ライフサイエンスデータベース統合推進事業(統合化推進プログラム)
研究開発実施報告書 様式

2023 年度 研究開発実施報告

概要

研究開発課題名	ヒトゲノム・病原体ゲノムと疾患・医薬品をつなぐ統合データベース
開発対象データベースの名称(URL)	KEGG MEDICUS (https://www.kegg.jp/kegg/medicus/)
研究代表者氏名	金久 實 (70183275)
所属・役職	京都大学化学研究所 特任教授 (2024 年 3 月時点)

目次

§1. 研究実施体制

グループ名	研究代表者または主たる共同研究者氏名	所属機関・役職名	研究題目
研究代表者グループ	金久　實	京都大学・特任教授	ヒトゲノム・病原体ゲノムと疾患・医薬品をつなぐ統合データベース

§2. 研究開発対象とするデータベース・ツール等

(1) データベース一覧

【主なデータベース】

No.	名称	別称・略称	URL
1	KEGG MEDICUS		https://www.kegg.jp/kegg/medicus/

【その他のデータベース】

No.	名称	別称・略称	URL
1	Virus-Host DB		https://www.genome.jp/virushostdb/

(2) ツール等一覧

No.	名称	別称・略称	URL
1	医薬品相互作用チェック		https://www.kegg.jp/medicus-bin/ddi_manager

§3. 実施内容

(1) 本年度の研究開発計画と達成目標

本研究開発では、ヒトゲノムおよびウイルスその他の病原体ゲノムの情報を社会で活用するための基盤データベースとして、ネットワーク情報、疾患情報、医薬品情報を統合した KEGG MEDICUS の機能拡張と高品質化を行う。とくにネットワークデータベースの拡張により、数多くの疾患と分子間相互作用・反応ネットワークのゆらぎとの関連づけを行う。これにより KEGG MEDICUS を医薬品情報だけでなく疾患情報としても国際的に最高品質のデータベースとし、ゲノムと疾患・医薬品をつなぐ実用的価値のあるデータベースとする。

研究開発項目としては、「ネットワークデータベース」、「疾患データベース」、「医薬品データベース」、「解析ツール」の 4 つを設定し、ネットワークデータベースの中でもとくに「ウイルスタンパク質のオーソロググループ（当初計画にあるオーソログクラスターの名称をオーソロググループへ変更したが、内容的には同じである）」を主要なサブ項目としている。今年度も継続してヒト遺伝子バリアントや病原体遺伝子がシグナル伝達・代謝といった分子間相互作用・反応ネットワークにどのようなゆらぎを与え、どのような疾患と関連しているかについての知識を集約した NETWORK データベースを拡張する。これは同時に DISEASE データベースに疾患と分子ネットワークの関連を付与するものであり、DISEASE データベース高品質化の一環でもある。また解析ツールの研究開発項目では、ウイルスオーソロググループのデータを拡張し、ウイルスと真核生物および原核生物の間でどのような遺伝子や遺伝子クラスター（ゲノム上で保存された類似遺伝子の並び）がやりとりされているかといった解析の準備を行う。

(2) 進捗状況

① ネットワークデータベース

今年度の進捗状況は、表 1 に示した KEGG MEDICUS データ数の推移で 2023/4/1 から 2024/4/1 の部分に反映している。NETWORK データベースはネットワーク要素（N 番号エントリ）とネットワークバリエーションマップ（nt 番号エントリ）から構成される。ネットワーク要素はシグナル伝達や代謝などに関与する一次元的な分子のつながりで、レファレンスとなる通常のネットワークとゆらいだネットワークがあり、後者はさらにヒト遺伝子バリアント、病原体遺伝子、環境因子などゆらぎの種類で区別される。ネットワークバリエーションマップは通常のネットワークの下にゆらいだネットワークをアライメント表示したもので、どのようなゆらぎがどのような疾患に関連しているかが示されている。また通常のネットワーク要素は PATHWAY データベースと関連づけられ、パスウェイマップ上でその位置をハイライト表示できるようにしている。

表 1. KEGG MEDICUS（https://www.kegg.jp/kegg/medicus）データ数の推移

	2019/4/1	2020/4/1	2021/4/1	2022/4/1	2023/4/1	2024/4/1
KEGG NETWORK (N)	690	1,011	1,312	1,408	1,349+702	1,396+1,099
(nt)	88	114	128	133	151	140
KEGG VARIANT	245	415	441	456	928	1,328
KEGG DISEASE	2,298	2,402	2,498	2,551	2,627	2,701
(nt linked)		143	174	175	417	739
KEGG DRUG	10,955	11,240	11,448	11,873	12,101	12,368
KEGG DGROUP	2,206	2,274	2,318	2,384	2,426	2,462

(N) – Network element
(nt) – Network variation map

前期の統合化推進プログラムでは、NETWORK データベースは疾患パスウェイマップに詳細情報を付与する観点で開発が行われ、がん、神経変性疾患、ウイルス感染症といった特定の疾患が主な対象であった。今期の統合化推進プログラムでは、DISEASE データベースの様々な疾患エントリに分子ネットワーク情報を付与する観点で開発を行っており、代謝やシグナル伝達などのパスウェイごとに関連する疾患を蓄積している。KEGG MEDICUS のトップページから KEGG NETWORK のリンクをたどると、ネットワークバリエーションマップの一覧が表示されるが、右側の Disease view は前期開発分、左側の Pathway view が今期開発分である。今年度はネットワークバリエーションマップを全体的に見直し、前期開発分の多くを今期開発分にまとめたため、表1では nt 番号エントリが 151 から 140 に減少している。実際にはネットワーク要素の N 番号エントリやバリアントのエントリの数は大きく増加している。なお N 番号エントリで+をつけて表示した数は非公開の（内部的には N9 で始まる N 番号の）エントリで、ここに含まれる情報はバリアントエントリで公開されているため、ネットワークバリエーションマップでは N 番号のカラムがなく、簡略化されている。

今年度はネットワークデータの表示法について 2 つの改良を行った。1 つはネットワークバリエーションマップの表示で、以前はネットワーク要素を構成する遺伝子シンボルや矢印などの記号を画像化してアライメント表示を行っていたが、アライメントデータがほしいという要望があったこともあり、アライメントは各ネットワーク要素を 1 行とした html テーブルで表示することとした。もう 1 つはパスウェイマップ上でのネットワーク表示に関するものである。代謝や多くのシグナル伝達パスウェイでは、ネットワーク要素は個々のタンパク質や化合物のつながりであるので、パスウェイマップ上で対応する位置を定義することが比較的容易にできる。これに対し分子集合体が関与する様々な細胞プロセスのパスウェイでは、箱や丸と線を用いたパスウェイマップだけで情報の流れを表現することが難しく、しばしばイラストレーションがつけられている。今回の改良では例えば map04820 Cytoskeleton in muscle cells にあるように、ネットワーク要素は箱や線だけでなく付随するイラストレーションでもハイライト表示できるようにした。

② ウイルスタンパク質のオーソロググループ

本研究ではヒト遺伝子バリアントとともに病原体遺伝子を、生体内分子ネットワークに対するゆらぎ物質として取り上げ、どのような疾患とどのように関わっているかの知識を集約している。病原体としてはとくにウイルスに着目しているが、ウイルス遺伝子やウイルスタンパク質に関して実験データに基づく知識は非常に限られている。そこで手作業で定義している KO（KEGG Orthology）を補完するため、またウイルスタンパク質の全体像を把握するため、計算手法によるオーソロググループ VOG（Virus Ortholog Group）の生成を行っている。なお VOG は当初は VOC（Virus Ortholog Cluster）と呼んでいたが、クラスターは遺伝子クラスターのように染色体上で保存された遺伝子の並びとして使っており、KO クラスターや VOG クラスターを考えるためにも、この名称変更を行った。

計算手法は昨年度に報告した通り、KEGG GENES に対し SSEARCH プログラムで全対全のアミノ酸配列比較をして作成される各遺伝子ごとのベストヒット配列リスト（GFIT データと呼んでいる）を用いている。配列類似性の尺度としては、アライメントされた部分 Overlap のアミノ酸一致度 Identity を、比較する配列長の違いで補正した

Modified identity = Identity * min (1, Overlap*2/(Length1 + Length2))

を用いている。この補正については昨年度から少し変更し、隔月の NCBI RefSeq Release とともに VOG 更新を行う体制を確立した。年度末時点の VOG データは表 2 の通りであった。

表 2. RefSeq Release 223 から生成した VOG (Virus ortholog group)

配列類似度の閾値	30%	50%	70%
グループの数	50,667	76,301	87,377
グループに含まれるタンパク質数	605,594	551,101	494,713
ウイルスタンパク質の総数	676,533		

VOG 生成手順については論文発表済みであり[1]、後述する通り VOG に関する解析ツール開発を行っている。また VOG は新たな KO を定義することにも使われており、KO アノテーションがついたウイルスタンパク質の割合は 5%から 7%に増加した。

③ 疾患データベース

DISEASE データベースの高品質化は本研究計画の中核となる部分である。表 1 において疾患エントリ(H 番号エントリ)数の増加よりも、ネットワークバリエーションマップと関連づけられた(nt-linked)エントリ数の増加が、本研究の進捗状況を最もよく反映している。DISEASE データベースの特徴として Gene、Network、Drug の 3 つのフィールドがあり、これらは以下のように構築している。Gene フィールドの疾患遺伝子については OMIM や MEDGEN を参考にして文献情報を調べ、手作業で入力し常に見直しも行っている。Network フィールドの分子ネットワーク情報は、NETWORK データベース構築の際にネットワークバリエーションマップから手作業でリンクづけした H 番号の逆引きであり、治療薬の情報(Drug フィールド)は医薬品データベース構築の際に適応症から手作業でリンクづけした H 番号の逆引きである。疾患データベースの高品質化は KEGG MEDICUS 全体の高品質化と深く関連している。

これら主要フィールド以外に、疾患名の階層化、ICD-11 疾患分類との対応づけ、独自の疾患カテゴリ分類も継続して行った。疾患エントリにはもともと疾患をどのレベルで定義するかの階層が内在し、総称名の下に複数の個別名があるといったことがしばしばある。このような親子関係は内部用の疾患エントリ編集画面で定義されており、公開されている疾患エントリではこれらはサブグループ名として表示され、その逆引きとしてスーパーグループ名も表示されるようになっている。

表 1 の通り年度末の時点で 739 の疾患に分子ネットワーク情報がつけられている。これらの疾患をカテゴリ別に見ると、先天性代謝異常症 231、先天性形成異常 199 で、単一遺伝子疾患と思われるものが半分強をしめている。他に 10 以上の疾患が含まれるカテゴリは、がん 38、神経変性疾患 13、その他の神経系疾患 64、原発性免疫不全症 18、その他の免疫系疾患 52、ウイルス感染症 12、細菌感染症 11、循環器系疾患 17、内分泌代謝疾患 53、血液疾患 35、筋骨格疾患 21 などであった。分子ネットワーク情報は疾患の分子メカニズム理解のためにも、医薬品開発で標的探索等のためにも有用な情報であり、今後ともできるだけ多くの疾患に付与していく。

④ 医薬品データベース

医薬品データベースはすでに高品質化が達成されており、KEGG MEDICUS の中で最も広く利用されているデータベースである。実際に社会で使われている医薬品添付文書などのデータに基づいているため、社会的価値が高い実用的なデータベースでもある。医薬品データベースは、日米欧での医薬品有効成分を

蓄積した DRUG データベース、様々な観点で医薬品をグループ化した DGROUP データベース、それに外部から導入している医薬品添付文書のデータから構成されている。DRUG データベースでは有効成分ごとに医薬品エントリ(D 番号エントリ)を作成しており、日本(PMDA)、米国(FDA)、欧州(EMA)で承認された新薬、および日本(JAN)、米国(USAN)、欧州(INN)に登録された新薬名(開発中のものを含む)に対応する有効成分を直ちに取り込む更新体制は確立している。また各エントリに対して、医薬品の標的分子(Target フィールド)、薬物代謝酵素と薬物トランスポーターの基質(Enzyme フィールド)および阻害・誘導(Interaction フィールド)、効能とくに適応疾患(Efficacy および Disease フィールド)といった独自のアノテーションを付与する品質管理を今年度も継続して行った。

DGROUP データベースは医薬品を Chemical、Structure、Target、Metabolism、Class の 5 つの観点でまとめたグループである。Chemical は塩や水和状態などの違い以外は同一の化学構造をもち、有効成分として実質的に同じ化学物質のグループである。Structure は基本骨格、Target は標的分子、Metabolism は薬物代謝が類似のグループで、医薬品相互作用(併用禁忌や併用注意)を定義する際に使用している。さらに、Class は医薬品添付文書のページで関連する商品一覧を提示するための実用的なグループであり、効能・効果が類似の幅広いグループを定義できるようになっている。これら医薬品グループエントリ(DG 番号エントリ)の追加・更新も継続して行った。

医薬品添付文書については従来からの更新体制を継続し、商品ごとに D 番号、DG 番号、ATC 分類などの対応づけを行っている。日本の医療用および一般用医薬品添付文書は、毎月 1 度、日本医薬情報センター(JAPIC)からデータの提供を受けて内部データベースに登録している。同じタイミングで米国 FDA の National Drug Code (NDC) database から米国の医療用医薬品のリストを取得し、NIH/NLM の DailyMed データベースへのリンクや検索用データを内部データベースに登録している。実際の添付文書データは内部データベースには登録せず、DailyMed を参照する形にしている。

⑤ 解析ツール

解析ツールの研究開発項目では、ウイルスオーソロググループ(VOG)に関するツール群を開発している。昨年度には全体計画を前倒しして、VOG 検索ツールの開発と Taxonomy マッピングツールで VOG 対応を行い公開した。VOG は類似度の閾値 30%、50%、70%で作成しており、それぞれ 3、5、7 で始まる 6 桁の数字を ID としている。ただし、この ID は隔月の VOG データ更新で変わるので、ツール群はグループに属するウイルス遺伝子 ID を指定することで利用できるようにしている。今年度は各ウイルス遺伝子エントリのページにまず Vog ボタンをつけ、3 つの閾値での類似グループが表示されるようにした。その後、ウイルスゲノム上での遺伝子の並びに関するツール開発を始め、Vog cluster ボタンからゲノム上で前後 5 つずつの遺伝子を含めた 11 個の遺伝子を調べて、同じ VOG(閾値 30%)をもつ他ゲノムの遺伝子をテーブル形式で表示できるようにした。方法は異なるが KEGG 生物種(cellular organisms)の各エントリにつけられている Gene cluster ボタンと同様に、Vog cluster は類似遺伝子の並びの保存領域である。当初の名称であるウイルスオーソログクラスターをウイルスオーソロググループに変更したのは、このようにクラスターをゲノム上で保存された遺伝子の並びとして使うためである。また今後 VOG を使ってウイルスと真核生物・原核生物がどのように遺伝子あるいは遺伝子クラスターをやりとりしているかを解析できるようにする計画である。そこで今年度は VOG を KEGG 生物種にも拡張したデータセットを作成し、定期的に更新できるようにした。

§4. 成果発表等

(1) 原著論文発表

① 論文数概要

種別	国内外	件数
発行済論文	国内(和文)	0 件
	国際(欧文)	1 件
未発行論文 (accepted, in press 等)	国内(和文)	0 件
	国際(欧文)	0 件

② 論文詳細情報

1. Zhao Jin, Yoko Sato, Masayuki Kawashima and Minoru Kanehisa, "KEGG tools for classification and analysis of viral proteins", Protein Science, vol. 32, No. 12, e4820, 2023 (DOI: 10.1002/pro.4820).

(2) その他の著作物(総説、書籍など)

該当なし

(3) 国際学会および国内学会発表

① 概要

種別	国内外	件数
招待講演	国内	0 件
	国際	1 件
口頭発表	国内	0 件
	国際	0 件
ポスター発表	国内	0 件
	国際	0 件

② 招待講演

〈国内〉
該当なし
〈国際〉
1. Minoru Kanehisa, Establishing a self-sustaining database for a sustainable society, HFSP Symposium at ISMB/ECCB 2023, Lyon, France, July 26, 2023

③ 口頭講演

〈国内〉
該当なし

〈国際〉
　該当なし

④ ポスター発表

〈国内〉
　該当なし
〈国際〉
　該当なし

(4) 知的財産権の出願 （国内の出願件数のみ公開）

出願件数

種別		件数
特許出願	国内	0 件

(5) 受賞・報道等

① 受賞

1. 小林財団 小林賞、金久 實、2024 年 2 月 29 日

② メディア報道

1. テレビ大阪ニュース、生命科学分野の研究を評価、2024 年 2 月 29 日

③ その他の成果発表

1. Clarivate Analytics Highly Cited Researchers 2023、金久 實、2023 年 11 月 15 日

§5. 主要なデータベースの利活用状況

(1) アクセス数

① 実績

表 1 研究開発対象の主要なデータベースの利用状況（月平均）

DB 名	種別	2023 年度	2022 年度 (参考)	2021 年度 (参考)
KEGG MEDICUS	訪問者数	3,032,459	2,862,801	1,964,636
	訪問数	6,762,179	6,088,038	3,744,592
	閲覧ページ数	14,667,077	10,985,920	6,570,287

② 分析

KEGG MEDICUS へのアクセス数は 2022 年度に大幅増であったが、2023 年度も増加傾向が続いている。

(2) データベースの利用状況を示すアクセス数以外の指標

Web of Science で引用回数が 1,000 回以上ある KEGG 論文 17 件を引用した論文は 61,949 件あった（2024 年 5 月 20 日現在）。このうち発行年が 2023 年は 7,174 件、2024 年は 2,407 件であった。

(3) データベースの利活用により得られた研究成果（生命科学研究への波及効果）

上記の 61,949 件のうち Web of Science で最新の Hot Papers と指定されたものが 16 件あり、そのうち Nature 系列誌 5 件を含む 10 件は以下の通りであった。

1. Nelson, MR; Liu, P; (...); Huang, YD. The APOE-R136S mutation protects against APOE4-driven Tau pathology, neurodegeneration and neuroinflammation. Nature Neuroscience 2023 Dec;26(12):2104-2121.
2. Tang, DD; Chen, MJ; (...); Wang, YW. SRplot: A free online platform for data visualization and graphing. PLoS One 2023 Nov 9;18(11):e0294236.
3. Camargo, AP; Roux, S; (...); Kyrpides, NC. Identification of mobile genetic elements with geNomad. Nature Biotechnology 2023 Sep 21 (Early Access).
4. Chklovski, A; Parks, DH; (...); Tyson, GW. CheckM2: a rapid, scalable and accurate tool for assessing microbial genome quality using machine learning. Nature Methods 2023 Aug;20(8):1203-1212.
5. Satam, H; Joshi, K; (...); Malonia, SK. Next-Generation Sequencing Technology: Current Trends and Advancements. Biology (Basel) 2023 Jul 13;12(7):997.
6. Zhao, L; Zhang, H; (...); Liang, QQ. Network pharmacology, a promising approach to reveal the pharmacology mechanism of Chinese medicine formula. Journal of Ethnopharmacology 2023 Jun 12:309:116306.
7. Shen, SQ; Zhan, CS; (...); Luo, J. Metabolomics-centered mining of plant metabolic diversity and function: Past decade and future perspectives. Molecular Plant 2023 Jan 2;16(1):43-63.

8. Hosomi, K; Saito, M; (...); Kunisawa, J. Oral administration of Blautia wexlerae ameliorates obesity and type 2 diabetes via metabolic remodeling of the gut microbiota. Nature Communications 2022 Aug 18;13(1):4477.

9. Noor, F; ul Qamar, MT; (...); Aljasir, MA. Network Pharmacology Approach for Medicinal Plants: Review and Assessment. Pharmaceuticals (Basel) 2022 May 4;15(5):572.

10. Moses, L and Pachter, L. Museum of spatial transcriptomics. Nature Methods 2022 May;19(5):534-546.

(4) データベースの利活用によりもたらされた産業への波及効果や科学技術のイノベーション(産業や科学技術への波及効果)

上記の 61,949 件を Web of Science のカテゴリで分類した件数と割合は以下の通りで、産業や科学技術への波及効果を含むと考えられるものにアステリスクをつけた。

Biochemistry Molecular Biology	9,323	(15.0%)
*Biotechnology Applied Microbiology	7,923	(12.8%)
Genetics Heredity	7,709	(12.4%)
Multidisciplinary Sciences	7,616	(12.3%)
Microbiology	6,731	(10.9%)
Biochemical Research Methods	4,928	(8.0%)
Mathematical Computational Biology	4,837	(7.8%)
Plant Sciences	3,609	(5.8%)
Oncology	2,908	(4.7%)
Cell Biology	2,736	(4.4%)
*Computer Science Interdisciplinary Applications	2,312	(3.7%)
*Medicine Research Experimental	2,004	(3.2%)
Chemistry Multidisciplinary	1,831	(3.0%)
*Pharmacology Pharmacy	1,747	(2.8%)
Biology	1,615	(2.6%)
*Environmental Sciences	1,486	(2.4%)
Immunology	1,454	(2.3%)
Statistics Probability	1,426	(2.3%)
Neurosciences	1,114	(1.8%)
*Food Science Technology	926	(1.5%)
Biophysics	907	(1.5%)
Ecology	882	(1.4%)
Evolutionary Biology	878	(1.4%)
*Marine Freshwater Biology	860	(1.4%)
Endocrinology Metabolism	775	(1.3%)
*Chemistry Medicinal	753	(1.2%)

§6. 研究開発期間中に主催した活動(ワークショップ等)

(1) 進捗ミーティング

年月日	名称	場所	参加人数	目的・概要

(2) 主催したワークショップ、シンポジウム、アウトリーチ活動等

年月日	名称	場所	参加人数	目的・概要

以上

別紙1　既公開のデータベース・ウェブツール等

No.	正式名称	別称・略称	概要	URL	公開日	状態	分類	関連論文
1	KEGG MEDICUS		ゲノムの情報と疾患・医薬品との関連を、生体システムを構成する分子ネットワークを通して統合的に理解し、ヒトゲノム情報および病原体ゲノム情報の有効利用を促進するためのリソースである。また日本と米国のすべての医薬品添付文書も統合されており、一般社会にとっても有用なリソースである。	https://www.kegg.jp/kegg/medicus/	2010/10/1	維持・発展	データベース等	Minoru Kanehisa, Miho Furumichi, Yoko Sato, Masayuki Kawashima and Mari Ishiguro-Watanabe, “KEGG for taxonomy-based analysis of pathways and genomes”, Nucleic Acids Research, vol. 51, No. D1, pp. D587-D592, 2023 (DOI: 10.1093/nar/gkac963)
2	医薬品相互作用チェック		与えられた医薬品リストの中に併用禁忌・併用注意に該当する相互作用があるかを判定するツール。KEGG MEDICUSにある医薬品添付文書に記載された相互作用をすべて抽出し、標準化したデータセットを用いている。	https://www.kegg.jp/medicus-bin/ddi_manager	2016/4/1	維持・発展	ツール等	