

ライフサイエンス分野の統合データベース整備事業

ライフサイエンス知識の階層化・統合化事業

19年度 研究成果報告書

平成20年3月

国立大学法人京都大学 化学研究所 金久實

本報告書は、文部科学省の科学技術試験研究委託事業による委託業務として、京都大学が実施した、平成19年度の「ライフサイエンス知識の階層化・統合化事業」の成果を取りまとめたものです。

従って、本報告書の著作権は、文部科学省に帰属しており、本報告書の全部または一部の無断複製等の行為は、法律で認められたときを除き、著作権の侵害にあたるので、これらの利用行為を行うときは、文部科学省の承認手続きが必要です。

1. 委託業務の目的

本計画は現在すでに世界有数のバイオ情報サービスとなっているゲノムネットを京都大学の事業と位置づけ、化学研究所バイオインフォマティクスセンターにおいて分子情報を中心とした統合データベースを構築する。日本語での統合DB検索システムを半年後に提供し、革新的なウェブ技術とKEGGにおいて人手で構築された知識の体系を融合して、平成22年までにライフサイエンス分野における世界最高水準の知的情報基盤を確立する。

2. 平成19年度の実施内容

2. 1 実施計画

(1) 共通基盤技術開発

統合データベースを構築する基盤技術はこれまでのKEGGプロジェクトですでに確立しているので、本計画では統合データベースを利用するための技術開発が中心となる。利用の形態は大きく分けてキーワード検索と、類似性検索や解析・予測などのソフトウェア利用がある。これらを2つの業務項目とし、平成19年度は以下の開発を行う。

①知識処理技術開発

ソフトウェア利用については、平成19年度は化合物に関するソフトウェアを集約的に整備する。特に化合物の化学構造比較、化学反応予測など、化学研究所バイオインフォマティクスセンターの研究成果をもとにしたソフトウェアを実用化し、ゲノムネットサービスとして順次公開する。

②ウェブ技術開発

キーワード検索の基本的なものはすでにDBGETシステムで運用しているので、日本語での利用環境作りとして、入力した日本語キーワードの翻訳機能と検索結果画面の英単語を日本語に翻訳するための支援ツールを開発する。これらを実装したゲノムネットサービスを平成19年10月に運用開始する。また現行の検索システムに代わる革新的なウェブ検索システムは3年間の開発期間を設定し、その第1段階の開発研究を開始する。

(2) 統合データベース開発・運用

本計画では既存データベースのデータ間のつながりを蓄積した LinkDB データベースが統合化の基本となる。一方では、我が国において公共的に利用可能なレポジトリが存在しない医薬品や化合物について、新規データベースの開発も行う。これらを2つの業務項目とし、平成19年度は以下の開発を行う。

①統合データベース開発

LinkDBデータベースは、これまでゲノムネットのDBGETシステムで運用している内部データベースが対象であったが、外部データベースについてもデータベース間のクロスレファレンス情報を2項関係ファイルとして取り込むことができる枠組みを開発し、検索コマンドの高機能化を行う。またDBGETシステムやデータベースの日々更新を行うSEQNEWシステムの高機能化も行う。LinkDBの検索効率を高めるために、

含まれるデータの階層化・重複除去作業に着手し、塩基配列データベースでこれをまず実現して公開する。

②医薬品・化合物データバンク開発

医薬品および化合物情報について平成19年度は、KEGG DRUGにある薬の名称、化学構造、薬効、ターゲットなどの情報と、JAPICが提供する医薬品の添付文書情報、特に副作用情報を統合し、日本語での医薬品・化合物データベースの最初のバージョンとして提供する。

(3) プロジェクトの総合的推進

分担機関である京都大学は、中核機関である情報・システム機構の全体戦略に従い連携して本事業を推進する。

医薬品・化合物に関する外部有識者を含む技術検討会を開催して、プロジェクトの推進に資する。本プロジェクトの成果は直ちにゲノムネットサービス (<http://www.genome.jp/>) に反映し、利用者の意見を収集して今後の展開に資する。

2. 2 実施内容 (成果)

本統合データベースプロジェクトはゲノムネット (www.genome.jp) をKEGGと分離して開発・運用するために提案し実施してきた。KEGGは現時点ではゲノムネットの主要サービス (www.genome.jp/kegg/) であるが、京都大学と東京大学の金久研究室が別予算で構築しており、KEGG独自のウェブサイト (www.kegg.jp) も存在する。本計画ではゲノムネットを京都大学の事業と位置づけ、DBGET/LinkDBシステムを中心に統合化を行うものである。KEGGは統合化の対象データベースの中心であり、またケミカル情報解析ツールの一部はこれまでKEGGの中で開発されていたものを引き継いで本計画で開発している。初年度は当初計画で掲げたDBGETの一括検索、日本語支援機能の導入、LinkDBの高機能化に重点を置いて開発し、すべて公開済である (図1参照)。また採択後に中核機関との話し合いの中で追加事項として計画に取り入れた医薬品・化合物のデータベース開発では、KEGGとの重複を避けるため手作業でのデータ収集・統合ではなく、JAPICなど他データベースの導入とLinkDBを用いた統合化を行った。初年度の実施内容は以下の通りである。

(1) 共通基盤技術開発

①知識処理技術開発

ゲノムネット利用形態のうちソフトウェア利用については、平成19年度は化合物に関するソフトウェアを集中的に整備した。特に化学研究所バイオインフォマティクスセンターの研究成果をもとにして実用化した化合物類似構造検索ツールSIMCOMP、糖鎖類似構造検索ツールKCaM、化学構造変化に基づく反応予測ツールe-enzymeをゲノムネットのケミカル情報解析ツールとして整備し公開した (図2参照)。また、マイクロアレイデータからの糖鎖構造予測ツールGECS (Gene Expression to Chemical Structure) を開発し、平成20年4月1日の公開に向けて整備した。SIMCOMPについては、サイズの大きな化合物に対する検索効率の悪さが問題となっていた。そこで、化合物構造を比較する際のグラフ表現と比較アルゴリズムについて複数の方法を調査・検討し、現状のものと比較した。その結果、高速化の目処がたったため、平成20年度中に高速化を実現する予定である。

図1. ゲノムネットの日本語版ホームページ。ゲノムネットは1992年より京都大学化学研究所で開発・運用を行っているライフサイエンス分野の統合情報リソースである。ゲノムネットの1日あたりの総アクセス数は100万件程度、1日あたりアクセスのあったユニークIPアドレス数は1万5千程度(大学等の組織ではproxy経由で多数の利用者が同一アドレスでアクセスしているので実際の利用者数は2万~3万と推定される)で、ライフサイエンス分野では我が国最大のまた国際的にも有数の情報サービスとなっている。平成19年度に本プロジェクトにおいて赤字で囲んだ部分、すなわち日本語支援の辞書ツールの新規開発、DBGETの全データベース(統合データベース)一括検索機能開発、多数の外部データベースを含むLinkDBの高機能化、ゲノムネット医薬品データベースの新規開発、化合物関連情報の解析ツールの整備を行った。

ゲノムネット - 計算ツール

http://www.genome.jp/ja/gn_tools_ja.html

GenomeNet

KEGG KEGG2 PATHWAY BRITE DRUG DBGET

環境設定 辞書ツール ヘルプ [English | Japanese]

Search 統合データベース for Go Clear

ゲノムネット
ゲノムネットとは
お知らせ
謝辞

KEGG
KEGGの概要
リリース情報

統合データベース
統合DBの概要
DBGETの概要
リリース情報
データベース増加図

医薬品データベース
利用法

研究支援データベース

計算ツール
その他のツール

フィードバック

ゲノムネット計算ツール

ゲノムネットでは以下の3つのカテゴリーで計算サービスを提供しています。配列解析の標準的なプログラム以外はすべて京都大学化学研究所バイオインフォマティクスセンターで開発されたものです。

配列解析

| | |
|-----------------|---------------------------|
| BLAST | 配列類似性検索 |
| FASTA | |
| MOTIF | 配列モチーフ探索 |
| CLUSTALW | 配列のマルチプルアライメント、および進化系統樹解析 |
| MAFFT | |
| PRRN | |

ゲノム情報解析

| | | |
|--------------------|-------------------------------------|--|
| KAAS | ゲノムまたはESTコンティグの自動アノテーションとバスウェイマッピング | Moriya et al. (2007) |
| EGassembler | 大量のESTデータからコンセンサスコンティグ自動生成 | Masoudi-Nejad et al. (2006) |
| GENIES | カーネル法での多様なオミクスデータ統合による遺伝子ネットワーク予測 | Yamanishi et al. (2005) |
| GECS | マイクロアレイデータからの糖鎖構造予測 | Kawano et al. (2005) Suga et al. (2007) |

ケミカル情報解析

| | | |
|----------------|----------------|--|
| SIMCOMP | 化合物類似構造検索 | Hattori et al. (2003) |
| KCaM | 糖鎖類似構造検索 | Aoki et al. (2004) |
| e-zyme | 化学構造変化に基づく反応予測 | Kotera et al. (2004) Oh et al. (2007) |

» その他のツール

京都大学化学研究所バイオインフォマティクスセンター

図2. ゲノムネット計算ツールは配列解析、ゲノム情報解析、ケミカル情報解析に大別され、配列解析以外はすべて京都大学化学研究所バイオインフォマティクスセンターの研究成果を実用化したものである。本プロジェクトでは赤枠で囲んだ部分の新規開発または機能向上を行った。

②ウェブ技術開発

もう1つの利用形態であるキーワード検索の基本的なものはすでにDBGETシステムで運用しているので、日本語での利用環境作りとして、入力した日本語キーワードの翻訳機能と検索結果画面の英単語を日本語に翻訳するための支援ツールを開発し、平成19年10月1日に運用を開始した(図3参照)。また、革新的なウェブ検索システムについては第1段階の開発研究として、ゲノムネットで提供する全データベースに対する一括検索機能を開発し、7月1日に運用を開始した(図4参照)。同時にゲノムネットにインストールされていない外部データベースへのリン

クや等価なエントリー間をつなぐリンクを含む LinkDB 新バージョンの検索機能を実装した。その結果、一部の外部データベースも内部データベースと同様にキーワード検索できるようになった。

ライフサイエンス辞書 (LSD) を用いた日本語支援ツール

LSD 和英辞書の使い方
「辞書ツール」をクリックして起動する。日本語キーワードを入力し、提示された英語のキーワードをクリックして、統合DB検索ボックスへ移す。

LSD 英和辞書の使い方
単語または熟語をマウスで選択 (ハイライト) 表示させ、shift-D を押すと、ポップアップウィンドウに和訳が表示される。あとは選択箇所を変更するだけで和訳が表示される。

図 3. ライフサイエンス辞書 (LSD) を用いた和英辞書ツールと英和辞書ツールの使用例。

全データベース一括検索機能

統合データベースを選択しキーワードを入力するとメニューに表示されているデータベース全てに対するキーワード検索となる。

検索結果は各データベースの最初の数エントリーずつ表示され、複数のデータベースが見渡せるようにしている。KegDrawやJmolなど構造表示ツールへの直接リンクも付けている。

図 4. ゲノムネット統合データベースの全データベース一括検索の例。

(2) 統合データベース開発・運用

①統合データベース開発

LinkDB データベースでは、これまでゲノムネットの DBGET システムで運用している内部データベース（表1のカテゴリー1とカテゴリー2）が検索の対象であったが、本プロジェクトにおいて、外部データベースについてもデータベース間のクロスリファレンス情報を2項関係ファイルとして取り込み、検索ができるような枠組みの開発を行った。取り込むべき外部データベースの最初の対象を、DBGET システムで既に運用している内部データベースから参照されているデータベースとして開発を行い、これにより、内部データベースに加え 500 以上の外部データベースが LinkDB の検索対象となった。また、検索コマンドの高機能化を行い、外部データベースと内部データベースを区別なく検索する機能および等価なリンクを扱うための機能を実装した。DBGET システムおよび SEQNEW に関しては全データベース一括検索用の改良等高機能化を図った。また、LinkDBの検索効率を高めるため、データの階層化・重複除去作業に着手した。その第一段階として、塩基配列データベース GenBank, EMBL, DDBJ を対象とした重複除去作業を行い、DBGET システムにおいて INSDC という一つのデータベースとして検索できるようにした。キーワード検索結果画面からは、上記3つのデータベースすべてにリンクが張られており、どのデータベースのエントリーも直接検索できる。

表1. ゲノムネット統合データベースのカテゴリー

| カテゴリー | bget bfind blink brite | 数 | 内訳 |
|--------------|------------------------|------|------------------------------------|
| 1. KEGG | ○ ○ ○ ○ | 18 | KEGGを構成するコアデータベース |
| 2.ミラーしているDB | ○ ○ ○ × | >16 | RefSeq, UniProt等の主要DB |
| 3.検索可能な外部DB | × ○ ○ × | 2 | INSDC(DDBJ/GenBank/EMBL), InterPro |
| 4.リンクのみの外部DB | × × ○ × | >500 | www.genome.jp/dbget/linkdb.html参照 |
| 5. PubMed | ○ × ○ × | 1 | |

(注) bget: エントリー取得、bfind: キーワード検索、blink: リンク検索、brite: 機能階層検索

②医薬品・化合物データバンク開発

医薬品および化合物情報について、JAPIC が提供する医薬品添付文書情報のうち、医療用医薬品に関するものをゲノムネット医薬品データベース第1版として平成19年9月1日に公開し、さらに一般用医薬品に関するものも含めたバージョンを第2版として平成20年1月28日に公開した（表2にアクセス数を、図5に概略を示した）。医療用医薬品データベースとKEGG DRUG 中の対応する医薬品へのリンク付けを行い、医薬品の名称、化学構造、薬効、ターゲット情報との統合を実現した。また、参考文献のうち可能なものについては PubMed や J-STAGE へのリンク付けも行っている。

表2. ゲノムネット医薬品データベース (<http://www.genome.jp/kusuri/>) のアクセス数

| | アクセス数 | ユニークIP数 |
|----------|---------|---------|
| 2007年10月 | 41,622 | 699 |
| 2007年11月 | 55,943 | 1,126 |
| 2007年12月 | 57,035 | 2,221 |
| 2008年1月 | 95,870 | 3,614 |
| 2008年2月 | 393,331 | 9,764 |
| 2008年3月 | 267,892 | 10,478 |

(注) ゲノムネット全体の月単位のユニークIP数は 200,000で、その15%が国内からである。従って国内利用者の約1/3がゲノムネット医薬品データベースを利用していることになる。

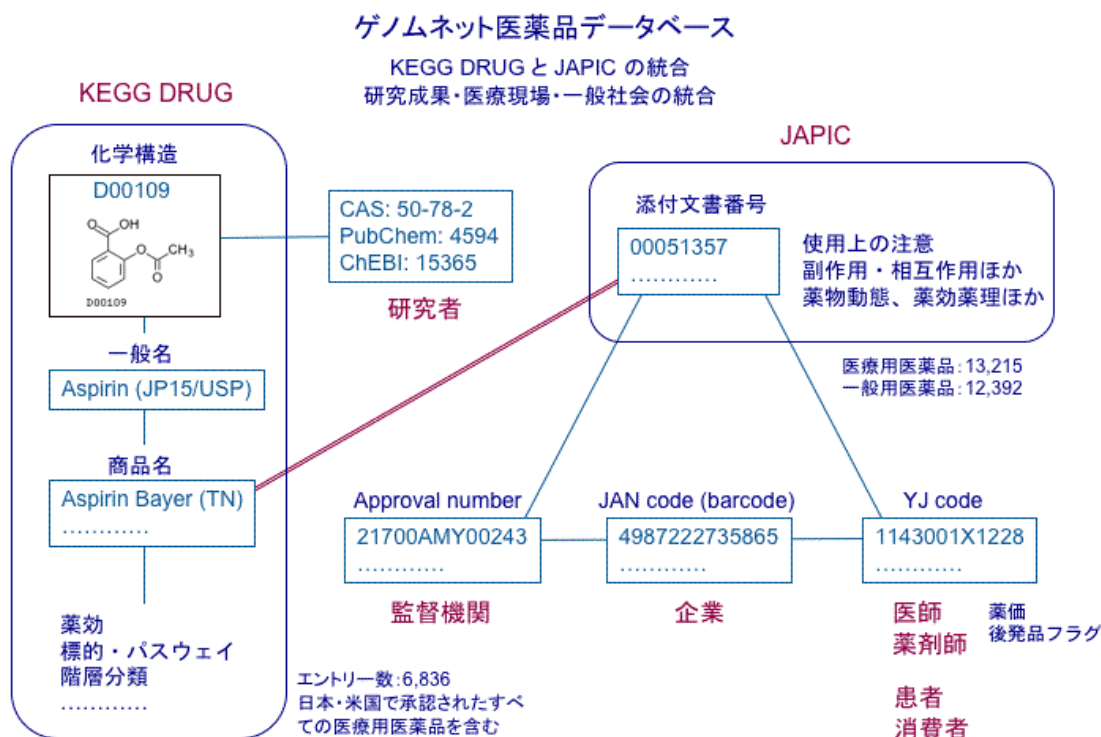


図5. ゲノムネット医薬品データベースの概略。

医薬品・化合物データベースに関しては、有料サービスが多く、無料サービスの場合でも知財がからむ場合が多い。JAPIC も本プロジェクトのデータベースとしてサービスするにはライセンス契約が必要である。したがって、多様なデータベースの統合を進めるには、上記 LinkDB を用いたリンク情報を中心とした形態が有効であると考えられるため、今後は LinkDB を中心とした統合化を進める予定である。

(3) プロジェクトの総合的推進

プロジェクトの推進にあたっては、中核機関である情報・システム研究機構の全体戦略の下、中核機関と日頃より連携して進めてきた。本プロジェクトの開始時点において本学は、Google-likeな全文検索システム化を含む分子情報統合データベースシステムの構築を担当することとしていたが、中核機関との協議の中で中核機関と本学の役割分担を整理し、中核機関では全文検索などの一般的技術の開発を担当することとし、本学では、化合物・医薬品を中心とする分子情報を利用した検索機能の高機能化を主に担当して統合データベースを構築することとした。このため来年度以降に予定していた革新的なウェブ検索システムの第2段階以降の開発は中止することとなった。

また、中核機関との連携を図りつつ進めてきたほか、日頃から頻度高く技術検討委員と意見交換を行い、助言を得ることができた。特に、京都大学薬学研究科の業務協力者である藤井信孝教授からは JAPIC との協力関係構築において非常に多大な助言を得ることができ、プロジェクトの推進に資することができた。また、大阪大学蛋白質研究所の中村春木教授からは、化合物2次元構造データを3次元立体構造化したデータベース LIGAND BOX との連携について助言を得ることができ、化合物・医薬品データベースの統合化推進に資することができた。本プロジェクトは主

に京都大学化学研究所で推進しているが、薬学研究科の金子周司教授とはライフサイエンス辞書を用いたゲノムネットの日本語化に関して緊密に連携することができた。

本年度の成果については、すべてゲノムネット統合データベースとして公開しており、中核機関の横断検索のインデクシング対象となっている。化合物データベースに関しては、上記(2)②に記述したように、知財がからむ場合が多いので、中核機関に知財を移譲する形での統合ではなく、相互リンクでの統合を進めている。

ゲノムネットには1日あたり100万件のアクセスと2万人の利用者があり、世界有数のバイオインフォマティクスサービスとなっている。ゲノムネットを今後とも京都大学化学研究所バイオインフォマティクスセンターの事業として発展させるために、本年度はサーバーなどの設備投資、ソフトウェア開発と運用に重点投資を行った。また、本データベースのより一般ユーザーへの利用拡大と普及のために、平成20年1月にはゲノムネットデータベース利用講習会を開催した。講習会参加者からは、特に反応中の化合物構造変化パターンに関するコメントを得ることができ、RPAIR データベースや e-zyme の改良に資することができた。また、サイボウズを用いたゲノムネットフィードバックのページを通して、利用者からの意見収集・質問応対を行っている。

2. 3 成果の外部への発表

(1) 論文寄稿

和誌、医薬品の統合データベース、金久實他、蛋白質核酸酵素、52、12、1486-1491

(2) 講演

国内、医薬品情報統合データベースの開発、伊藤真純他、BMB2007、2007.12.11-15、ポスター発表

(3) データベースの公開

ゲノムネット医薬品データベース

<http://www.genome.jp/kusuri/>

研究の最先端と医療の現場さらには一般社会をつなぐ日本語の医薬品統合データベース。JAPIC 医薬品添付文書情報（医療用医薬品 13,973 件、一般用医薬品 12,658 件、平成20年4月現在）を検索可能。KEGG DRUG の構造情報やターゲット情報と統合している。また、PubMed や J-STAGE など文献データベースへのリンクも付加している。医療用医薬品は平成19年9月、一般用医薬品は平成20年1月より公開している。

DBGET/LinkDB: ゲノムネット統合データベース検索システム

http://www.genome.jp/ja/gn_dbget_ja.html

平成18年度までに DBGET/LinkDB として開発してきたシステムを、日本語

支援環境の整備、LinkDBの拡張、新たな検索システムの開発という観点から改良したもの。全データベース一括検索と外部データベースを含むLinkDB検索を平成19年7月に、日本語支援環境を平成19年10月に公開した。

(4) データベース基盤システム、ツールの公開

SIMCOMP

<http://www.genome.jp/tools/simcomp/>

類似化合物検索システム。グラフ比較に基づいた精度の高い類似度計算を実現している。検索速度に問題があったため、平成19年度には高速化についての調査を行い、平成20年度に高速化を実現する。

e-zyme

<http://www.genome.jp/tools/e-zyme/>

化学構造変化に基づく反応予測システム。基質と生成物を与えると、その間の反応パターンを予測、EC番号との対応付けなどを行う。テンプレートとなる反応パターンの充実が課題であったため、平成20年度に反応パターンデータベースを整備し、化合物データとリンクさせる。また、平成21年度以降に、複数反応ステップの予測システムを実現する。

KCaM

<http://www.genome.jp/tools/kcam/>

糖鎖類似構造検索システム。糖鎖に特徴的な木構造のための動的計画法を実装したシステムであり、ユーザーインターフェースを他のシステムと統一した。今後は、以下の糖鎖構造予測システムとの連携を計画している。

GECS

<http://www.genome.jp/tools/gecs/>

遺伝子発現データから化合物構造を予測するシステム。ゲノム情報と化合物情報を結ぶためのシステムとして開発している。平成19年度は糖転移酵素のリストから合成可能な糖鎖構造を予測するシステムを開発し、平成20年4月に第1版を公開した。今後は、ユーザーインターフェースなどを改良するとともに、脂質など他の化合物のためのシステムを開発し統合する。

2. 4 活動

平成20年1月30日、31日

ゲノムネットデータベース講習会

発表者：五斗進他、開催場所：東京大学

概要：PCを用いた実習形式での講習会。ホームページ上で一般から20名の参加

者を募った。大学、公的機関の研究所、企業から幅広く集まった。

2. 5 実施体制

| 研究項目 | 担当機関等 | 研究担当者 |
|---|--|--|
| 1. 共通基盤技術開発 (1) 知識処理技術開発 (2) ウェブ開発技術 | 京都大学化学研究所 京都大学化学研究所 京都大学化学研究所 京都大学化学研究所 | ◎○金久實 山西芳裕 馬見塚拓 瀧川一学 |
| 2. 統合データベース開発・運用 (1) 統合データベース開発 (2) 医薬品・化合物データベース開発 | 京都大学化学研究所 京都大学化学研究所 京都大学大学院薬学研究科 京都大学化学研究所 京都大学化学研究所 京都大学化学研究所 京都大学化学研究所 | ○ 五斗進 伊藤真純 金子周司 服部正泰 時松敏明 藤田征志 奥田修二郎 |
| 3. プロジェクトの総合的推進 | 京都大学化学研究所 | ◎○金久實 |

注1. ◎:課題代表者、○:サブテーマ代表者

注2. 本業務に携わっている方は、全て記入。