

TogoID による生命科学データベースの横断利用

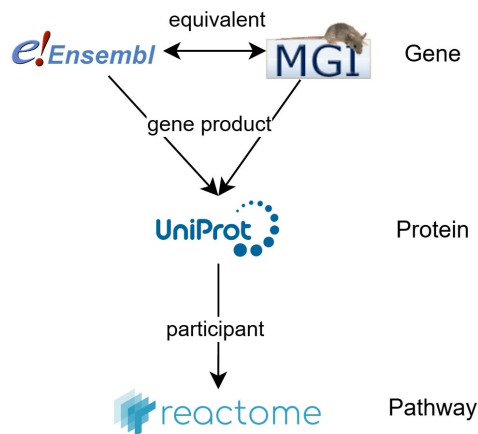
ライフサイエンス統合データベースセンター (DBCLS) 特任研究員
池田秀也

概要

- 生命科学データベース ID 変換ツール TogoID の紹介
 - データベースの ID 間の対応関係を…
 - ウェブ UI の操作で取得
 - REST API や RDF を使用してプログラムから取得

生命科学データベース利用における ID 変換の必要性

- 同じものを扱った複数のデータベースが存在
 - 遺伝子の場合、NCBI Gene, Ensembl, HGNC (ヒト), MGI (マウス), ZFIN (ゼブラフィッシュ), FlyBase (ショウジョウバエ), WormBase (センチュウ), TAIR (シロイヌナズナ), SGD (酵母), etc.
 - ID を入力とする解析ツールが、特定の DB の ID しか受け付けないことがある
- 必要な情報が、ID の由来するデータベースとは異なるデータベースにある
 - 手元の ID リストに関連する情報を他のデータベースから取得したい
 - 例: 遺伝子のコードするタンパク質が属するパスウェイについて知りたい
- 個々のレコードの ID 対応は各データベースに記載されているが、解析で得られた大規模な ID リストを一括で変換するのにはツールが必要



生命科学データベース ID 変換サービス TogoID

<https://togoid.dbcls.jp/>

ウェブ上で誰でも利用可能なサービスとして公開

- カテゴリーを問わずにデータベースを収録し、多様な意味的関係の取得をサポート
- データベース間の意味的な関係を整備し、ユーザーにわかりやすく提示
 - オントロジーを利用し、RDF 版も提供
- 3 つ以上のデータベースにまたがる多段階的な変換をサポート
- ID に付けられたラベルを取得する機能
- ラベルを入力として ID に変換する機能

The screenshot displays the TogoID web interface. At the top, the 'TOGO ID' logo is visible. Below it, a search bar contains the text 'ENSG00000136997'. To the right of the search bar are buttons for 'Submit', 'Input from text file', and 'Reset'. Below the search bar, there are examples: 'Refseq RNA', 'Ensembl gene', and 'UniProt'. The main content area is divided into two columns. The left column shows 'Ensembl gene' with a count of 4 and 'HGNC gene symbol' with a count of 4. The right column shows 'Affymetrix probeset' with a count of 11, 'ChEMBL target' with a count of 0, 'Ensembl protein' with a count of 24, 'Ensembl transcript' with a count of 30, 'HGNC' with a count of 4, 'MGI gene' with a count of 0, and 'NCBI Gene' with a count of 4. A central column lists relationships: 'is target of', 'has gene product', 'is transcribed to', 'is nearly equivalent to', and 'is nearly equivalent to'. A blue dot is positioned next to the 'is target of' relationship. On the far right, a 'History' panel shows a list of datasets last updated on 2025-06-10, including 'miRBase mature' and 'RNAcentral, LncBook, Rfam, FlyBase transcript, FlyBase protein, ZFIN transcript'.

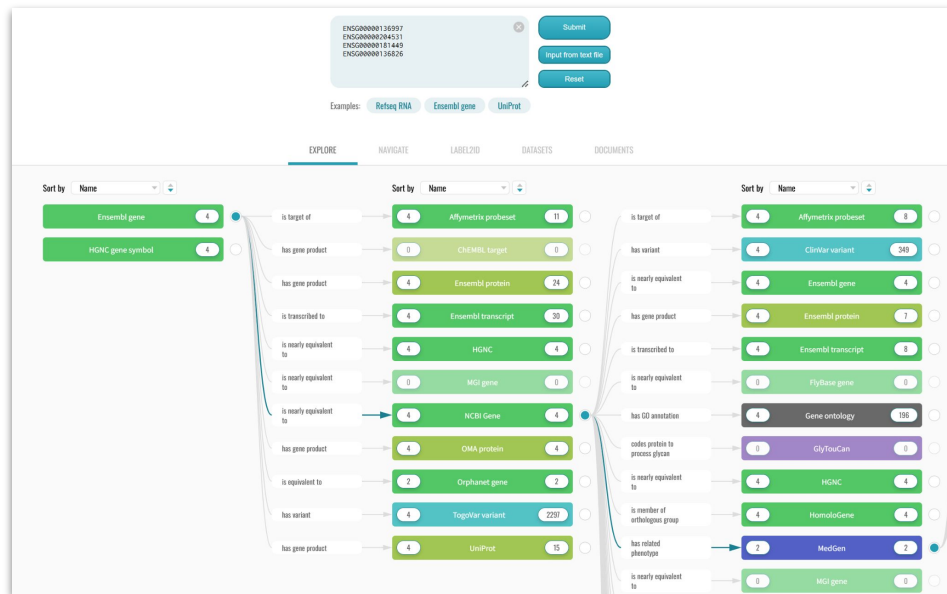
対象データベース

- 生命科学に関わる幅広いカテゴリーのデータベースをカバー
 - 遺伝子・タンパク質・化合物・パスウェイ・疾患...
 - 2025 年 6 月現在 76 データベース
 - 各データベースが記述している、そのデータベースのレコードと外部データベースのレコードの ID 関係を抽出し収載

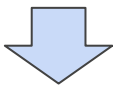
Category	Datasets
Gene	NCBI Gene, Ensembl gene, RefSeq genomic, GenBank/ENA/DBJ, HGNC, Consensus CDS, FlyBase gene, LncBook gene, LRG, MGD gene, MGI gene, OMIM gene, Orphanet gene, RGD, SGD, TAIR, VGNC, WormBase gene, Xenbase gene, ZFIN gene
Transcript	Ensembl transcript, RefSeq RNA, miRBase, miRBase mature, FlyBase transcript, RNCentral, Rfam, ZFIN transcript
Protein	UniProt, RefSeq protein, ChEMBL target, Ensembl protein, OMA protein, FlyBase protein
Structure	PDB
Domain	InterPro, Pfam, PROSITE, SMART
Probe	Affymetrix probeset
Genome	Assembly INSDC, Assembly RefSeq
Proteome	UniProt proteome
Variant	ClinVar variant, dbSNP, MGI allele, TogoVar variant
Genotype	MGI genotype
Reaction	Reactome reaction, Rhea
Pathway	Pathbank, PubChem pathway, Reactome pathway, WikiPathways
Interaction	IntAct
Ortholog	COG, HomoloGene, OMA group
Phenotype	Disease ontology, HP Phenotype, MedDRA, MedGen, MeSH, MONDO, MP, NANDO, NCIt disease, OMIM phenotype, Orphanet phenotype
Classification	HP Inheritance
Organism	MGD organism, Taxonomy
Cell Line	Cellosaurus
Anatomy	NCIt tissue, UBERON
Function	Enzyme nomenclature, Gene ontology
Annotation Rule	PROSITE ProRule
Compound	ChEBI compound, ChEMBL compound, PubChem compound, PubChem substance, HMDB, InChIKey, IUPHAR ligand, PDB CCD
Lipid	LIPID MAPS, SwissLipids
Glycan	GlyTouCan, GlycoMotif
Drug	ATC classification, DrugBank
Experiment	GEA, SRA experiment
Sequence Run	SRA run
Submission	SRA accession
Project	BioProject, BioProject umbrella, GEO series, INSDC master, JGA Dataset, JGA Study, NBDC Human Database, SRA project
Analysis	SRA analysis
Sample	BioSample, GEO sample, SRA sample
Literature	PMC, PubMed

EXPLORE モード: 探索的な ID 変換

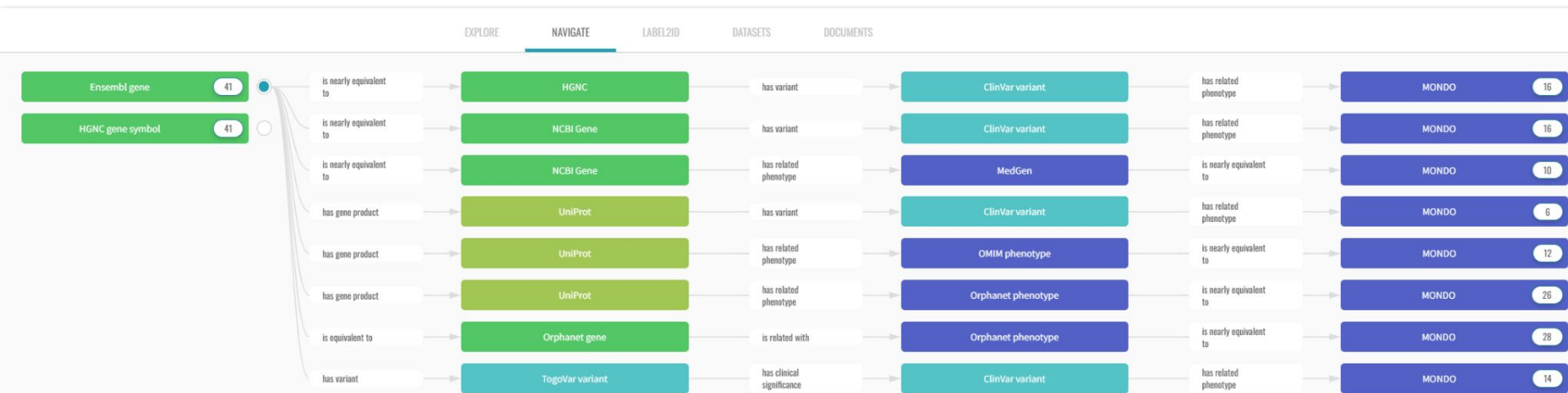
入力した ID のデータセットから直接つながる ID に変換し、そこから更に他のデータセットに変換する、という多段階的な変換が可能



NAVIGATE モード: 始点と終点を指定した ID 変換



- 始点と終点をつなぐ変換パスを検索
- データセット間の関係を確認しながら、目的に合ったパスを選択する



様々な利用例

- OMIM (疾患) → MeSH (疾患)
 - [Deep multiple instance learning on heterogeneous graph for drug – disease association prediction](#)
 - drug-disease association に関する複数の既存のベンチマークデータセットを使って評価を行う際に、データセット間の ID を揃えるのに TogoID を使った
 - 疾患データベースやオントロジーは数多いが、ID 変換を簡便に行えるツールは意外とない
- ChEMBL Compound (化合物) → PDB CCD (化合物) → PDB (タンパク質立体構造)
 - 中に低分子化合物が含まれる、タンパク質(複合体)立体構造
- BioSample (実験サンプル) ↔ SRA experiment (実験メタデータ) ↔ SRA Run (実験データ)
- GlycoMotif (糖鎖モチーフ) → GlyTouCan (糖鎖) → UniProt (タンパク質) → ClinVar (疾患関連バリエーション) → MONDO (疾患)
 - 「「モチーフ構造を持つ糖鎖」を合成する酵素」が持つ多型」が関与する疾患
 - 糖鎖 DB では酵素の基本情報は記述しているが、その多型情報などは他の DB を参照する必要がある。各 DB が直接記述していない関係も、TogoID 上で簡単につなぐことができる

ラベル表示機能

- 遺伝子シンボルや疾患名などを表示可能
 - バックエンドに用いている仕組み ([grasp](#)) はラベル以外の情報にも拡張可能
 - 遺伝子タイプ (protein-coding, long non-coding, pseudo, etc.) など様々なアノテーションも表示できるようにし、かつフィルタリングに使えるようにする、といった仕組みを近日公開予定

Preview

NCBI Gene		MedGen	
ID		ID	
☒ Show Labels		☒ Show Labels	
4609	MYC	C0006413	Burkitt lymphoma
6657	SOX2	C1859773	Anophthalmia/microphthalmia-esophageal atresia syndrome

ラベルを ID に変換する機能

- 遺伝子の場合は目的の生物種を指定し、変換先の ID を絞り込む
- 疾患名や化合物名の場合は曖昧一致の許容度を指定
- Convert IDs を押すと、変換結果の ID が入力窓に移動し、TogoID のインターフェースで他の ID に変換可能

The screenshot displays the TogoID web application interface. At the top, there's a header with navigation links: EXPLORE, NAVIGATE, LABEL2ID (active), DATASETS, and DOCUMENTS. Below the header, there's a search bar with a dropdown menu set to 'NCBI Gene'. To the right of the search bar, there's a 'Species' dropdown menu set to 'Homo sapiens (ID: 9606, human.)' and an 'OR' button followed by a text input field containing '9606'. Below these, there's a 'Select label types' section with three checkboxes: 'NCBI Gene symbol' (checked), 'Synonym', and 'Official symbol'. An 'EXECUTE' button is located to the right of these checkboxes. Below the 'EXECUTE' button, there's a 'Report' section with two radio buttons: 'Matched labels' (selected) and 'Include unmatched labels'. Below the 'Report' section, there's an 'Action' section with four buttons: 'Convert IDs', 'Copy to Clipboard', 'Download as CSV', and 'Download as TSV'. Below the 'Action' section, there's a 'Showing' section with a table of results. The table has four columns: 'Input', 'Match type', 'Symbol', and 'ID'. The 'ID' column has a dropdown menu set to 'ID'. The table contains five rows of results.

Input	Match type	Symbol	ID
FUT8	NCBI Gene symbol	FUT8	2530
ATP6	NCBI Gene symbol	ATP6	4508
OCT4	Synonym	POU5F1	5460
OCT4	Synonym	POU5F1P3	642559
OCT4	Synonym	POU5F1P4	645682

REST API

<https://togoid.dbcls.jp/apidoc/>

- プログラムから呼び出して使用
- ID を入力するようなアプリケーションを開発するときなどに、ID 変換部分をTogoID に担わせることで、異なる DB の ID を入力として受け付けられるように

The screenshot shows the 'Report' section of the TogoID web interface. It includes radio buttons for 'All converted IDs', 'Source and target IDs', 'Target IDs', and 'All including unconverted IDs' (which is selected). There are also radio buttons for 'Expanded' (selected) and 'Compact' format. Below these are buttons for 'Download as CSV', 'Download as TSV', 'Copy to clipboard', 'Copy API URL' (highlighted with a red box), and 'Copy CURL'. A 'Preview' section below shows a table with three columns: 'Ensembl gene', 'NCBI gene', and 'MedGen'. Each column has a dropdown menu for 'ID' and a 'Show Labels' toggle. The table contains four rows of data.

Ensembl gene	NCBI gene	MedGen
ENSG00000136997	4609	C0006413
ENSG00000204531	5460	
ENSG00000181449	6557	C1859773
ENSG00000136826	9314	

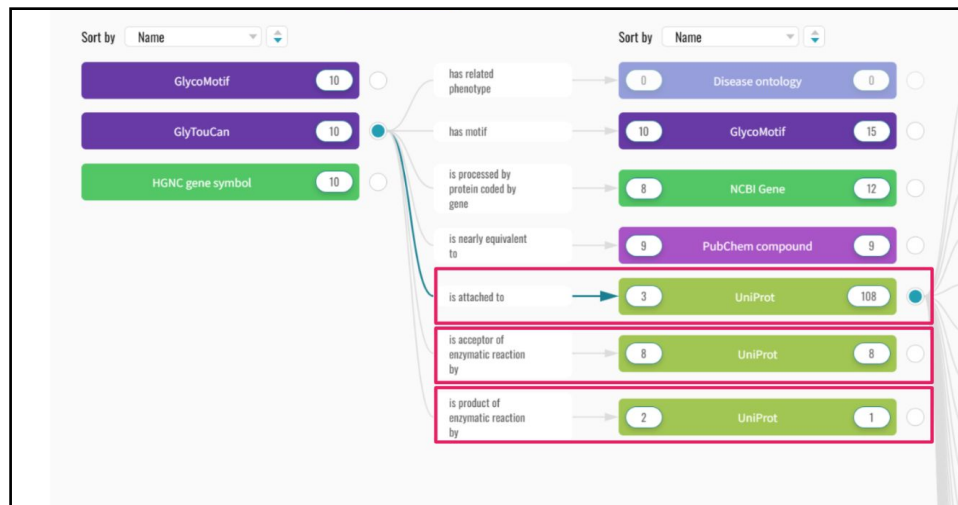


UI 上で変換を実行してから
Copy API URL を使うのが楽

https://api.togoid.dbcls.jp/convert?route=ensembl_gene%2CTIO_000001%2Cncbigene%2CTIO_000011%2Cmedgen&ids=ENSG00000136997%2CENSG00000204531%2CENSG00000181449%2CENSG00000136826&report=full&format=json

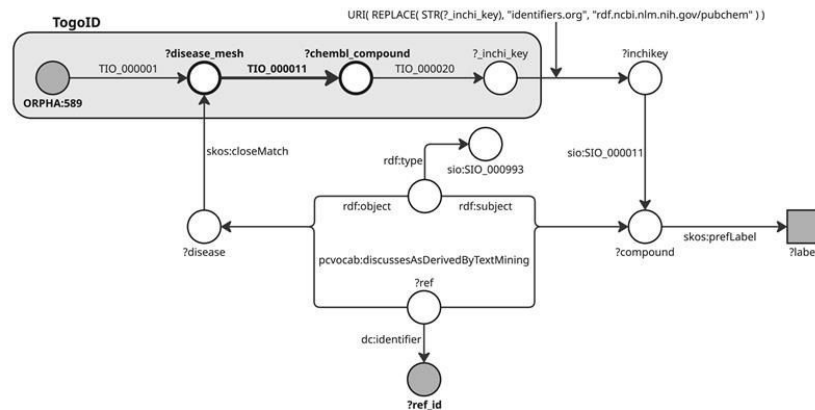
TogoID Ontology

- TogoID では幅広いカテゴリーの ID 間の変換に対応
 - → ID 間がどのような意味で関係しているのかが自明でない
- リンクの生物学的な意味を明示するためにオントロジーを作成
 - GUI 上でデータセット間をつなぐ線に意味を表示
 - 同一のデータセットペア間に複数種類の意味的關係を扱える
 - 糖鎖がタンパク質を修飾する
 - 糖鎖が酵素反応の基質・生成物になる



RDF 版

- RDF Portal で提供
 - <https://rdfportal.org/primary/sparql>
- 利用例
 - <https://is.gd/oHEBj3>
 - ある疾患 [ORPHA:589](#) (重症筋無力症、Myasthenia gravis) と、ChEMBL で関連が記述されている化合物を、PubChem の共起関係データにおいて関連を記述している文献数の多い順に取得
 - ID 対応の部分は TogoID で整理されたデータを用いている



DBCLS 山本さんよりご提供

ID ペア取得プログラムの開発と公開

TogolD-config <https://github.com/dbcls/togoid-config>

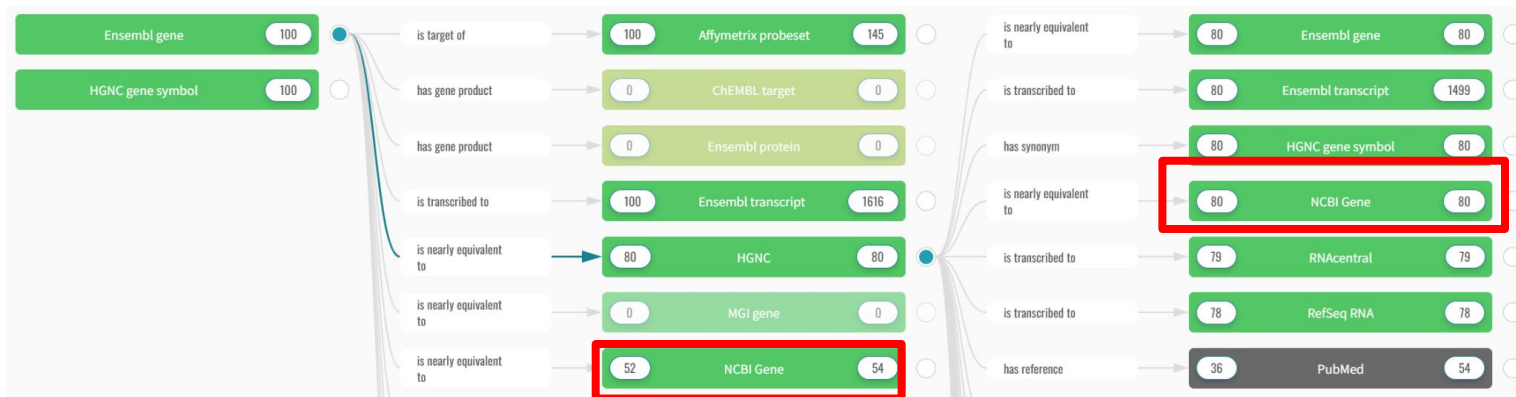
- 変換元と変換先のペア毎に、ID の対応関係を抽出し 2 列のタブ区切り形式で出力するプログラムを作成
 - オリジナルのデータベースが提供しているファイルのパーズなど
- データセットごとに、ID の正規表現などを記述
 - 正規表現は、入力された ID が由来するデータセットを推測するのに使用。
 - ID には表記揺れがあるので、実用される表記法をカバーするように正規表現を記述。
 - 例えば OMIM の場合、同じレコードに対する ID の表記に
"107741", "OMIM:107741", "MIM:107741"
がある。正規表現は `^(O?MIM:)?\d{6}$`
- 作成したプログラムを用いて週に一回の定期更新

D5VRB9	6IAD
D5VRB9	6IA8
D5VRB9	6IAZ
A9A677	3S0X
O52980	1IX5
P43410	6HF7

uniprot-pdb の抜粋

ID 変換において注意すべき点

- 外部データベースのレコードを「同一のもの」として捉える根拠はデータベースごとに異なる
 - 同じものが得られると期待される変換であっても、経由するデータベース次第で結果が異なることがある
 - ↓ ヒトの lncRNA の Ensembl gene エントリーの ID 100 件を NCBI Gene に変換した例。直接つながるのは 52 件だが、HGNC を経由すると 80 件取れる。
 - ID 変換を行った結果、情報の欠損が起こり得るということは認識しておくべき



まとめ

- ID 変換という、データベースの統合利用における基本的な操作を様々な形でサポート
 - 論文中の遺伝子 ID リストを他のデータベースの ID のリストに変換して再解析する
 - 関連情報を調査したい ID リストを入力し、関連先のデータベースの ID のリストを UI 上で出して個々に確認する
 - プログラムの中で、ある ID に関連する他の ID を取得するために API を利用する
 - データベース開発の中で、ID 変換の部分を TogoID の API や RDF に担わせる