

統合データベース探索フレームワークTogoDXの開発とその応用

○高月 照江¹⁾ 池田 秀也¹⁾ 井手 隆広²⁾ 大田 達郎¹⁾ 小野 浩雅¹⁾ 片山 俊明¹⁾ 川島 秀一¹⁾ 申 在紋¹⁾ 建石 由佳²⁾ 千葉 啓和¹⁾ 豊岡 理人^{2,5)} 内藤 雄樹¹⁾ 仲里 猛留^{1,3)} 信定 知江^{2,4)} 藤原 豊史¹⁾ 三橋 信孝¹⁾ 箕輪 真理¹⁾ 守屋 勇樹¹⁾ 山本 泰智¹⁾ 八塚 茂^{2,3)} 五斗 進¹⁾

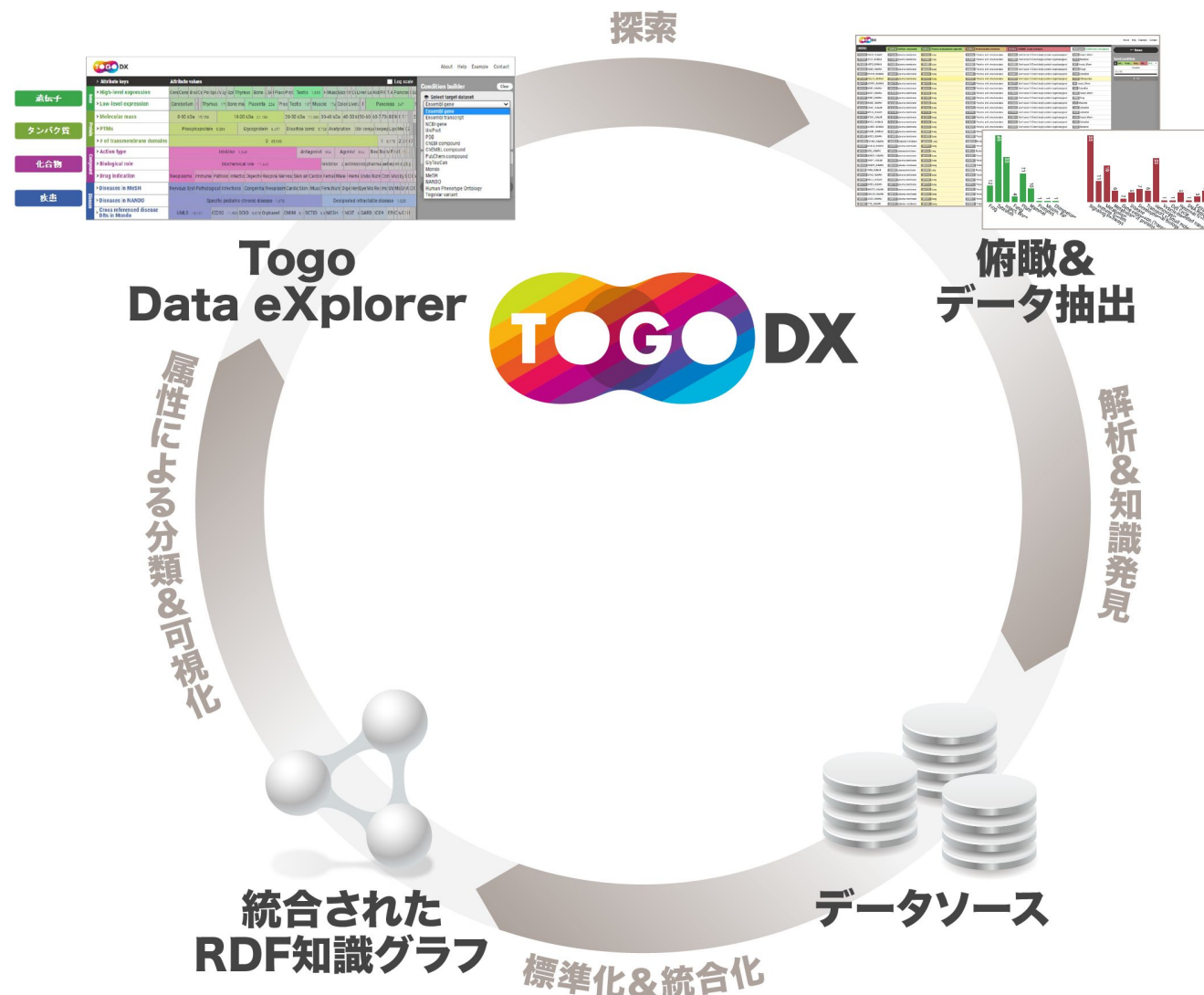
1)情報・システム研究機構 ライフサイエンス統合データベースセンター 2)科学技術振興機構 NBDC事業推進部 3)現:製品評価技術基盤機構 バイオテクノロジーセンター 4)現:理化学研究所 生命医科学研究センター 5)現:富山国際大学 現代社会学部

●TogoDX <https://togodx.dbcls.jp/human/>



TogoDX とは

TogoDX (Togo Data eXplorer)は、知識グラフ (Knowledge graph)化された様々なデータベースを統合的に探索するためのフレームワークです。異なるデータベースに由来する多数のデータセットを、同じものを示すIDの互換性や、異なるデータ間のつながりを整理したID関係に基づいて統合しています。TogoDXでは、データセットが持つ多様な属性(attribute)によって絞り込みを行い、データサイエンスに有用なデータを柔軟に抽出する新しい仕組みの提供を目指しています。



開発の経緯

NBDC/DBCLSではこれまでデータベース利用者に委ねられていた、データベースを統合的に利用する場合の

データベースごとに異なるインターフェース

データベース間を繋ぐリンク情報の欠如

データベースごとに異なる出力形式

と言った、問題を解決するために、**知識グラフ**による生命科学分野のデータベース統合を進めてきました。

知識グラフの技術基盤として**RDF** (Resource Description Framework)を用いることで、以下のような利点を得ることができます。

複数のデータセットが共通のURIで連結される

各データとそのつながりの意味が表現できる、すなわち、知識が表現できる

RDFを用いてデータベースを統合することにより、異なるデータセットをまたいだ情報の検索や取得が可能になります。

一方で、統合されたデータ量は非常に大きく複雑なものとなり、検索にはある程度専門的な知識が必要です。そこで、**統合されたデータを高速に処理し、ウェブブラウザ上の操作で探索することができるTogoDXを開発しました。**

TogoDX/Humanとは

TogoDX/HumanはTogoDXのフレームワークを用いて、ヒトに関するデータを統合的に探索するためのインターフェースを提供します。遺伝子、タンパク質、化合物、疾患などの情報を国内外のデータベースから収集・統合しています。本サイトは、[NBDC/DBCLS](#)が開発・提供しています。

TogoDX/Humanの特長

TogoDX/Humanでは、さまざまなデータベースから収集・統合したヒトに関する情報をワンストップで探索することができるアプリケーションです。現在、遺伝子、タンパク質、化合物、疾患などの多様なデータベースに由来する50以上の属性(例: 遺伝子の発現情報やタンパク質の局在情報、疾患別の分類など)を用いて、データの探索を行うことができます。



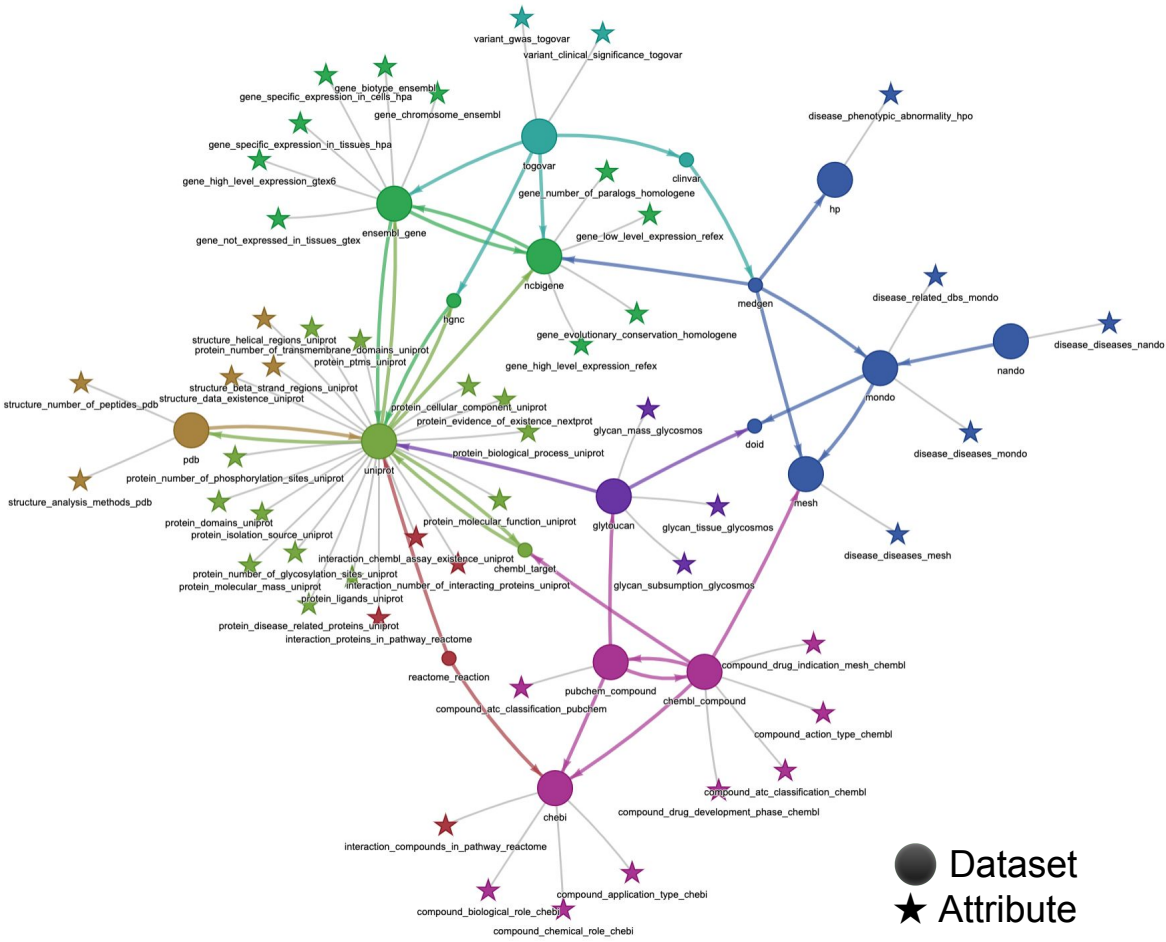
The screenshot displays the TogoDX/Human web application. The main area is a large table with columns for different biological attributes. The table is organized into sections: Gene biotype, Chromosome, # of exons, # of paralogs, Evolutionary divergence, High-level expression, Low-level expression, Expressed in tissues, Not expressed in tissues, Specifically expressed in tissues, Specifically expressed in cell types, Transcription factors, Protein domains, Cellular component, Biological process, Molecular function, Ligands, Molecular mass, PTMs, # of transmembrane domains, # of phosphorylation sites, # of glycosylation sites, Disease-related proteins, Tissues w/ expression reported, and Evidence of existence. Each section contains a list of values with counts. For example, under 'Gene biotype', 'protein coding' has 19,944 entries, 'lncRNA' has 16,889, and 'processed pseudogene' has 10. The right sidebar contains a 'Condition builder' section with a 'Select target dataset' dropdown, a 'Map your IDs' section with a text input and a 'Try' button, and an 'Add filters' section with a 'Map attributes' button and a 'View results' button.

<https://togodx.dbcls.jp/human/>

TogoDX/Human データの概要

統合されたデータベースとデータセット間のIDリンク関係

	カテゴリー	DB名	Attribute数
	Gene	Ensembl, HomoloGene, RefEx, GTEEx, HPA	10
	Protein	UniProt, neXtProt	13
	Protein structure	PDB, UniProt	5
	Interaction	Reactome, ChEMBL, UniProt	4
	Chemical compound	ChEMBL, ChEBI, PubChem	8
	Glycan	Glycosmos	3
	Disease	Mondo, MeSH, NANDO, HPO	5
	Variant	TogoVar (ClinVar, GWAS Catalog)	2



検索事例

The screenshot shows the TogoDX search interface. On the left, there are several filter panels: 'Attribute keys', 'Biological role', 'Drug indication', 'PubChem ATC classification', 'ChEMBL ATC classification', 'Molecular mass of glycans', 'Tissues', 'Subsumption', 'Diseases in Mondo', 'Diseases in MeSH', 'Diseases in Nando', 'Cross referenced disease DBs in Mondo', and 'Phenotypic abnormality'. The main area displays a list of results with columns for 'orig_dataset', 'orig_entry', 'orig_label', 'dest_dataset', 'dest_entry', and 'value'. The results are filtered by 'Abnormality of the eye' and 'Mendelian disease'.

●目に異常のある、遺伝性疾患とその遺伝子を検索する

選択条件としてMondoの遺伝性疾患、HPOの目の表現型異常を選び、遺伝子を選択項目として選んで実行をする



Ensembl gene	Human Phenotype Ontology	Phenotypic abnormality	Mondo	Diseases in Mondo
ENSG00000112619 PRPH2	0030642	Abnormality of the eye	0008210	Mendelian disease
			0011974	Mendelian disease
			0013137	Mendelian disease
			0024561	Mendelian disease
ENSG00000183638 RP1L1	0030636	Abnormality of the eye	0013316	Mendelian disease
			0032940	Mendelian disease
ENSG00000113139 CRYGS	0010922	Abnormality of the eye	0007284	Mendelian disease
			0012456	Mendelian disease

The screenshot shows the Ensembl gene detail page for ENSG00000112619. It displays the gene symbol PRPH2, its description as a protein coding gene, its location on chromosome 6, and its tissue specificity (Retina). It also shows cell specificity (Retinal cone cell, Retinal rod cell) and expression information.



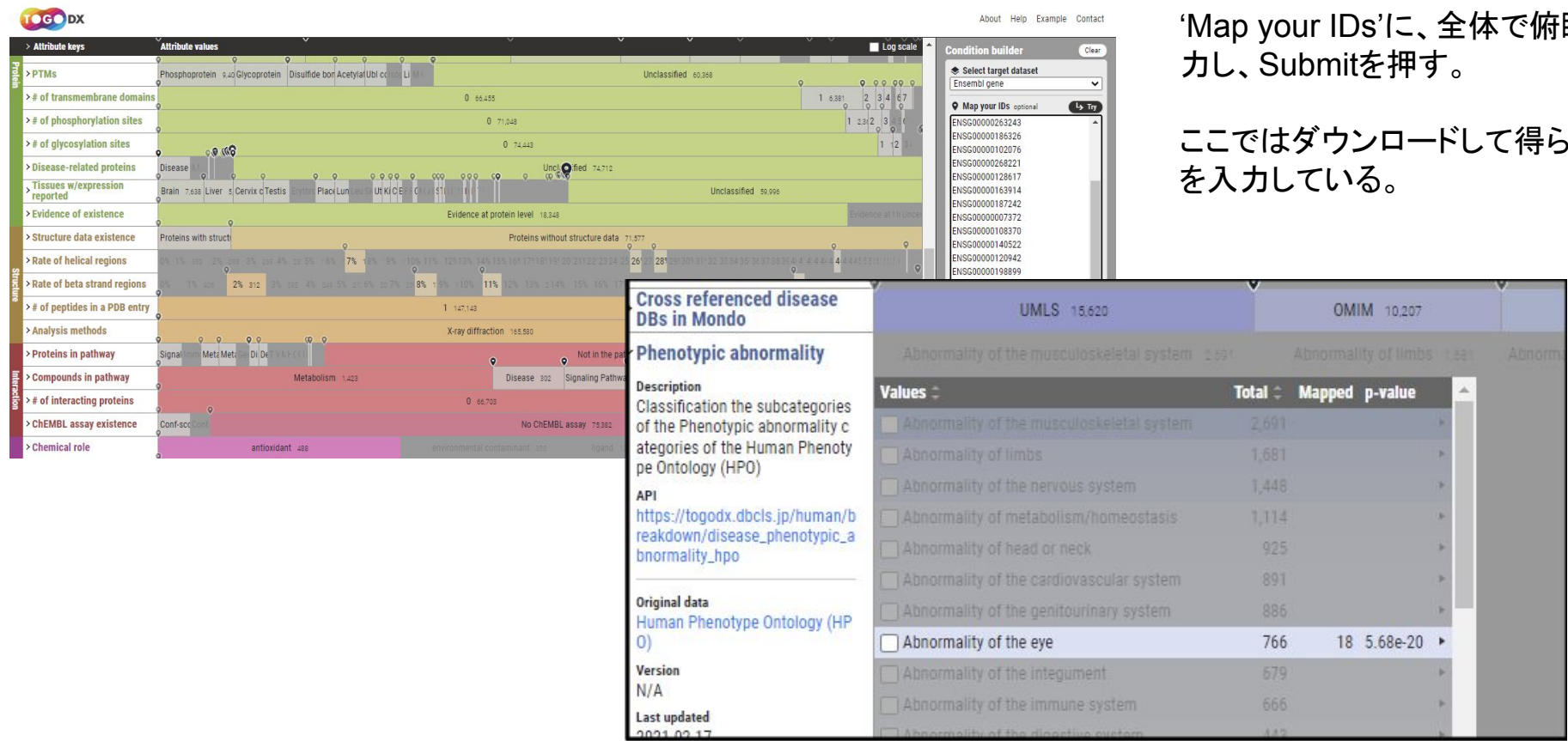
結果として、21個の遺伝子が該当し、その結果が表示される。項目のグレーのアイコンを押すと、それぞれのデータの詳細が表示される。結果はTSVでダウンロードできる。



検索事例

- 目に異常のある、遺伝性疾患とその遺伝子を検索する(2)

次にダウンロードした結果を利用して、データの俯瞰を行ってみる。



‘Map your IDs’に、全体で俯瞰してみたい IDを入力し、Submitを押す。

ここではダウンロードして得られた 21 個の遺伝子を入力している。

該当するデータがHitした場合は、詳細を見ると、Hit数と、p-valueが表示される。

まとめと展望

●まとめ

TogoDX は、知識グラフ化された様々なデータベースを統合的に探索するための汎用的なフレームワークです。統合されたデータを高速に処理し、ウェブブラウザ上の操作で探索する事が可能です。

TogoDX/Humanは、TogoDXのフレームワークを用いて、ヒトに関するデータを統合的に探索するためのインターフェースを提供しています。

●今後の展望

- ①機能向上を目指したUIの開発(カスタマイズ機能、オントロジービューア等)
- ②更なるデータセット(Attribute)の追加
- ③Human以外へのデータを適用したTogoDXの作成とその支援(TogoDX-TPP: ポスター番号10, 16)