Drug repositioning を志向した異種生物学データベースの統合・比較に関する基礎的検討

Basic study on the comparison and integration of heterogeneous biological database for drug repositioning

田中　達也　今井　健

Tatsuya Tanaka[1], and Takeshi Imai[1]

[1] 東京大学大学院 医学系研究科
[1]Graduate School of Medicine, The University of Tokyo

Abstract: There are many existing databases in biology, and it is important to utilize these databases to predict new medical, pharmaceutical and biological events. Although many attempts have been made to integrate databases to make new predictions, they have focused on developing highly accurate machine learning methods, and few studies have been taken the characteristics and biological mechanisms into consideration when constructing databases. In particular, many of the existing studies equally treat data low to high likelihood of actual occurrence in database, which may well affect the prediction results. Furthermore, they often include data on indirect associations, which we believe it difficult to make accurate predictions based on biological mechanisms. In this paper, I propose a database that takes biological mechanisms into account by comparing the features of existing databases, examining the optimal database structure and reorganizing them. I focus on the research for drug repositioning candidates that show anti-malignant tumor effect and constructed a database with a graph structure that includes 2,675 drugs, 18,880 proteins, and 28,873 edges. As a result, it is confirmed that this database contains 69 drugs that can show anti-malignant tumor effect.

1. はじめに

近年、様々な分野でこれまで構築・蓄積されてきたデータベースを再利用し、活用する動きが注目を集めており、実際に医学・薬学・生物学の分野においても化合物の化学構造データベースや体内で分子間相互作用データベース、疾患の遺伝子関連データベースなどを活用した研究が多く行われている。特に最近ではより多くの情報を加味した上での推論を行うため、複数のデータベースを統合して利活用する研究が複数報告されている[1]。また国内でもLinked Open Data(LOD)チャレンジ [2] のように、複数の異なるデータベースソースを RDF 等の形式で表現し、これを統合的に利用することで別の価値を創造するような試みも多く行われている。

最近、医学・薬学・生物学領域では、「既存の薬剤を既知の適応症以外の適応症に用いる」いわゆるDrug repositioning が注目を集めている。これも Drug repositioning 候補を効率的に導出する方法の１つとして、関連する既存の医学・生物・薬学分野のデータベースを統合し、横断的に探索する方法が有効であると考えられる。しかし、このようなデータベース統合のためにはいくつかの課題が存在する。

第１に、探索による未知の Drug repositioning 候補の導出のためには、薬剤の作用機序を表現したデータベースが必要となる。既存の複数のデータベースからこの作用機序に相当する情報を抽出し、統合する必要がある。

第２に、各データベースはそれぞれ異なる目的をもって作成されている為、これらの統合に際しては、対象とする分野、データ数と粒度、統合のための ID を考慮しなければならない。“対象とする分野”は、データベースごとに特化しており、同じ薬剤のデータベースであっても構造やその化合物の性質のみに注目したものもあれば、標的タンパク質のみに焦点をあてたデータベースや臨床的な適用例に注目したものまで様々なものがある。 “データ数と粒度”もデータベース統合の際には重要で、薬剤の作用機序

を探索する上で必要に応じた適度な粒度を選定する必要がある。また“ID”も、複数のデータベースを統合する際に重要な役目を果たす。例えば、薬剤を含むデータベースを扱う場合、薬剤の ID には PubChem Compound ID や ChEBI ID, PharmGKB ID, Drug Bank ID など多くの ID が存在するが、これらを統一して同一のものと認識する為には、最も網羅的に適切な粒度で作成したデータベースの ID でマッピングされていることが求められる。

これまでも医学・薬学・生物学分野においては多くのデータベースが構築されてきたが、Drug repositioning に２次的に活用する観点から、データベース統合に関する知見を整理した先行研究はあまり存在しない。そのため、本研究では、データベース統合に際し重要と考えられる項目に注目し、既存のデータベースを比較・検討することで、抗悪性腫瘍効果を示す未知の Drug repositioning 候補予測に用いることの出来るデータベースを構築する上での基礎的な検討を行うことを目的とする。具体的には、チロンシキナーゼ阻害による抗腫瘍効果を示す未知の Drug repositioning 候補予測のためのデータベースを構築し、実際の既存の薬剤の含有率を確認した結果を考察した上で、今後の展望について議論を行う。

2. Drug Repositioning 分野の先行研究

新規の薬剤の探索を行うことは非常に長い時間やコストがかかることが知られているが、既存の薬剤を既知の適応症以外の適応症に用いる Drug repositioning は時間やコストを大幅に短縮する解決策の１つとなり得る。近年では既存データベースを活用した Drug repositioning の事例が多く報告されており、薬剤がどの疾患、標的タンパク質に関連しているのかを予測する研究がなされている。

多くの既存研究では、複数のデータベースを統合してグラフ構造のデータベースとして扱い、そこから新たな予測を行うことを目的としており、最近報告された研究では STITCH database, The Comparative Toxicogenomics Database を統合してグラフ構造のデータベースを作成し、Graph Convolutional Neural Network やその他の手法により薬剤が標的とし得るタンパク質を探索することにより先行研究と比較して優れた結果を示している。[3]他にも同様の研究はなされているが、その多くがグラフを探索する手法や機械学習の手法に焦点を当てて精度の良い予測を行うことを目的としており、データベースの構成に関してはあまり考慮がなされていない研究が多くみられる。

既存研究におけるデータベースの構成として薬剤の作用機序を意識していない場合が多く、例えば既存研究の多くは疾患関連遺伝子のデータベースを用いているが遺伝子が疾患の直接的な原因ではない場合が多い為、適切な推論を行うことを妨げる要因になり得る。

そこで本論文では適切な組み合わせの既存のデータベースを統合することで、薬効を示し得る薬剤作用機序の経路を表現することに注目した。

3. 方法・結果

3.1 方法概要

今回薬剤の作用機序を表すデータベースの構築を考えているため、（ｉ）「薬剤がどの標的タンパク質と結合して作用を及ぼすのか」（ii）「標的タンパク質からどの生物学的経路をたどって疾患に影響を及ぼすタンパク質に到達するのか」（iii）「どのタンパク質が疾患に関与しているのか」の３点に注目したデータベースがそれぞれ必要である。

化学構造、標的タンパク質をまとめた代表的なデータベースとして Drug Bank database[4], PubChem database[5], PharmGKB database[6], Binding DB[7], Therapeutic Target database[8], STITCH database[9], ChEBI database[10], KEGG Drug database[10]の８種類が挙げられ、生体内でのタンパク質相互作用をまとめた代表的なデータベースとして KEGG pathway database[11], Reactome database[12], BioCyc database[13], Gene Ontology[14], BioGRID database[15], STRING database[16]の６種類が挙げられる。

多くの Drug repositioning に関する既存研究では薬剤作用機序の考慮がなされていないが、本論文では可能な限り正確な薬剤作用機序を考慮した上でデータベースの構成を検討した。

3.2 使用 ID の選定

今回は主に薬剤、タンパク質、疾患についての複数のデータベースを統合した上で構築を行う為、はじめにそれぞれ何の ID を用いるかを統一する必要がある。

まず薬剤に用いる ID を決めるために以下にそれぞれの薬剤データベースに含まれる総薬剤 ID 数を示す。

表 1: 各薬剤データベースの総薬剤 ID 数比較

データベース数	各データベース内の総薬剤 ID 数
Drug Bank database	14,315
PubChem database	109,908,766
PharmGKB database	3,449
BindingDB	5,726
Therapeutic Target Database	236,469
STITCH database	389,393
ChEBI database	59040
KEGG drug database	11,576

薬剤ではデータベースごとに様々な分類方法が異なる為、ID 数においても大きな違いが見られ、主な違いは化合物ごとに異なる ID で区別しているか､同じ化合物の中でも水和物か無水物かで詳細な情報も含めて区別しているかである。今回は薬剤の主成分の化合物としての役割を区別したい為、水和物か無水物かなどの詳細情報は割愛することとし、その上で可能な限り網羅的に化合物ごとに ID を付与している Drug Bank, ChEBI の ID が適していると考えられた。さらにその中でも他の多くのデータベースでも用いられている ChEBI の ID 体系が化合物を統一する際に有用であると考えられた。

次にタンパク質に使用する ID の検討を行う為に薬剤データベースに含まれる標的タンパク質に付与されている ID を以下に示す。

表 2: 標的タンパク質マッピング ID 比較

データベース名	標的タンパク質のマッピング ID
Drug Bank	Uniprot ID, NCBI Gene ID
PubChem	Uniprot ID, NCBI Gene ID
Binding DB	Uniprot ID, NCBI Gene ID
Therapeutic Target database	Uniprot ID, NCBI Gene ID
STITCH database	Uniprot ID, NCBI Gene ID
KEGG Drug	Uniprot ID, NCBI Gene ID

タンパク質では今回示した全てのデータベースでタンパク質データベースの Uniprot ID または、遺伝子データベースの NCBI Gene ID を用いているが､今回は主にタンパク質に注目する為、タンパク質データベースである Uniprot ID を用いることとした。

疾患については後にも述べるが、疾患の主なデータベースの ID である ICD コードや OMIM ID にマッピングされた信頼性の高い薬剤-適応症のデータベースが存在しないことから、解剖学治療化学分類法に基づき薬効をまとめた ATC コードの薬効を用いることとした。

3.3 データベースの選定

(ⅰ)「薬剤がどの標的タンパク質と結合して作用を及ぼすのか」を表すのに用いるデータベースとして、どのデータベースを用いるのが最適かを検討するために、まずは薬剤に関連するデータベースである Drug Bank, PubChem, PharmGKB, BindingDB, Therapeutic Target database, STITCH database, ChEBI, KEGG drug の８種類のデータベースにおいて、今回必要としている化学構造データ・構造類似性・標的タンパク質データ・標的タンパク質との結合強度データの 4 つに関する有無を比較した結果、以下のようになった。(表中の✔は当該データを含むもの、―は含まないものを示す。)

表 3: 各データベースの内容比較

データベース名	化学構造	構造類似性	標的タンパク質	標的タンパク質との結合強度
Drug Bank databse	✓	–	✓	–
PubChem database	✓	–	✓	✓
PharmGKB databse	✓	–	–	–
Binding DB	–	–	✓	✓
Therapeutic Target database	✓	–	✓	✓
STITCH database	✓	✓	✓	✓
ChEBI database	✓	–	–	–
KEGG Drug databse	✓	–	✓	–

この表から、まず必須である標的タンパク質データを含むデータベースは Drug Bank, PubChem, Binding DB, Therapeutic Target database, KEGG Drug であることがわかった。次にこの５種類に含まれる標的タンパク質のデータの内、承認薬のデータ数の比較を行った結果を以下に示す。

表 4: 標的タンパク質データ数比較

データベース名	承認薬に対する標的タンパク質データ数
Drug Bank	2,170
PubChem	207
Binding DB	1020
Therapeutic Target database	730
KEGG Drug	4,975

Drug Bank と KEGG Drug がより多くの薬剤-標的データベースを含んでおり包括的なデータベースだと考えられるが分類方法に注目すると、Drug Bank は主成分の化合物ごとに ID が決められているのに対して KEGG Drug は同じ化合物であっても名称が異

なるものに対しては区別して ID が決められている。今回は薬剤の主成分となる化合物に注目したいので、ChEBI ID へのマッピングも可能である点も踏まえて Drug Bank を用いるのが良いと考えられた。

次に、(ⅱ)「標的タンパク質からどの生物学的経路をたどって疾患に影響を及ぼすタンパク質に到達するのか」を表すタンパク質相互作用データベースについて検討を行う。代表的なデータベースとして KEGG pathway, Reactome, Biocyc, Gene Ontology , NCI/Nature PID, BioGRID, STRING database などが挙げられる。それぞれ代謝 Pathway に注目したものや疾患 pathway に注目したものなどがあるが、今回の研究では未知の薬剤作用機序予測を行うことを最終的な目標として考えており、同程度の信頼性・粒度であればデータ数が多い方が適切である。そこで、KEGG pathway, Reactome, Biocyc, Gene Ontology , NCI/Nature PID, BioGRID の全てのデータを含んだ STRING database が最適であると考えられた。

最後に(ⅲ)「どのタンパク質がどの疾患に関与しているのか」を表現するデータベースとして多くの既存研究では OMIM database などの疾患関連遺伝子をまとめたデータベースを使用しているが、遺伝子から疾患に至るまでを直接結びつけるには作用機序として間接的である。そこで、本来は「遺伝子がコードするタンパク質が細胞レベルでどのような役割を果たし、その細胞レベルでの Phenotype がどの疾患に繋がるのか」を表すことが理想である。すなわち、“タンパク質”と“細胞レベルでの Phenotype”と“疾患”を結びつけるデータベースを構成するために、「“タンパク質”-“細胞レベルでの Phenotype”」及び「“細胞レベルでの Phenotype”-“疾患”」の組み合わせが必要である。「“タンパク質”-“細胞レベルでの Phenotype”」は遺伝子の役割ごとにアノテーションが行われている Gene Ontology が最適であるが「“細胞レベルでの Phenotype”-“疾患”」に関するデータベースは私の知る限り存在しない。その為、手動ではあるが Gene Ontology の“細胞レベルでの Phenotype”の項目に相当する項目を薬理学書グッドマン・ギルマンよりデータを抽出することとし、今回は“Cell Cycle”, “Cell Death”, “Cell Population Proliferation”が「抗悪性腫瘍効果」につながり得る“細胞レベルでの Phenotype”であると考えた。

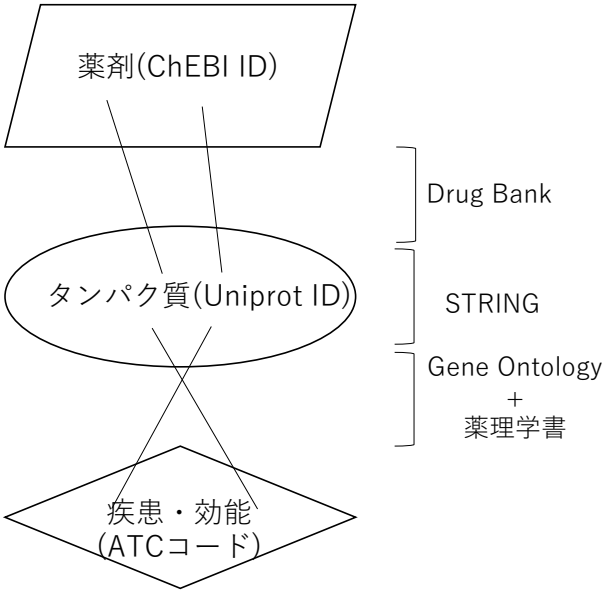
3.4 評価に用いるデータベース

将来的にデータベースを用いて薬効の予測を行うことを考えている為、評価を行う際に薬剤-適応症の組み合わせに関するゴールドスタンダードデータベースが必要である。しかし薬剤添付文書などの文章形式での薬剤-適応症データは存在するが適応症を何らかの ID にマッピングしてまとめた信頼性の高いデータベースは存在しない。その為、適応症ではなく薬効を体系的にまとめた ATC コードを利用することで、薬剤-ATC コードのデータを評価の際のゴールドスタンダードのデータとした。

3.5 最終的なデータ構成

以上をまとめると、薬剤-タンパク質間は“Drug Bank”を用い、タンパク質間相互作用は“STRING database”を用い、タンパク質-薬効間は“Gene Ontology”と“薬理学による知識”を利用して「抗悪性腫瘍効果」に対して効果を示す薬剤作用機序データベースを構築した。

図 1: 構築したデータベースの概略図



その結果、承認薬剤データは 2,675 種類、タンパク質は 18,880 種類、薬剤-タンパク質相互作用データは 9,993 種類、タンパク質間相互作用データは 18,880 種類、タンパク質-効能データは 193 種類となった。実際にチロシンキナーゼ阻害作用により「抗悪性腫瘍効果」を示し ATC コードで確認される薬剤は 70 種類存在し、その中の 69 種類が実際に今回のデータベースで「抗悪性腫瘍効果」を説明することが出来た。具体例として Imatinib ではチロシンキナーゼを標的タンパク質として PI3K/Akt の経路を介して抗悪性腫瘍効果を示し得ることが確認され、これは実際の Imatinib の作用機序としても考えられる経路である。

5. 考察

今回の結果は、実際にチロシンキナーゼ阻害により「抗悪性腫瘍効果」を示す 70 種類の薬剤の内 69 種類が今回構築したデータベースにより「抗悪性腫瘍効果」までの経路を説明できることを示している。説明することが出来なかった Dacomitinib という 1 種類の薬剤に関して薬剤-標的タンパク質のデータが含まれていなかった為であると考えている。しかしながら、このような薬剤-標的タンパク質のデータが存在しない薬剤に関しても、薬剤構造類似性を考慮することで探索対象に含めることが可能であると考えられ、今後の課題である。

また、既存の薬剤の内 69 種類が「抗悪性腫瘍効果」まで経路を結びつけることが出来たがそれぞれの薬剤において複数の作用機序が考えられ、どの経路が最も適切なのかを判断することは出来ない。その為実際に起こり得る経路を優先順位付けする必要があり、その方法として既存研究の多くはデータベースのグラフ構造自体の特徴を考慮した方法が多く報告されている。しかしこの方法では実際のデータの因果関係の強さを利用しておらず生物学的な情報を十分加味出来ていないと考えられる。そこで実データを利用した上で因果を推定することで、生物学的なデータに基づいたより正確な予測を行うことができるのではないかと考えており、今後はデータベースを構成する個々のオブジェクト間の因果強度を推定する手法の開発に進めていく予定である。

その他にも臨床現場で Drug repositioning の可能性が示唆されている薬剤について今回のデータベースや臨床データを用いて適切な説明・検証が出来るのかを検討してく予定である。

参考文献

[1] Anastasis Oulas, George Minadakis, Margarita Zachariou et al: “Systems Bioinformatics: increasing precision of computational diagnostics and therapeutics through network-based approaches”, Briefings in Bioinformatics, Vol. 20, No. 3, pp. 806-824, (2019)

[2] Linked Open Data チャレンジ Japan https://2020.lodc.jp/

[3] Wei Wang, Xi Yang, Chengkun Wu et al: “GCInet:graph convolutional network-based chemical-gene interaction in an integrated multi-relational graph”, BMC Bioinformatics, Vol. 21, pp. 544, (2020)

[4] Wishart DS, Knox C, Guo AC et al: “Drugbank: a comprehensive resource for in silico drug discovery and exploration”, Nucleic Acids Res, Vol. 34, No.1, ppD668-D672

[5] Kim S, Chen J, Cheng T, et al, “PubChem in 2021: new data content and improved web interfaces”, *Nucleic Acids Res*., Vol. 49, ppD1388-D1395, (2021)

[6] M. Whirl-Carrillo, E.M. McDonagh, J. M. Hebert et al, “Pharmacogenomics Knowledge for Personalized Medicine”, Clinical Pharmacology and Therapeutics, Vol. 92, pp414-417, (2012)

[7] ilson,M.K., Liu,T., Baitaluk,M. et al, “BindingDB in 2015: A public database for medicinal chemistry, computational chemistry and systems pharmacology”, Nucleic Acids Research, Vol. 44, ppD1045-D1063, (2016)

[8] Y.X. Wang, S. Zhang, F. C. Li et al, “Therapeutic target database 2020: enriched resource for facilitating research and early development of targeted therapeutics”, Nucleic Acids Research, Vol. 48, ppD1031-D1041, (2020)

[9] Szklarczyk D, Santos A, von Mering C et al, “STITCH 5: augumenting protein-chemical interaction networks with tissue and affinity data”, Nucleic Acids Research, ppD380-D384, (2016)

[10] Hastings J, Owen G, Dekker A et al, “ChEBI in 2016: Improved services and an expanding collection of metabolites”, Nucleic Acids Research, pp D1214-D1219, (2015)

[11] Kanehisa, Furumichi M., Tanabe M et al, “KEGG: new perspectives on genomes, pathways, diseases and drugs”, Nucleic Acids Research, pp D353-D361, (2017)

[12] Jassal B, Matthews L, Viteri G et al, “The reactome pathway knowledgebase”, Nucleic Acids Research, pp.D498-D503, (2020)

[13] Peter D Karp, Richard Billington, Ron Caspi et al, “The Biocyc collection of microbial genomes and metabolic pathways”, Briefings in Bioinformatincs, Vol. 20, pp.1085-1093, (2017)

[14] Gene Ontology Consortium, “The Gene Ontology resource: enriching a GOld mine”, Nucleic Acids Research, Vol. 49, pp.325-334, (2020)

[15] Chris Stark, Bobby-Joe Breitkreutz, Teresa Reguly et al, “BioGRID: general repository for interaction datasets”, Nucleic Acids Research , Vol. 34, pp.535-539, (2006)

[16] Szklarczyk D, Gable AL, Lyon D et al, “STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets”, Nucleic Acids Research, Vol. 47, pp.D607-613, (2018)