

# 5. TogoID: データベース統合の基盤となる ID変換サービス

○池田秀也<sup>1</sup>、千葉啓和<sup>1</sup>、藤原豊史<sup>1</sup>、五斗 進<sup>1</sup>、井手隆広<sup>2</sup>、川島秀一<sup>1</sup>、箕輪真理<sup>1</sup>、  
三橋信孝<sup>1</sup>、守屋勇樹<sup>1</sup>、内藤雄樹<sup>1</sup>、仲里猛留<sup>1,3</sup>、信定知江<sup>2,4</sup>、大田達郎<sup>1</sup>、小野浩雅<sup>1</sup>、申在紋<sup>1</sup>、  
高月照江<sup>1</sup>、建石由佳<sup>2</sup>、豊岡理人<sup>2,5</sup>、山本泰智<sup>1</sup>、八塚茂<sup>2,3</sup>、片山俊明<sup>1</sup>

1 情報システム研究機構データベース共同利用基盤施設ライフサイエンス統合データベースセンター (DBCLS)

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



# ID変換の必要性

-----  
バイオインフォマティクスで様々なDBを活用するにはデータベースID間のリンクが重要

- 等価なものに付けられたID間の変換
  - 例: NCBI Gene ID ↔ Ensembl ID
  - 使いたい解析ツールが、手元のIDを受け付けてくれない場合など
- 関連する情報の取得
  - バリアント → 遺伝子
  - 遺伝子 → トランスクリプト
  - トランスクリプト → タンパク質
  - タンパク質 → 立体構造
  - 立体構造 → 相互作用
  - 相互作用 → 化合物・医薬品
  - 化合物・医薬品 → パスウェイ
  - パスウェイ → 疾患

# 既存のID変換サービス

---

データベースID間のリンク情報を提供する既存のサービスの例

- 国内: LinkDB (ゲノムネット), Biodb.jp
- 海外: BioMart, UniProt ID mapping, Ensembl, Bio2RDF

既存のリンク情報の課題

- 対象としているデータベースのカバレッジが限られる
- 各データベースの毎年・毎月・毎日など更新への追従
- 対話的に操作するUIと、プログラムから自動化して利用するAPIの両方が欲しい

類似のものとしてIDの転送サービスがあるが

- ページを開いてみるまで転送先は不明: PURL, Identifiers.org
  - OK: 転送ルールだけ記述しておけばよいので維持管理は容易
  - NG: 事前に転送先のIDを知っておくには、データとして維持管理しておく必要がある

# TogoIDで実現したいこと

----

データベースのカバレッジを確保

- ライフサイエンス統合データベースセンターにおけるデータ統合のハブとして
- LINCなど生命医科学ドメインのニーズに応じて対応データベースを拡張する

ウェブ上で対話的に操作して変換し結果をダウンロード

- 始点となるIDから探索的に接続先のデータベースをたどる
- 始点および終点となるデータベースを指定して経路を探索する

プログラムによる自動処理を実現

- 上記と同じ機能をウェブサービス(API)としても提供

安定的なサービスの提供と定期的な更新

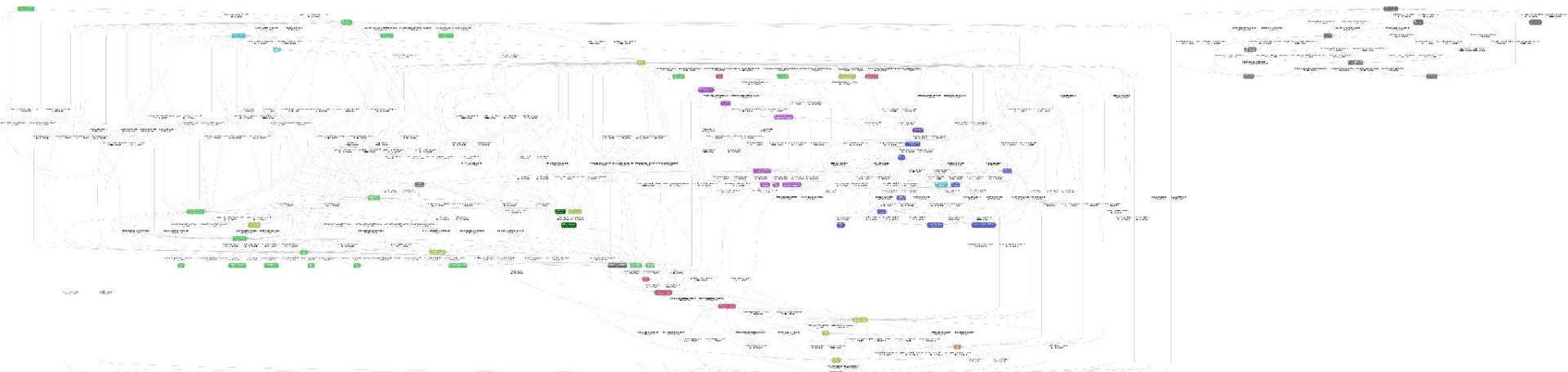
- サービスをクラウドで提供することでダウンタイムを解消
- データベース毎の更新頻度に合わせたアップデートを自動化 (TogoID-config)

# TogoIDの対象データベース選定

- 遺伝子・タンパク質・化合物・パスウェイ・疾患など、対象とするDBをリストアップ
- NBDCデータベースカタログとの対応付け
- データ取得元・データ形式・更新頻度・ライセンスなどを調査
- 各DBからID変換可能なDBを調査
- ID体系（正規表現パターンなど）の調査

The screenshot shows a Google Docs spreadsheet with a large table of database information. The table has multiple columns, including database names, IDs, and various attributes. The rows are color-coded in a grid pattern, likely representing different categories or statuses of the databases. The spreadsheet is titled 'TogoID' and is being viewed in a browser window.

# TogoIDによるリンク情報の集約と管理



2022/9 現在

- 65 データベース
- 162 データベースペア

# Togoid ontology の作成

- Togoid では同じ実体を指すものの間の ID 変換(例: NCBI gene <-> Ensembl Gene)だけでなく、何らかの意味でつながる ID 間の変換を幅広く対応
- 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
  - UI 上で表示
  - RDF 版でも predicate として使用
- <https://togoid.dbcls.jp/ontology>

# TogoIDのウェブインターフェイス



63962027  
1046  
63961921  
102060930  
100472895  
497025  
84148

Submit  
Input from text file  
Reset

Examples: Refseq RNA   Ensembl gene   Uniprot

EXPLORE   NAVIGATE   DATASETS   DOCUMENTS

The screenshot shows a network of ontologies. The left column lists various ontologies with their counts: ChEBI compound (10), ClinVar variant (10), HGNC gene symbol (10), HomoloGene (10), MISO organism (10), MedDRA (10), NCBI gene (10), PubChem compound (10), PubChem substance (10), PubMed (10), RGD (8), Taxonomy (10), HGNC (4), and PDB (1). The middle column shows relationships: 'is target of' (2), 'is nearly equivalent to' (1), 'is nearly equivalent to' (4), 'is member of orthologous group' (4), 'has related disease' (2), 'has gene product' (5), 'is transcribed to' (1), 'has GO annotation' (5), 'is transcribed to' (1), 'is nearly equivalent to' (4), 'is nearly equivalent to' (2), 'has gene product' (1), 'is transcribed to' (7), and 'is gene of organism' (10). The right column lists more ontologies: Affymetrix probe set (3), Ensembl gene (5), HGNC (4), HomoloGene (4), MedGen (4), Ensembl protein (8), Ensembl transcript (3), Gene ontology (118), miRBase (1), OMIM gene (4), RefSeq genomic (2), RefSeq protein (10), RefSeq RNA (15), and Taxonomy (6). Relationships on the right include 'is related with' (4), 'is nearly equivalent to' (8), 'is nearly equivalent to' (1), 'is nearly equivalent to' (4), 'is nearly equivalent to' (4), and 'is nearly equivalent to' (0).

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

- IDリストを入力（もしくはファイルアップロード）
- 自動判定されるDBを確認して選択
- 変換先のDBを選択
- 必要なら数ステップ先の変換先DBまで選択
- : オントロジーで定義された、リンクの意味





# TogoIDのウェブインターフェイス

The screenshot shows the TogoID web interface. At the top, the 'TOGO ID' logo is visible. Below it, a 'Results' window displays a query route: 'NCBI gene' (ID: 10) -> 'has related disease' (ID: 2) -> 'MedGen' (ID: 4) -> 'is nearly equivalent to' (ID: 4) -> 'MONDO' (ID: 4). The 'has related disease' and 'is nearly equivalent to' relationships are highlighted with red boxes. Below the route, there are options for 'Report' (All converted IDs, Source and target IDs, Target IDs, All including unconverted IDs) and 'Action' (Download as CSV, Download as TSV, Copy to clipboard, Copy API URL). A 'Preview' section shows a table of results:

NCBI gene IDs	MedGen IDs	MONDO IDs
84148	C5436525	MONDO_0033547
88	C2677338	MONDO_0012808
88	C5231445	MONDO_0032852
88	C5203349	MONDO_0032853

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

- IDリストを入力（もしくはファイルアップロード）
- 自動判定されるDBを確認して選択
- 変換先のDBを選択
- 必要なら数ステップ先の変換先DBまで選択
- : オントロジーで定義された、リンクの意味
- プレビューして変換状況を確認
- 問題なければ変換表をダウンロード

# TogoIDのAPIによるプログラムからの自動変換処理

## Togo ID API <sup>2.0.0</sup>

[ Base URL: api.togoid.dbcls.jp ]

Schemes

HTTPS

### convert

GET /convert Convert IDs

Parameters Try it out

Name	Description
<b>ids</b> <sup>required</sup>	A comma-separated list of source IDs.
string (query)	
<input type="text" value="ids"/>	
<b>route</b> <sup>required</sup>	A comma-separated list of datasets, starts with the source dataset and ends with the target dataset.
string (query)	
<input type="text" value="route"/>	
<b>report</b>	The output type can be selected from the following: 'target' (includes target IDs only), 'pair' (includes source and target IDs), 'all' (all converted IDs including source, target, and intermediate IDs of the route), or 'full' (all IDs including unconverted IDs).
string (query)	
Available values: all, pair, target, full	
Default value: target	
<input type="text" value="target"/>	

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

- ?ids=5460,6657,9314,4609
  - 変換元のIDリストをカンマ区切りで渡す
- &route=ncbigene,ensembl\_gene
  - 変換ルートをカンマ区切りのデータベース名で渡す
- &format=json
  - 取得するデータ形式を指定 (csv, tsv, json)
- &include=target
  - 変換先IDだけ (target)
  - 変換元IDと変換先ID (pair)
  - 中間の変換ルートすべてのID (all)
- &offset=0&limit=10000
  - 大量に取得する場合のオフセット・リミット値

詳細: [TogoID API 2.0.0](#)

# TogoIDを構築して分かった課題

## 元々のデータベースに内在する問題

- IDの表記ゆれが激しい
  - PDB: 1G0M, 1g0m
  - Gene ontology: GO:0019907, GO\_0019907
  - Orphanet: ORPHA:101078, ORD0:101078, Orphanet:101078
- 1つのDBに複数のID体系が混在 (ゲノム, 遺伝子, トランスクリプト, タンパク質…)
  - Ensembl: ENSG00000186283, ENST00000638000, ENSP00000365411
  - RefSeq: NG\_004671, NM\_001199636, NP\_001171968

TogoIDでは名前空間を分けて管理



実用される表記法をなるべく拾うような正規表現で対応、DATABASESタブに例示

# TogoIDを構築して分かった課題

---

## 変換後のIDが発散する問題

- 1対多、多対多（変換後にその先の変換を続けると対応数が爆発する）
  - 遺伝子→Gene ontologyによる分類→タンパク質
  - タンパク質→Pfamなどの機能ドメイン→立体構造

生物種で絞り込む、などが考えられるが未対応

## どこまでID変換でやるのか問題

- ID変換の意味（セントラルドグマ、相互作用ネットワーク、関連文献）
  - ウェブページのクリックで辿れるもの全部OKというわけではない…

## リンクの意味を標準化する必要性（逆向きの変換も含め）

- セマンティック・ウェブ技術によるリンク関係のオントロジー整備
- 同じペアでも関係が同じとは限らない
  - 例) 糖鎖-タンパク質 (糖鎖を代謝する酵素？糖鎖で修飾されるタンパク質？)

# Togoidの対象データベースの追加

変換元DBと変換先DBのペア毎に、IDの対応関係を抽出するプログラムを作成 ← TSVを生成

Togoid-config

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

11099	ENSG00000071794
11114	ENSG00000126012
11115	ENSG00000012817
:	

内容

- Rakefile 自動更新手順 (← 前処理が必要なら追記する)
- bin/ 各種取得・変換スクリプト群
- config/ db1-db2ごとの変換規則群
  - dataset.yaml データベース一覧 (← まだ載ってなければDBを追記する)
  - db1-db2/config.yaml 更新手順 ← 上記プログラムの実行方法を記載する
- input/ 共通の前処理入力データ置き場
- output/ 生成される出力IDペア置き場
  - tsv/db1-db2.tsv タブ区切りファイル
  - ttl/db1-db2.ttl RDF版ファイル

# Publication




— — —

*Bioinformatics*, 38(17), 2022, 4194–4199  
<https://doi.org/10.1093/bioinformatics/btac491>  
Advance Access Publication Date: 8 July 2022  
Original Paper

OXFORD

Databases and ontologies

## TogID: an exploratory ID converter to bridge biological datasets

Shuya Ikeda <sup>†</sup>, Hiromasa Ono <sup>†</sup>, Tazro Ohta , Hirokazu Chiba , Yuki Naito ,  
Yuki Moriya , Shuichi Kawashima , Yasunori Yamamoto , Shinobu Okamoto,  
Susumu Goto  and Toshiaki Katayama \*

Database Center for Life Science, Joint Support-Center for Data Science Research, Research Organization of Information and Systems, University of Tokyo Kashiwanoha-campus Station Satellite 6F, Kashiwa, Chiba 277-0871, Japan

\*To whom correspondence should be addressed.

<sup>†</sup>The authors wish it to be known that these authors contributed equally.

Associate Editor: Peter Robinson

Received on April 7, 2022; revised on June 8, 2022; editorial decision on July 5, 2022; accepted on July 7, 2022

<https://doi.org/10.1093/bioinformatics/btac491>

# ご意見お待ちしております！

----

- ご意見・ご要望はこちらから <https://dbcls.rois.ac.jp/contact.html>
- ウェブUIの機能面
  - ここが使いにくい/使いやすい
- 対象データベース
  - このIDを変換したい