塩基配列(INSD)および遺伝子発現データ(GEO)のデータバンク目次

小笠原　理[1]、渡邊　康司[2]、栗原英輔[2]、森山丈士[2]、大久保　公策[1,2]

1. 国立遺伝学研究所　生命情報・DDBJ研究センター
National Institute of Genetics, Center for Information Biology, DNA Data Bank of Japan
2. 情報・システム研究機構機構・ライフサイエンス統合データベースセンター
Research Organization of Information and Systems, Database Center for Life Science

要約

塩基配列データベース共同体 (INSDC)やNCBI GEO (Gene Expression Omnibus)のような、いわゆる一次データベース(バンク型データベース)には、どのような研究がどのような目的でどの程度の規模で行われたかといった、有用な情報が含まれている。そこで、データバンクのデータをプロジェクトごとにまとめるなどの分類をおこない、それを「目次」として用いることにより、より意義深い切り口でデータの俯瞰、検索、解析、取得が可能となる。われわれはこの目的のために DNAバンク目次およびGEO目次を構築してきた。本発表ではこれらの開発の進展および利用方法について報告する。

1. DNAバンク目次

INSDCが提供する塩基配列データベース(INSD)の大きさは現在約一億レコードに上る。これらのレコードはもともとは様々な研究プロジェクトに由来しているのだが、INSDでは由来するプロジェクトと関係なく配列を単位として互いに独立なエントリとして扱われている。そこで我々はINSDの各エントリ中のリファレンス情報等をもとに、全エントリを研究プロジェクト単位に束ねた。その結果INSDを構成している約一億レコードは約59万プロジェクトにまとめられた。これによりエントリをプロジェクト単位で取得することが可能となり、全文検索速度も向上した。さらに研究プロジェクトを研究の型別に分類したことにより、検索結果を研究タイプ別に俯瞰できるようになり、また、それぞれの型の研究がいつ頃どれだけ配列を登録したか、最近はどのような型の研究が盛んに配列を登録しているのかがこのデータベースから読み取れるようになった。現在は検索機構の高速化などをおこなっている。

2.遺伝子発現バンク(NCBI GEO)目次

GEOは遺伝子発現情報に関するデータバンクであるが、このような進展の早い実験領域の1次データバンクはえてして利用者には難解である。少しでも利用しやすいようにデータの整理パイプラインを作成しDNAバンク同様の目次を作成した。

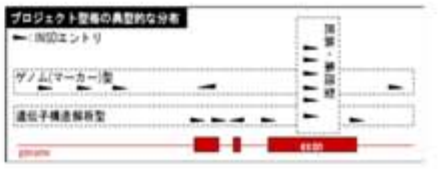
作成方法（研究の型別分類）

(1) 生物別ではないディビジョンのエントリがプロジェクト内で最多 => そのディビジョンに分類

(2) moltypeにより分類
- mRNA　トランスクリプトーム型
- その他RNA　機能性RNA・RNAゲノム型
- DNA circluar　オルガネラゲノム型

(3) 調査個体(/isolate)、調査系統(/strain)が多種類か? => 民族・集団型

(4) 研究対象は特定の遺伝子か?
- 対象の記載なし　ゲノム（マーカー）型
- 対象が多種類　エクソン構造型
- 対象が１種類　免疫、嗅覚、エクソン構造型



1. DNAデータバンク(INSD)目次

http://lifesciencedb.jp/ddbj/

キーワード検索

(DDBJのARSA web APIを使用。図はDDBJからのレスポンス待ちの画面)

キーワード検索結果の一覧表示画面

検索結果はDDBJのディビジョンごとに分類、さらにプロジェクト目的別に細分されて表示される。

登録先特許庁単位で結果を分類表示

プロジェクトの詳細表示

2. GEO目次

遺伝子発現データに関しても塩基配列データと類似のインタフェイスを提供することにより、どのようなデータが登録されているのか一覧できるようにした。(右側の図を参照)

データバンク目次と同様、データがとられた生物種をDDBJディビジョン単位で分類し、発現データの測定方法でデータを細分した。

http://lifesciencedb.jp/geo/