達経路中に RT を含んでいる場合、両者には関係がない可能性があることを示している。

以上、科学技術用語シソーラスの RT を細分類化（オントロジー化）することで、RT のままでは判断することができなかった分子、生命現象、疾患間の詳細な関係が明らかになることが示された。なお、このような疾患と遺伝子産物や化合物の関係を収集、整理することは、疾患の治療法や創薬の開発や提案に有用かつ重要な知見を与えると考えられる。

3.3 外部データによるナレッジグラフの拡張と推論

前述のように科学技術用語シソーラスの概念は、科学技術領域を幅広く対象とする一方、各分野について深く掘り下げた情報の収集、整理はできていない。そこで、科学技術用語シソーラスの生命科学分野で情報の収集整理が十分にできていない生命現象と遺伝子産物および化学物質の関係情報を Gene Ontology, MeSH, ChEBI (https://www.ebi.ac.uk/chebi/) を使って補うことを試みた。

Gene Ontology は生物学的プロセス、分子機能、細胞局在の各概念に対し、それらを制御する遺伝子産物の情報を整理している。MeSH は特定の機能や構造上の特徴を有する生体分子の抽象概念（例、血液凝固因子）とその構成メンバーの情報を整理している。ChEBI は化合物の生物学的ロール、化学的ロール、アプリケーションの概念を構造化したオントロジーである。これらが提供する遺伝子産物および化合物と生物学的概念の情報を、前述のラベルおよびシノニムのマッピングツールを使い、ナレッジグラフの構成要素（概念）とマッチさせ、これらのグラフへの取り込みを行った。

線維素溶解ナレッジグラフに追加された遺伝子産物および化合物の数は、Gene Ontology 由来の遺伝子産物が、Fibrinolysis（線維素溶解）に対して 11 件、blood coagulation（血液凝固）に対して 100 件、Platelet Aggregation（血小板凝集）に対して 42 件であった（図 2、水色・長方形ノード）。これら遺伝子産物と生命現象の関係は “has function”を使った。

また、MeSH 由来の遺伝子産物が blood coagulation factor（血液凝固因子）に対して 26 件、化合物が 1 件、Antifibrinolytic Agents（抗線溶薬）に対して 1 件、化合物が 8 件（図 2、薄緑・長方形ノード）、ChEBI 由来の化合物の 4 件がそれぞれナレッジグラフに追加された（図 2、白・長方形ノード）。これら遺伝子産物および化合物と血液凝固因子、抗線溶薬は、“subClassOf”の関係を使って繋いだ。

拡張された遺伝子産物、化合物の情報は、前述の疾患関連遺伝子として、ナレッジグラフを使った推論に活用されることが期待される。例えば、情報拡張の前は、血栓塞栓症の疾患関連遺伝子として推論されるは、CLEC2 (UniProtKB:Q9P126) など 6 件だけであったが、拡張後は、PLAU (UniProtKB:P00749) など 107 件の遺伝子産物がその候補遺伝子として推論された。なお、本研究でナレッジグラフから発見した疾患関連遺伝子、例えば、血栓塞栓症関連遺伝子のすべてが実際に疾患と関係がある遺伝子として特定されるわけではない。あくまでも “候補遺伝子”という位置づけである。ただし、数万ある全遺伝子から少数の候補遺伝子を抽出しその情報を活用することは、効率的な生命科学研究の実施に貢献するなどその利点は大きいと考えられる。

4 結論

オリジナルの科学技術用語シソーラスの RT を細分類、標準化した、すなわちオントロジー化したシソーラスのオントロジー構造を活用することで、生物学的概念に対する機能、ロール、性質の推論が可能になった。これまでに 1600 件以上の概念に対して、上位概念から下位概念へ機能、ロール、性質の継承（is-a 継承）、部分構造から全体構造への機能の継承が行われた。

他方、オントロジー化したシソーラスから構築したナレッジグラフを使い、疾患関連遺伝子および関連化合物の推論の可能性を検討した。線維素溶解ナレッジグラフを用いた検証では、シソーラスのオントロジー化によって、「ある疾患に対して、それに先行して生起する生命現象があり」かつ「その生命現象を制御する機能を有する遺伝子産物がある」などの記述が可能になる。さらにそれらから「その遺伝子産物が、生命現象を通して疾患を制御する」関係が導き出され、その遺伝子が疾患関連遺伝子になると考えられた。

Gene Ontology、MeSH、ChEBI が提供する生物学的概念と遺伝子産物および化合物の関係情報を、線維素溶解ナレッジグラフに取り込んだ例では、100 件以上の血栓塞栓症関連遺伝子および関連化合物が発見された。

5 今後の予定

本研究の構築したナレッジグラフの有効性を厳密に評価するためには、抽出された情報の精度や網羅性を示す必要があるが、評価用の適当な正解データセットが見当たらない。そこで、公的データベースの

データセットやテキストマイニングよって収集された疾患とその関連遺伝子の情報を提供するデータベース DisGeNET [4]を使って両者の比較を行った。本研究で構築したナレッジグラフで抽出した 107 件の血栓塞栓症関連遺伝子のリストと、DisGeNET が提供する 85 件の同疾患関連遺伝子のリストを比較したところ、両者で共通の遺伝子は 14 件のみであった。DisGeNET が提供する情報がすべて正しくまた網羅的である保証はないが、様々なシステムが提供する疾患関連遺伝子の情報およびそれらを比較した結果を公開することは、生命科学研究者をはじめ様々な分野の研究者、開発者に有益であると考える。

他方、RDF や生物学的なナレッジグラフに対して、機械学習によって疾患候補遺伝子, タンパク質-タンパク質相互作用、ドラックターゲットを予測、推論する取り組みが行われ、その成果や課題、可能性が活発に議論されている[5, 6]。今後、オントロジー化した科学技術シソーラスに対して機械学習的なアプローチを採ることを検討している。また現在 Turtle 形式で提供しているシソーラスデータを owl 形式に変換し、DL 推論の実行環境を整備する予定である。

6 データの公開

本研究で使用したオントロジー化した科学技術用語シソーラスの SPARQL エンドポイント (http://lod.hozo.jp/repositories/JstNbcdOnt) を 、 CC BY-NC のライセンスで公開する。

謝辞

本研究の一部は、科研費（17H01789：基盤研究(B)）の支援を受けて実施した。

参考文献

[1] Kushida T., K. Kozaki K., Tateisi Y., Watanabe K., Masuda T., Matsumura K., Kawamura T., and Takagi T.: Efficient construction of a new ontology for life sciences by subclassifying related terms in the Japan Science and Technology Agency thesaurus, Proceedings of 8th International Conference on Biomedical Ontology (ICBO 2017) , (2017) (in print)

[2] Kushida T., Y. Tateisi Y., Masuda T., Watanabe K., Matsumura K., Kawamura T., Kozaki K., and Takagi T.: Refined JST Thesaurus Extended with Data from other Open Life Science Data Sources, Joint International Semantic Technology Conference JIST 2017, LNCS 10675, pp. 35-48, (2017)

[3] http://geneontology.org/page/ontology-relations

[4] Piñero J., Queralt-Rosinach N., Bravo À., Deu-Pons J., Bauer-Mehren A., Baron M., Sanz F., and Furlong L.I.: DisGeNET: a discovery platform for the dynamical ex-ploration of human diseases and their genes. Database 2015, pp.1-17, (2015)

[5] Alshahrani M, Khan MA, Maddouri O, Kinjo AR, Queralt-Rosinach N, Hoehndorf R.: Neuro-symbolic representation learning on biological knowledge graphs. Bioinformatics. 33(17), pp.2723-2730, (2017)

[6] 片山俊明、川島秀一: 生命医科学 RDF データの機械学習・人工知能への応用、2017 年度 人工知能学会全国大会 (第 31 回), 4H1-OS-27a-2 , (2017)