9

アノテーションの質を高めるためのツール TogoAnnotatorの開発

○山本 泰智
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）
李 慶範
情報・システム研究機構 国立遺伝学研究所生命情報・DDBJセンター（DDBJ）
藤澤 貴智
情報・システム研究機構 国立遺伝学研究所生命情報・DDBJセンター（DDBJ）

TOGO ANNOTATOR

INSDCに登録するリソース名の正規化作業を
効率化するツールです。

本研究では遺伝子産物名などのアノテーションをリソース名と定義しています。

アノテーション作業を支援

- INSDCに対して、塩基配列と共に各遺伝子領域の生物学的機能などを記述するアノテーションを登録するさいには、一定のルールに基づく記述が求められる
- 自動アノテーションプログラムなどで得られた結果のままでは不適切な場合がある
- DDBJアノテーターは与えられたアノテーションを逐一確認し、適宜書き換えを提案
- タンパク質命名ガイドラインに当るなどしながら人手により行われており、時間と労力が必要

- 計算機を用いてアノテーション作業を効率化できないか検討
- その結果、機械的な書き換えで済む事例があることや、与えられたアノテーションに含まれる単語が、特定の辞書に含まれるか否かを示すことも有効であるという結論に至る

TogoAnnotatorの特徴

高速

クエリの構成やインデックスの格納方法を工夫し、数千以上のゲノムスケールの大量のアノテーションデータであってもストレスなく処理することが可能

TogoAnnotatorの構成

あらかじめ用意した一連の辞書を、入力文字列に対して順次適用

ブラックリスト

- 不適切な単語、例えば、起こりやすいスペルミスを格納しておき、適宜それを修正するための辞書

ホワイトリスト

- アノテーターが確認作業を必要としない語のリストで、マッチした場合にはその旨が示される

書き換え辞書

- 特定のアノテーションに対して、より適切なものに書き換える
- 多少の表記ゆれでも対応できるように、類似マッチも行なう
- 与えられたアノテーションが書き換え後のそれであれば、そのまま出力
- 一方で辞書中の、書き換え後のアノテーションに類似検索も含めてマッチしない場合は、辞書中の、書き換え前のアノテーションを検索
- その結果としてマッチするものがあれば、それに対応する、書き換え後のアノテーションを出力

TOGO ANNOTATOR

Alginate biosynthesis protein

Whiteリスト
Black リスト

マッチの順序	完全マッチ	類似マッチ
書き換え前	3	4
書き換え後	1	2

elastic

書き換え辞書

書き換え辞書

書き換え前	書き換え後
LptA	LptA protein

この例の場合

1. 辞書の書き換え後見出し語に完全マッチするものはあるか？
   No!
2. 辞書の書き換え後見出し語に類似マッチするものはあるか？
   Yes!

"**Alginate biosynthesis protein AlgA**",
"Alginate biosynthesis protein AlgF",
"Alginate biosynthesis protein AlgK",
"Alginate biosynthesis protein AlgX",
"Alginate biosynthesis protein Alg44",
...

上記1から4までの検索と、複数の入力に対する検索を一度のクエリで実現することにより、高速化を実現した。

Alginate biosynthesis protein AlgA

アノテーターはTogoAnnotatorの出力する書き換え提案を確認しながら最適なアノテーション を決める。

入出力はWebページで可能であるほか、REST APIによることも可能。

検索性能の評価

- 正解データ
  - 690件の入力文字列に対して、望ましいリソース名を上位10件人手で構築
  - ただし、リソース名のリストが10件に満たないものもある
- TogoAnnotatorによる検索結果と比較
- 評価軸
  - 再現率（Recall）
  - 正解率（Precision）
  - 順位を考慮したDiscounted cumulative gain（DCG）

結果

- DCG = 1（順位も含めて適切に検索） 45
- 0 < DCG < 1 452
  - Recall = 1 80
  - Recall < 1 372
  - Recall = 0 （DCG = 0）
- DCG = 0 193

- Precision = 1 2
  - 正解セットには10件全てではない事例もあるため、45にはならない。
- 0 < Precision < 1 495
- Precision = 0 （DCG = 0）

考察

- 0 < DCG < 1でRecall = 1の場合は、順位の調整が必要
- 0 < DCG < 1でRecall < 1の場合は、正解データの確認も必要
- DCG = 0 の場合は、書き換え辞書の不備とTogoAnnotatorの検索手法を確認が必要
- DCG = 0 の193件に対しては、クエリーおよび結果の返却を期待するリソース名が辞書にないことを確認し、今後の方針を確認した
  - ✓辞書にない場合 → 辞書リソースを追加する
    - ✓クエリーがGeneSymbol由来のリソース名で辞書にない
    - ✓クエリーに完全にマッチする名前が辞書にない
    - ✓クエリーがfamily/superfamily/domain/repeatを含むリソース名で辞書にない
    - ✓クエリーがncRNA遺伝子名などのCDS以外のリソース名で辞書にない
  - ✓辞書に存在しないのが正しい場合 → ブラックリスト、書き換え辞書で対応する
    - ✓クエリーが化合物などのリソース名

継続的な辞書リソースの高度化により性能の向上を目指す

今後の予定

- 今回の評価を受けた辞書の修正
  - 辞書に新たに追加すべきか、入力データの不備として扱うかを判断

- DDBJ登録配列の機能アノテーション査定の実運用化
  - 完全一致のアノテーションはスキップし、類似判定の高精度化のための辞書データ高度化のためのイテレーションを実施