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

２０年度　研究成果報告書

平成２１年３月

国立大学法人京都大学　化学研究所　　五斗進

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２０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。

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。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２０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、本計画では化合物をはじめとする分子データを扱う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。

①知識処理技術開発

平成１９年度に引き続き、化合物の化学構造と化学反応に関する解析技術を開発し、ソフトウェア化を行う。とく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、化学構造比較の高速化、化学反応ネットワーク予測、酵素番号の自動割り当てなどを行い、実用的なソフトウェアをゲノムネットサービスの一部として順次公開する。

②ウェブ技術開発

平成１９年度に日本語支援ツール開発は終了したので、今年度以降は残りの新規検索エンジン開発を中心に行う。具体的には現行のDBGETシステムのキーワード検索コマンドbfindをよりアップグレードした高速・高機能な検索エンジンの開発を行い、最初のバージョンを提供する。

（２）統合データベース開発・運用

本計画ではライフサイエンスの観点から、医薬品・化合物データベースの開発と運用を行う。また、統合化の基本となるLinkDBの開発・運用を継続して行う。

①医薬品・化合物データベース開発・運用

平成１９年度は医薬品を中心に開発を行い、ゲノムネット医薬品データベースとして公開した。JAPIC添付文書情報の更新とKEGG DRUGとのリンク付けなど毎月の更新作業を継続して行う。また糖鎖・脂質を含む化合物に関して、食品や環境物質といった生体システムや病原性との関連物質の知識を統合し、ゲノムネット化合物データベースとして提供する。

②LinkDB開発・運用

データベース間のリンク情報に関するLinkDB検索システムの高速化、データベースフォーマットの変更や新規データベースの追加対応のためのSEQNEWシステムの高機能化、分子情報データベースの日々更新作業を継続して行う。

（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。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とともに、中核機関での横断検索システムでの検索対象となるようにする。

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。平成２０年度は当初計画で掲げた化学構造・化学反応解析ソフトウェアの高機能化、化合物データベースの開発、DBGET/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）。また、中間評価および中核機関との調整を考慮し、検索エンジンに関しては化合物・医薬品により特化したものを検討した。平成２０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

平成１９年度に引き続き、化合物の化学構造と化学反応に関する解析技術を開発し、ソフトウェア化を行った。具体的には、平成１９年度に実施した調査に基づき化合物類似構造検索ツールSIMCOMPの検索効率を改善したプログラムを実装し、平成２０年７月１日に公開環境へと反映した。これにより、１０倍程度の高速化を実現できた。また、化合物部分構造検索ツールSUBCOMPにおける、特定の構造に対する問題点を改良したバージョンを実装し、平成２０年１２月１日に公開環境へと反映した。さらに、光学異性体を区別して検索する機能の追加を検討し、いくつかの検索オプションとともにSIMCOMPに実装し、平成２１年４月１日に公開環境へと反映した（図２）。SUBCOMPへの同機能の実装は、クエリに含まれる構造の検索機能（従来はクエリを含む構造の検索のみ）の実装とともに進めており、平成２１年度の早い段階で公開する予定である。

化学反応ネットワーク予測ツールに関しては、プロトタイプは完成し、現在高速化を進めている。平成２１年度中のウェブによるサービスを実現する。酵素番号自動割り当てツールE-zymeに関しては、反応パターンの階層的な表現を用いることによる改良を実現し、平成２１年１月１日に公開環境へと反映した（図３）。

図１．ゲノムネット統合データベース検索システムの日本語ページ。ゲノムネットでサポートしている化合物データベースをまとめて、ゲノムネット化合物データベースとして参照できるようにしている。検索ボックスでは「医薬品」を選択すると医薬品関係のデータベースを、「化合物・糖鎖」を選択すると医薬品以外の化合物関係データベースを横断的に検索することができる。平成１９年度から引き続き開発している医薬品データベースと化学構造・化学反応解析ソフトウェアへのリンクは左側の赤枠で示している。

図２．化合物類似構造検索ツールSIMCOMPの検索画面。構造比較用のグラフ作成方法として共有結合をベースにした方法を実装し、高速化を実現した。また、光学異性体の区別をできるようにし、特に何も指定しない場合は区別するようにした。この結果、グルコースとガラクトースの区別などができるようになるとともに、脂質などに見られる cis 結合と trans 結合の区別もできるようになった。

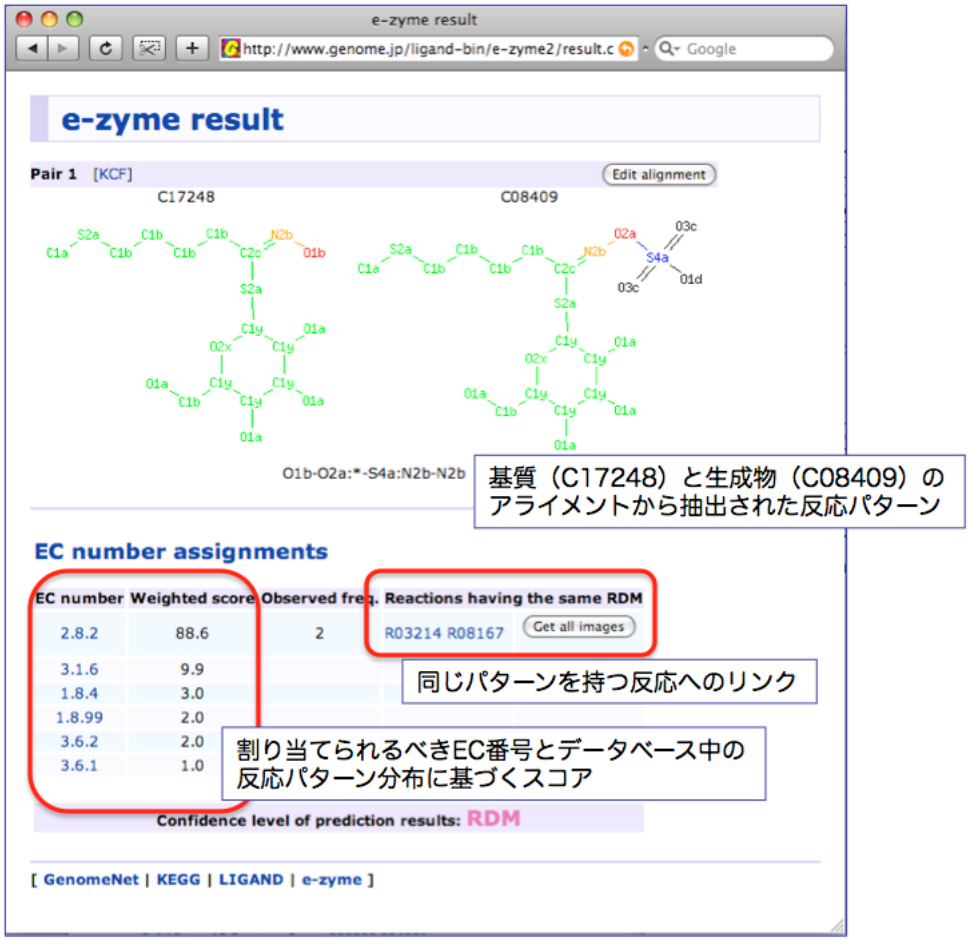

図３．新規反応に酵素番号を自動で割り当てるツールE-zymeの実行結果画面。基質と生成物をSIMCOMPでアライメントし、反応パターンを抽出したのち、データベース中から同じようなパターンを持つ反応を探し出して、酵素番号を割り当てる。平成２０年度はデータベース中の反応パターン分布に基づくスコアリング（Weighted score）を導入し、精度を上げるとともに、データベース中の類似反応を表示するインタフェースを追加した（Reactions having the same RDM）。

②ウェブ技術開発

中間評価および中核機関との調整を考慮し、高速・高機能な新規検索エンジンの開発として、化合物・医薬品により特化したものを検討した。平成２０年度は、構造検索と他の検索を組み合わせることによる高機能化を検討し、構造検索の結果からパスウェイ情報や階層的な機能分類情報へと関連検索ができるシステムを実現した（図４）。本システムは平成２１年４月に公開した。また、DBGETシステムのインデックス方式をgdbmからtinyCDB方式へと変更[1]し、インデクシングの高速化を実現した。

[1] gdbm は古くから UNIX のファイルシステム管理に利用されきたインデクシングシステムであり、多くのデータベース管理にも利用されている。しかし、データベースのサイズが大きくなるとインデックスも大きくなり、インデックス作成に時間がかかるという問題がでてきていた。tinyCDB ではデータベースの更新時にはデータの追加でなくインデックスの再作成が必要であるが、その分、インデックスサイズも小さく、作成時間も大幅に短縮されている。

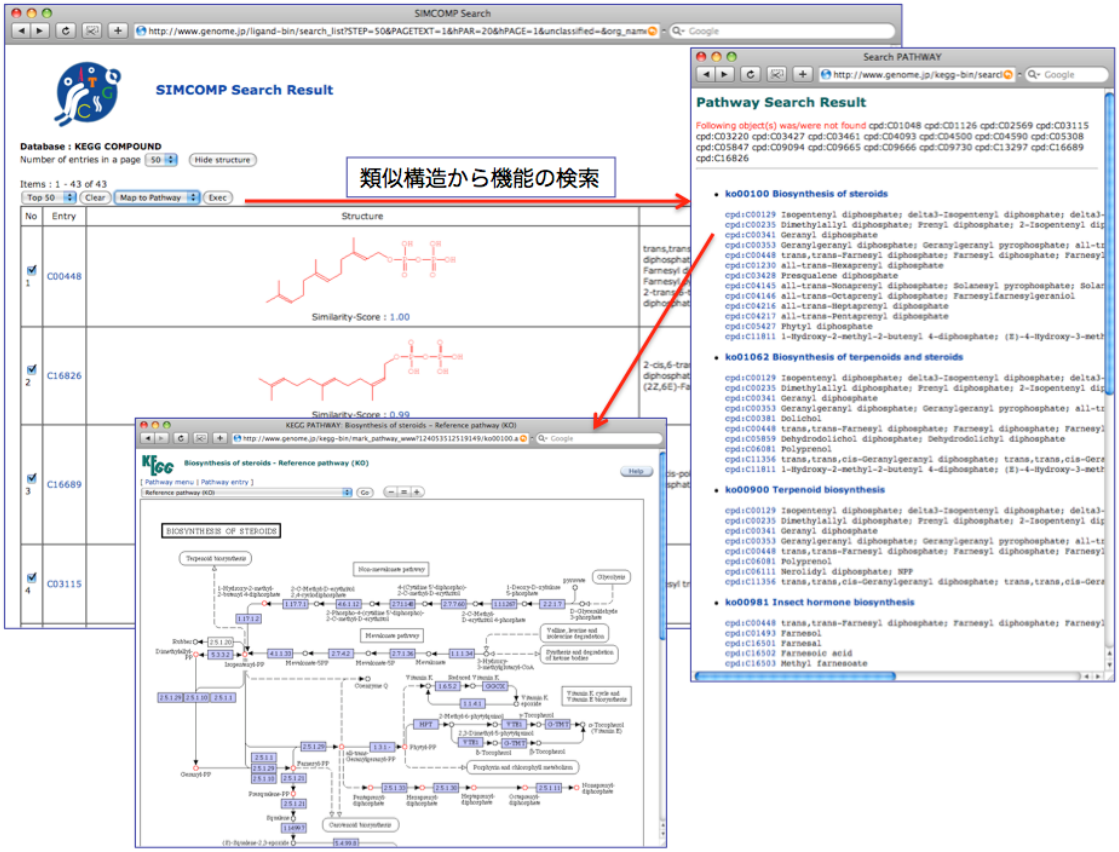

図４．化合物類似構造検索結果から機能の手がかりを探すためのリンク。現在は検索結果のリストに出てくる化合物を選択してKEGG PATHWAYとBRITEとの関係を検索できるようになっている。

（２）統合データベース開発・運用

①医薬品・化合物データベース開発・運用

平成１９年度に引き続き、JAPIC添付文書情報の更新とKEGG DRUG、PubMed、J-STAGEとのリンクづけを、毎月の更新に伴い継続的に行っている。また、ライフサイエンス辞書を利用した同義語・類義語検索のテスト版をゲノムネット医薬品データベースに実装した。ゲノムネット化合物データベースとしては、糖鎖・脂質を含む化合物データベースとして、LIPIDMAPS、KNApSAcKにKEGG COMPOUNDとGLYCANを含めた統合検索を実現した。また、化合物全般を扱うPubChem、ChEBI、立体構造情報を扱うPDB-CCD、3DMET、医薬品を扱うDrugBank、KEGG DRUGについても同様に統合検索を実現した（図１参照）。

②LinkDB開発・運用

ゲノムネット化合物データベースとして提供する７つを統合検索できるようにLinkDBおよびSEQNEWを改良した。また、JAPIC、DailyMed、LipidBank、LigandBoxの４つの医薬品・化合物データベースについては、他の医薬品・化合物データベースのエントリーとの対応関係を定義し、LinkDBによる統合検索を実現した（図５）。引き続き、MassBankや日化辞ウェブとの対応関係の実装についても検討している。

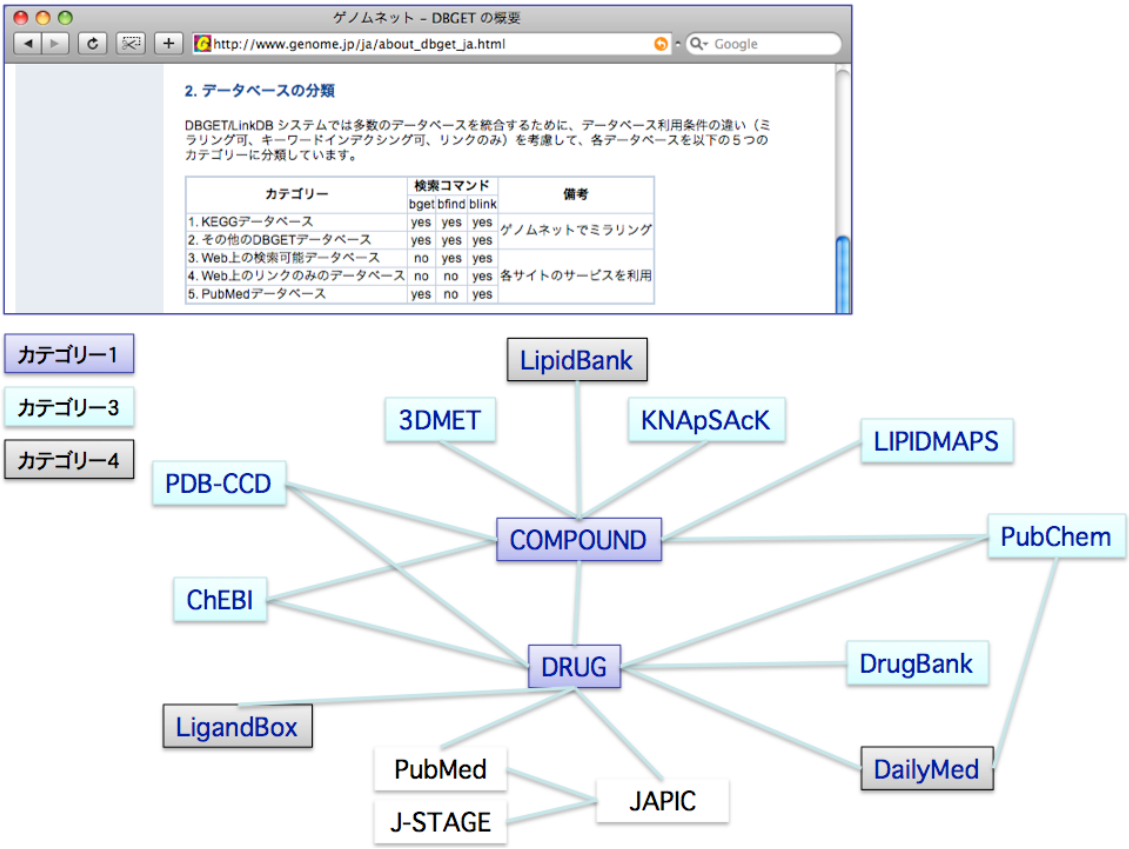

図５．ゲノムネットでサポートしている化合物・医薬品データベースとそのサポート状況。カテゴリー１から３までは通常のキーワード検索が可能。カテゴリー４のデータベースはLinkDBでのみ検索可能であるが、LipidBankについては平成２１年度にキーワード検索対応を予定している。COMPOUNDとDRUGを中心に各データベースの同一化合物に対してリンクが張られている。またJAPICやJ-STAGEといった日本語によるデータベースに対しても関連付けを行っている。

（３）プロジェクトの総合的推進

中核機関の全体戦略に従い、化合物・医薬品を中心とする統合データベースの開発を推進している。また、平成２０年度までの成果はすべてゲノムネット統合データベース、医薬品・化合物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。

ゲノムネット医薬品データベースへのアクセス数は順調にのびており、平成２１年２月のアクセス数が約８５万、ユニークIPアドレスにして約２万２千であり、これは平成２０年２月に比べて倍以上である（表１）。アクセス元のドメインを見てみると、大学・教育関係よりも製薬企業からのアクセスが多い傾向が見える。また、プロバイダ系のドメインも多い。これはロボットの影響もあるが、多数のアクセスポイントからのアクセスがあることを考慮すると、不特定多数の個人がアクセスしていると考えられる。また、平成２１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。

表１．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		アクセス数	ユニークIP数
2007年10月	41,622	699	2008年7月	565,126	16,481
2007年11月	55,943	1,126	2008年8月	528,769	15,656
2007年12月	57,035	2,221	2008年9月	607,950	16,593
2008年1月	95,870	3,614	2008年10月	775,339	19,494
2008年2月	393,331	9,764	2008年11月	759,223	20,161
2008年3月	267,892	10,478	2008年12月	862,632	21,206
2008年4月	297,048	11,129	2009年1月	952,436	22,346
2008年5月	306,584	13,225	2009年2月	862,742	22,745
2008年6月	372,326	13,650	2009年3月	1,006,818	22,639

（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。

２．３　成果の外部への発表

学会等発表実績

委託業務題目：「ライフサイエンス知識の階層化・統合化事業」(ライフサイエンス知識の階層化・統合化事業)

機関名：京都大学

1. 学会等における口頭・ポスター発表

発表した成果（発表題目、口頭・ポスター発表の別）	発表者氏名	発表した場所（学会等名）	発表した時期	国内・外の別
ゲノムネットの化合物情報データベース、口頭発表	五斗進	BIB2008	2008.12.19	国内

2. 学会誌・雑誌等における論文掲載

掲載した論文（発表題目）	発表者氏名	発表した場所（学会誌・雑誌等名）	発表した時期	国内・外の別
E-zyme: predicting potential EC numbers from the chemical transformation pattern of substrate-product pairs	Yamanishi, Y., Hattori, M., Kotera, M., Goto, S., and Kanehisa, M.	Bioinformatics	2009	国外
vurDB: a pathogen-specific sequence database of protein families involved in antigenic variation	Hayes, C.N., Diez, D., Jasonin, K., Honda, W., Kanehisa, M., Wahlgren, M., Wheelock, C.E., and Goto, S.	Bioinformatics	2009	国外

２．４　活動

平成２１年１月２９日、３０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３０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	京都大学化学研究所	小寺正明
	京都大学化学研究所	守屋勇樹
(2) ウェブ開発技術	京都大学化学研究所	○金久實
２．統合データベース開発・運用		
(1) 医薬品･化合物データベース開発・運用	京都大学化学研究所	服部正泰
	京都大学大学院薬学研究科	金子周司
	京都大学化学研究所	時松敏明
	京都大学化学研究所	Nelson Hayes
	京都大学化学研究所	中川善一
(2) 統合データベース開発	京都大学化学研究所	◎○五斗進
３．プロジェクトの総合的推進	京都大学化学研究所	◎○五斗進

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。

２．６　整備実績一覧

報告日：平成２１年４月２８日

整備実績一覧【代表機関名：京都大学】

（1）保有データ情報
特になし

（2）データ（又はＤＢ）の連結、統合化整備（※試験的、限定的公開済みのものも含む。）

通番	データ（又はＤＢ）の名称 ※URL があれば記述	公開／未公開	概要（データの種類（生物種）・数量（kB等）、本プロジェクトで実施した特徴点、進捗状況、今後の計画・課題などを簡潔にわかりやすく記述） ※　公開している場合は、開始年月、利用状況（平均利用者数、アクセス数、ダウンロード数等の数値的指標で記述） ※　必要に応じて画面コピー等の図表添付可
1	ゲノムネット医薬品データベース http://www.genome.jp/kusuri/	公開	研究の最先端と医療の現場さらには一般社会まで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 14,164 件、一般用医薬品 12,119 件、平成２１年３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 などの文献データベースへのリンクも付加している。公開は平成１９年度よりしているが、JAPIC データの更新に伴い毎月の更新を行っている。利用状況については成果報告書の 2.2 を参照のこと。
2	ゲノムネット化合物データベース http://www.genome.jp/ja/gn_cbget_ja.html	公開	平成１８年度に、全データベース一括検索、外部データベースを含む LinkDB 検索、日本語支援環境を整備した「DBGET/LinkDB: ゲノムネット統合データベース検索システム」を化合物・化学反応データベースの統合化に応用したもの。平成２０年度は DrugBank、PubChem、ChEBI、PDB-CCD、LIPIDMAPS、KNApSAcK、3DMET をキーワード検索対象データベースとして、DailyMed、LipidBank、LigandBox を LinkDB 対象データベースとして組み込んだ。

（3）ＤＢ基盤システム、ツール等開発成果物の概要（※試験的、限定的公開済みのものも含む。）

通番	ＤＢ基盤システム、ツール等の名称	公開／未公開	概要（主な機能・特徴点、進捗状況、今後の計画などを簡潔にわかりやすく記述） ※ プログラムプロダクトに限らず、データ形式共通化、標準化のための仕様書、共通規約等のドキュメントについてもリリースしているものは対象とする。（リリース済みドキュメントは参考として目次一覧、抜粋を添付） ※ 必要に応じて画面コピー等の図表添付可
3	SIMCOMP http://www.genome.jp/tools/simcomp	公開	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。平成２０年度は平成１９年度の調査に基づき、検索速度の高速化を実現した。また光学異性体の違いを認識するように改良した。
4	SUBCOMP http://www.genome.jp/tools/subcomp	公開	化合物部分構造検索システム。ビットストリングを用いた高速な部分構造検索システム。これまで研究レベルで開発されてきたものを、バグの修正などをした上で、ウェブの検索システムとして公開している。平成２１年度に光学異性体の違いを認識するための改良、および、クエリを含む構造の検索を実装して公開する予定である。
5	E-zyme http://www.genome.jp/tools/e-zyme	公開	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抽出し、EC 番号との対応づけなどを行う。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た。また、データベース中の EC 番号と反応パターンの分布を用いて精度の向上を実現したものを、平成２１年１月１日に公開した。

（4）その他の成果物（（2）、（3）に該当しないもの）

特になし