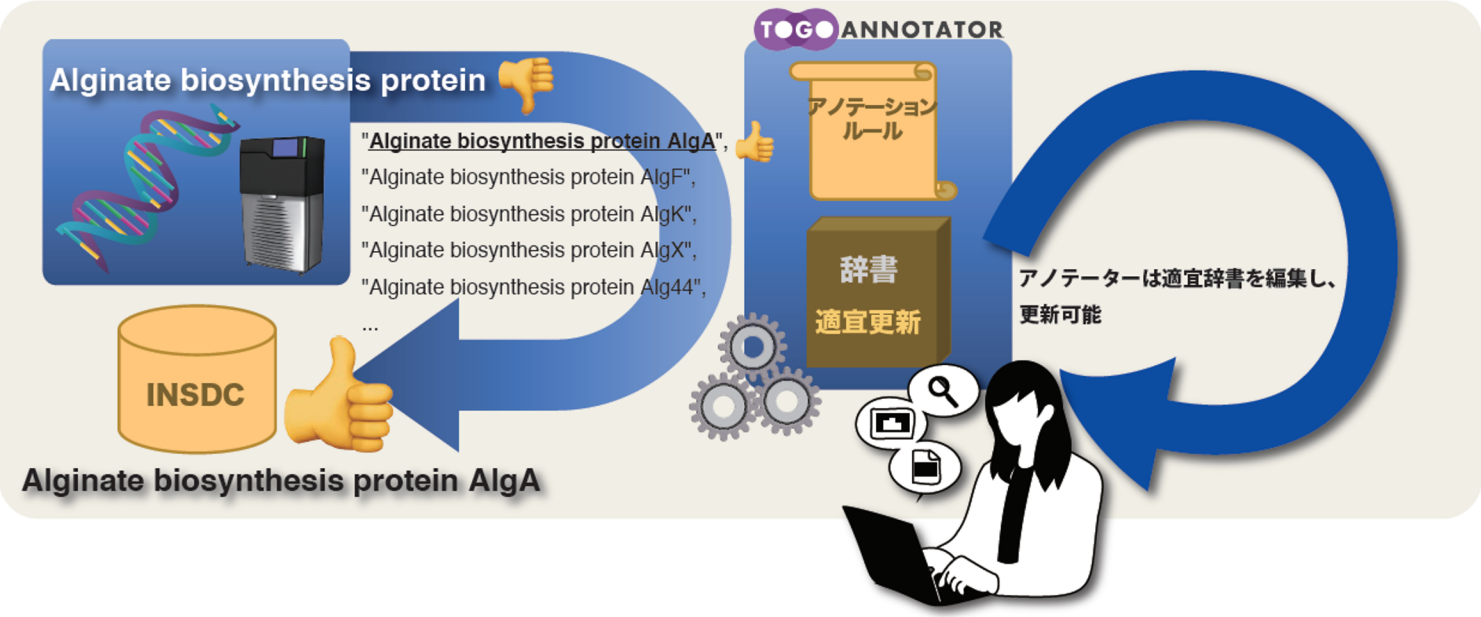
国際タンパク質命名ガイドラインに準拠したタンパク質名提案システムへの取り組み

○藤澤貴智[1]、李 慶範[1]、山本泰智[2]

1) 国立遺伝学研究所 生命情報・DDBJセンター 2) 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(DBCLS)

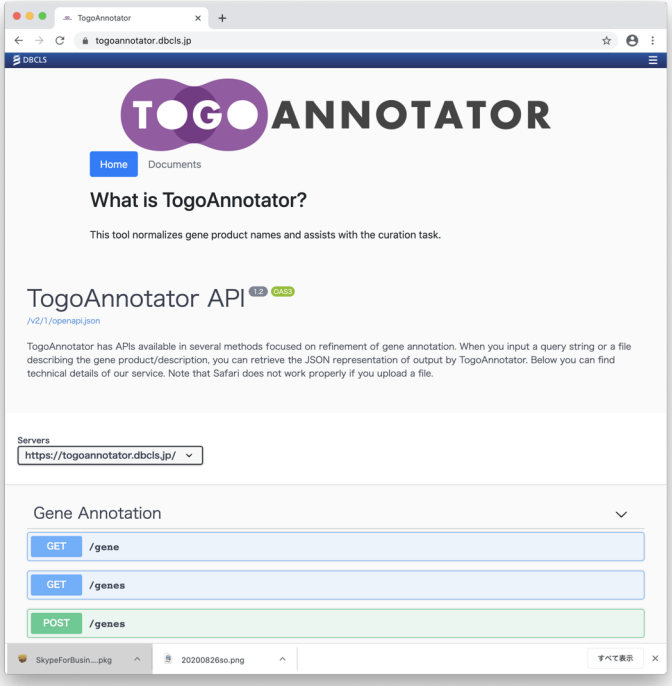
TogoAnnotatorとは？

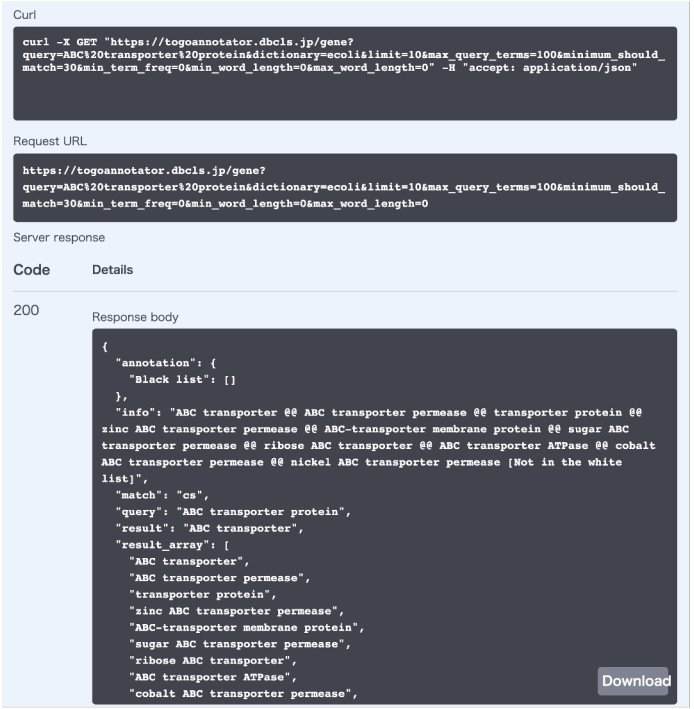
INSDCに登録する遺伝子産物名表記の正規化作業を効率化するツールです。



TogoAnnotator API

TogoAnnotatorはウェブサービスとして公開しており、APIの３つのメソッドが利用可能です。





URL: https://togoannotator.dbcls.jp

国際タンパク質命名ガイドラインとは？

国際タンパク質命名ガイドラインは、 European Bioinformatics Institute (EMBL-EBI)、National Center for Biotechnology Information (NCBI)、Protein Information Resource (PIR) および Swiss Institute for Bioinformatics (SIB)によって共同で作成され、タンパク質に名前を付けたい人が、データベース間でのタンパク質命名の一貫性を高め、データ検索を支援し、コミュニケーションを改善するために使用することを目的としています。
一貫したタンパク質の命名法は、コミュニケーションや文献検索や配列エントリーの検索に不可欠です。良いタンパク質名とは、他の種からのオーソログに帰属させることができ、適切な場合には公式の遺伝子命名法に従うことができるユニークかつ明確なタンパク質名です。名前をタンパク質配列に関連付けるプロセスには、配列からの機能同定・予測、名前の選択、フォーマットの適用など、さまざまな要素があります。このガイドラインでは、名前の選択と汎用フォーマットに関するガイドラインを示しており、配列からの機能同定・予測に使用される方法に関するベストプラクティスは対象とされていません。

International Protein Nomenclature Guidelines
https://www.ncbi.nlm.nih.gov/genome/doc/internatprot_nomenguide/

目的と手法

TogoAnnotatorが最適な遺伝子産物名を提案するために、以下の２つのプロセスにおいて国際タンパク質命名ガイドライン適用を計画しました。

1. 入力クエリに対して国際タンパク質命名ガイドライン項目に準拠しているかの情報を提供
2. 類似マッチの結果に対して、ガイドラインに準拠している辞書データが検索上位となるような最適化

各ガイドライン項目詳細について定義したチェックリストを作成

ガイドライン項目詳細の定義

id	code	message	bad example	good example	implementation
PN001	2-A	Use American spelling, not British spelling	uncharacterised protein catalyse	characterise	TRUE
PN002	2-A	Use protein names ending in 'in' (not 'ine')	maurocalcine	maurocalcin	TRUE
PN003	2-A	Avoid diacritics such as accents, umlauts etc.	protein spätzle 5	protein spaetzle 5	TRUE
PN004	2-A	Avoid pluralization for names based on domain and repeat content	ankyrin repeats-containing protein	ankyrin repeat-containing protein	TRUE
PN005	2-A	Avoid common words	protein IMPACT		TRUE
PN006	2-A	Avoid duplication			FALSE
PN007	2-B	Avoid using an abbreviation as the complete name	ACP	acyl carrier protein	TRUE
PN008	2-B	An abbreviation may be part of a protein name	(3R)-hydroxymyristoyl-ACP dehydratase	(3R)-hydroxymyristoyl-ACP dehydratase	FALSE
PN009	2-B	Protein name based on a protein symbol (PS) or gene symbol (GS): Prokaryote symbol guidelines			FALSE
PN010	2-B	Protein name based on a protein symbol (PS) or gene symbol (GS): Eukaryote symbol guidelines			FALSE

ガイドラインから**51**項目を定義しチェックリストを作成しました。そのうちの**35**項目について**TogoAnnotator**においてガイドラインの適合・不適合を判定するための実装を行いました。全項目については、以下に定義されています。

https://github.com/togoannotator/docs/blob/master/guideline_ja.md#ガイドライン項目詳細の定義

各ルール毎に整備した統制語の一例

- PN028: Delta should start with an upper case letter
  - Acyl-CoA Delta(11) desaturase
  - Delta(1)-pyrroline-2-carboxylate reductase
- PN050: Inactive protein
  - Inactive levansucrase
  - Inactive D-aminoacyl-tRNA deacylase
- PN012: Non-compliance chemical names and symbols
  - Na, Na(+)→ Sodium
  - Iron→Fe, Fe(2+), Fe(3+)

今後の課題

ガイドラインへの適否を定義し、辞書内の遺伝子産物名がガイドライン適合・非適合であるか判別するTogoAnnotatorへの実装については実施済みです。今後は、評価方法を確立して、類似マッチの結果に対して、ガイドラインに準拠している辞書データが検索上位となるような最適化を行う予定です。さらに、遺伝子名、タンパク質ファミリー、タンパク質名を考慮した機能アノテーションの判定や生物種特有のアノテーションを考慮した最適化のために実現するためには、キュレーションされたデータベースリソースの収集および適用が必要です。