

BH15.15/DDBJ

DDBJ 登録システム周辺の開発

- Annotated Sequence登録に関する情報定義(真島、李、藤澤)
 - Structured_Comment定義 [背景・導入と要約](#)
- DDBJ Validation Service開発
 - BioSample validationルール定義
 - OWL/RDFを利用したBioSample validation
 - TogoAnnotatorを利用したproductのvalidation
 - Web Application開発

BH15.15関連開発

- 微生物ゲノム配列解析パイプライン作成(鈴木治夫、板谷、横江、永田)
- 次世代シーケンサーのRDFデータスキーマの構築(理研:小林)
 - DDBJ DRAのXMLスキーマを参考にRDF化を試みる ← 藤澤さんご議論お願いします
 - 理研作業フローの確認
 - Excel(バイオリジスト) → RDF → DRA XML

1日目

目的

Structured COMMENTの入力情報についてユーザにDDBJ HPを通じてアウトプットする

- <http://www.ddbj.nig.ac.jp/sub/comment-j.html>
- http://www.ddbj.nig.ac.jp/sub/mss/annotation_file-j.html

日本語版の差し替え: コンテンツは準備完了。

英語版の差し替え: コンテンツをなるべく早く準備。

VPN 接続に失敗した上に昨晚の停電 (3/16 20:45- 30min) で DDBJ の

機能停止中なので、戻ったらなるべく早く反映する。

作業

Structured_CommentについてDDBJ定義情報として[スプレッドシート](#)に整理する

- [Structured COMMENT - DDBJ](#)
- <http://www.ncbi.nlm.nih.gov/genbank/structuredcomment>

定義

- Structured Commentの表記は？ ”structured COMMENT”
- では、structured COMMENT とは？
 - INSDC (or RefSeq) の「flat file 上の COMMENT block において ##[tagset_id]-START## と##[tagset_id]-END## の行に挟まれた行で ”:” で区切られた tag_name と tag_value の組で構造化されたデータ記述。」
 - 記述される内容は、通常、その flat file に示される配列データ全体にかかる補足的説明であることが期待される。
 - DDBJ annotation file では ST_COMMENT feature
- “tagset_id” とは
 - tag_name, tag_value を共有するデータセットの名称
 - DDBJ annotation file では ST_COMMENT feature 配下の tagset_id qualifier の値
- “tag_name”
 - tagset_id を共有する structured COMMENT において、共有されるデータ項目の名称

- DDBJ annotation file では ST_COMMENT feature 配下のユーザー定義 qualifier の名称
- “tag_value”
 - tag_name で示されたデータ項目に対応する個別の値
 - DDBJ annotation file では ST_COMMENT feature 配下のユーザー定義 qualifier に対応する value

作業方針

- まず、実の登録運用で INSDC 共有している以下のtagset_idについてまとめる
 - GenomeAssembly-Data
 - Assembly-Data
 - International Barcode of Life (iBOL)Data
- その後に、NCBIで扱っているtagset_idについても整理する
 - MlxS
 - HIV

Todo

- [スプレッドシート](#)のdescription、description_jaをすべて埋める
- exampleをすべて埋める
- 留意すべき点、制約、ルールなどをcommentカラムに記述する。
 - [Structured COMMENT - DDBJ](#) 記載 の情報を反映させる
- INSDCリソースにおける関連するqualifier/attributeなどの情報をrelated resourcesカラムに記述する
- Valuesの全リストを現在のエントリーからパースする【Done】

微生物ゲノム配列解析パイプライン情報共有（鈴木、藤澤）

- ゲノム微生物学会「シアノバクテリアに特化したCyanoBaseゲノムアノテーションパイプラインの構築」共有
 - メタデータ追加部分について、Prokka4ddbjコードを共有
- 他に、慶応ではゲノムアノテーション出力についてはinference/db_xrefでアノテーションを足したい
- Product nameの大文字／小文字変換問題について確認
- その他、共通リソースがあれば共有していく

TogoAnnotator辞書開発(山本、李、真島、藤澤)

辞書開発方針の確認

- before/afterのアノテーションセットとしては、以下の2つを準備する
 - INSDC-UniProt
 - UniProt-RefSeq
- UniProtのReviewed/UnReviewed および RefSeqのstatusで辞書データセットにフィルターをかけられるようにする
- RefSeqのstatus情報について調査する(Todo: 李)
- 辞書データからtaxon単位でのフィルターも設計上可能
 - アノテーションデータを分けたいtaxonをリストアップできるとよい
 - まずはCyanobacteria(taxid:1117)で試行

2日目

tagset_id、tag_valueの使用状況の確認

Structured COMMENT情報をDDBJ release 103 およびGenBank 210 wgs master エントリーからパースし

tagset_idの頻度情報およびDDBJ submissionにおいて対応している

Genome-Assembly-Data, Assembly-Data, International Barcode of Life (iBOL)Dataに限定してtag_valueの頻度情報を算出し、[スプレッドシート](#)に追加した。

RefSeq COMMENT分析。

ただし、NCBI 側でも document 整備が不十分な様子なので、憶測が入っているかもしれない。

[Structured COMMENT - RefSeq](#)

↑ **RefSeq status**情報の調査(李) をマージ

3日目

TODO

- Sequencing Technology, Assembly Methodの入力具体例をリストアップ
- DDBJ 用にStructured COMMENTのドキュメント策定する(真島、李)
 - 技術文書的な structured COMMENT とその要素の定義
 - Annotated Sequence/Structured COMMENTのバリデーションルール策定
 - 配列データ登録用の annotation file への記載方法: ST_COMMENT feature

- フラットファイルの読み方の説明: structured COMMENT
- ラップアップ用資料策定

技術文書的な **structured COMMENT** とその要素の定義 (新規)

どこかで正式に公開すべきかもしれないが、場所など未定

- 表記: structured COMMENT
- “structured COMMENT”
 - INSDC (and RefSeq) の「flat file 上の COMMENT block において ##[tagset_id]-START## と ##[tagset_id]-END## の行に挟まれた行で “::” で区切られた tag_name と tag_value の組で構造化されたデータ記述。」
 - ユーザー(データ登録者を含む)コミュニティが定義したデータセットを flat file を介して公開・共有することを可能にする。
 - ただし、実際には、INSDC 側の都合で記述させていることも多い。
 - 記述される内容は、通常、その flat file に示される配列データ (エントリ) 全体にかかる補足的説明であることが期待される。
 - DDBJ annotation file では ST_COMMENT feature の配下に記載される。
- “tagset_id”
 - tag_name, tag_value を共有するデータセットの名称。
 - DDBJ annotation file では ST_COMMENT feature 配下の tagset_id qualifier の値。
- “tag_name”
 - tagset_id を共有する structured COMMENT において、共有されるデータ項目の名称。
 - Flat file 上では COMMENT block の ##[tagset_id]-START## と ##[tagset_id]-END## の行に挟まれた行内で “::” で区切られた左側に表示される。
 - Tag_name を複数回、記載することはしない。tag_value 相当の値が複数ある場合は “,” 区切りで列挙する。
 - DDBJ annotation file では ST_COMMENT feature 配下のユーザー定義 qualifier の名称。
- “tag_value”
 - Tagset_id と tag_name で示されたデータ項目に対応する個別の値
 - 値が複数の場合は “,” で区切って列挙する。tag_name を複数回、記載することはない。
 - Flat file 上では COMMENT block の ##[tagset_id]-START## と ##[tagset_id]-END## の行に挟まれた行内で “::” で区切られた右側に表示される。
 - DDBJ annotation file では ST_COMMENT feature 配下のユーザー定義 qualifier に対応する value
- COMMENT としての見え方
 - COMMENT は 80 字で改行する。
 - “::” で区切る位置は固定長ではなく、データセット内で最も長い tag_name に合わせる可変長。
 - Tag_name が長い場合、[tag_name] + “::” + [tag_value] で 80 字に達した際に改行するが、“::” の後までスペースインデントする。

配列データ登録時の **annotation file Feature/Qualifier** としての記載ルール

- 1つの ST_COMMENT feature に対して、1つ tagset_id qualifier の記載が必須
- 1つの entry に ST_COMMENT feature を複数回、記載して良い。

- 1つの entry に同一の tagset_id を持つ ST_COMMENT feature を複数回、記載することはできない。
- 1つの ST_COMMENT feature 配下に同一の tag_name (ユーザー定義 qualifier) を複数回、記載することはできない。
- tag_value で示すべき値が複数ある場合は “;” で区切り、列挙する。
- いくつかの tagset_id の値と tag_name (qualifier 名) は辞書化してチェック (jParser) する予定。3月末実装。

下記既存サイトの書き換え: **2016.03.18** 済

配列データ登録用の **annotation file** への記載方法: **ST_COMMENT feature**

http://www.ddbj.nig.ac.jp/sub/mss/annotation_file-j.html#5-4

http://www.ddbj.nig.ac.jp/sub/mss/annotation_file-e.html#5-4

Structured COMMENT 説明ドキュメント

<http://www.ddbj.nig.ac.jp/sub/comment-j.html>

下記既存サイトの書き換え: **2016.03.22** 済

Structured COMMENT 説明ドキュメント

<http://www.ddbj.nig.ac.jp/sub/comment-e.html>