1. 課題（統合データベースプロジェクト当初課題名）

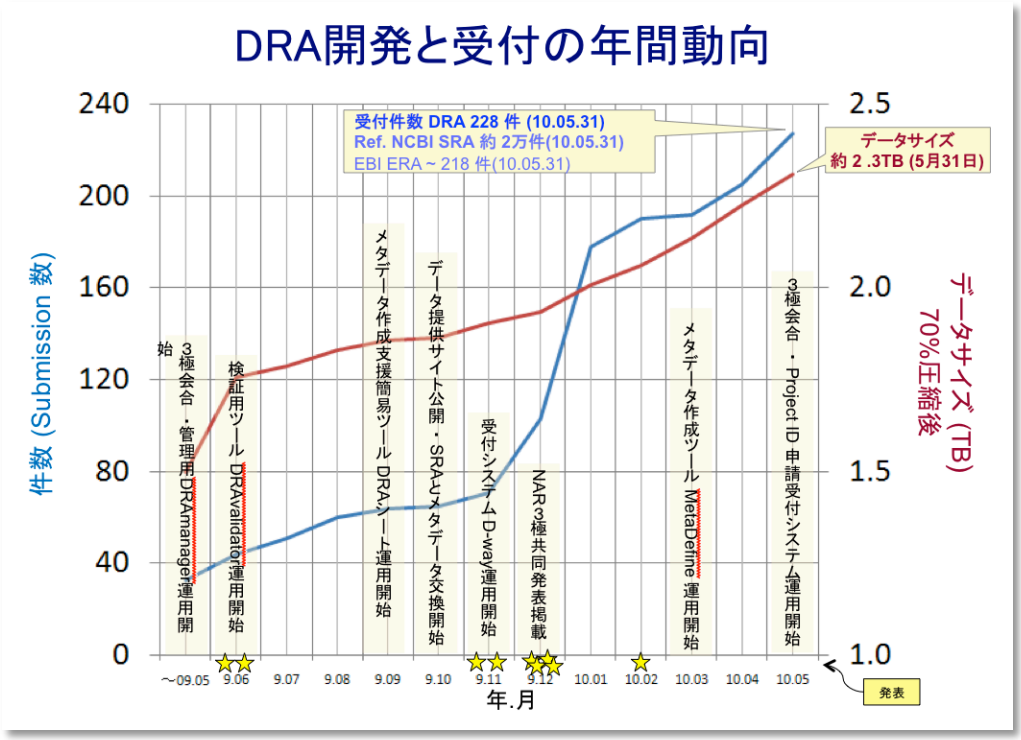
塩基配列アーカイブのデータベース構築と統合への貢献

2. 今年度の目標（2010 年度 JST BIRD の研究計画書から）

- **受付システムの高度化と運用**：利用者インターフェースの充実とデータの検証機能を強化し、また、新たに定義した「fastq データより豊富な情報を含む」SRA ファイルに対応する。このため、国際協調の一環として NCBI が開発している SRA toolkit を導入する。
- **公開システムの高度化と運用**：メタデータの多様な項目の複合条件での検索と検索結果のダウンロードを実現する。SRA データのダウンロードも実現する。
- **管理システムの高度化と運用**：各種統計情報の自動生成、データのバージョン管理、DRA 側でのデータ検証、データ更新の自動反映、公開日管理の機能をシステム化して、得られる統計情報をシステム改善に活かしていく。
- **リードアーカイブ登録支援の強化**：リードデータを解析して DRA への登録支援を行なう解析パイプラインを開発する。リードデータの注釈付け処理や、解析結果のユーザの登録支援を行なう。
- **国際連携**：米国 NCBI の SRA ならびに欧州 EBI の ERA と国際塩基配列データベース実務者会議の機会も活かして、国際協調のもとに DRA の構築・運用を行う

3. 現状（5月末まで）

DDBJ Read Archive の開発と受付の年間動向（2010 年5月末まで）は下図のとおり。



（現状（5月末まで））

1) データ受付公開実績:受付件数 228 件、公開件数 27 件

2) データ登録環境の拡充

(1) 登録者のデータ登録作業の効率をあげるために、Project ID 申請フォームの登録受付システム D-way への組み込み、ユーザ評価をもとにしたメタデータ作成ツール MetaDefine の改善、を行った。

(2) ユーザからの要望や研究技術の進歩に対応するため、DRA/ERA/SRA 共同で新しいメタデータの XML スキーマ version 1.2 を作成した。

3) パイプラインの構築:高次解析部と基礎解析部のテスト版構築

高次解析の接続のテスト版を構築した。サンプルコードとして、ゲノムの large insertion/deletion を検出するプログラム BreakDancer をパイプラインに組み込んだ。具体的には、front-end (GUI部), back-end (データベース部＋クラスタ解析部)を開発した。

4) 国際連携

(1) DRA と NCBI SRA との間で公開されている全てのメタデータと fastq ファイル(約 30 TB) のデータ交換を実現した。

(2) 5月18-21日に EBI で開催された国際実務者会議における情報交換と議論

① DRA の担当者が参加し、DDBJ/EBI/NCBI 間でデータベース運営上の問題について話し合った。

② NCBI, EBI ならびに DDBJ のアーカイブを SRA（Sequence Read Archive）と総称することになった。なお、日欧はローカルには今後もそれぞれ DRA (DDBJ Read Archive)、ERA(EBI Sequence Read Archive)を使用するが NCBI のアーカイブは SRA と呼ぶので、注意が必要。

③ DDBJの貢献

EBI と NCBI には MetaDefine のようなメタデータ作成のための簡便なツールがなく、登録者はメタデータを XML 形式で作成しなければならず、大きな負担となっている。そのような状況のなか、ERA の担当者が DRA ウェブサイトから公開されている MetaDefine を評価し、そのソースコードの提供可否について打診があった。会議において、DRA が EBI と NCBI にソースコードを提供することで合意した。

4. 6～9月ならびに年度末までの予定

下半期は DRA の運用を継続しつつ、以下の下線部について整備または開発を行う。

1) スキーマ更新への対応

DRA システムを新 XML スキーマ version 1.2 に対応させる。

2) データ提供

3 極から公開されている全てのメタデータと fastq データの FTP によるダウンロード提供をまもなく開始する。

3) SRA ファイルへの取り組み

シークエンシングデータを 3 極の統一フォーマットである SRA ファイル形式で取り扱うため、SRA Toolkit を共同開発した。これを受け、SRA ファイルのデータ交換に取り組む。具体的には、DRA で受け付けたデータを Toolkit で SRA ファイルに変換し、メタデータ、データ更新のためのアクセッション番号リストとともに公開する。また、ERA/SRA のデータを fastq（塩基配列 ＋ Quality value）よりも多くの情報を含んでいる SRA ファイル（fastq 相当 + メタデータ + 基本統計量 + md5 など）形式で取得する。

4) 登録件数増加への対応

次世代シークエンサの普及に伴い DRA への登録件数は増加している。登録者によるデータ登録、アノテータによるデータチェックのより一層の効率化を図るため MetaDefine と D-way の拡充を継続していく。また、データ保存のため、年度内に 200 TB 以上のハードディスクが必要になる見込みである。

5) データ検索システムと DRA 統計情報への取り組み

9月以後データ受付、交換、公開システム開発が一通り完了した後、メタデータの検索システム開発に着手する。また、DRA の動向を表現する統計指標を獲得表示するシステム開発にも着手する。

6) SRA ファイル

NCBI SRA では fastq と SRA ファイルを両方公開しているが、部分的に情報が重複しているため SRA ファイルに一本化する予定である。その場合、利用者は SRA ファイルをダウンロードし、ローカルで Toolkit により fastq ファイルなどを取り出す、という作業が必要になる。このため、NCBI SRA は利用者の負担を軽減するため FuSE system の導入を検討している（http://fuse.sourceforge.net/）。DRA においても9月以後このシステムの導入について検討予定。

7) パイプラインの拡充

(1) Single Sign On(SSO)により、高次解析部と基礎解析部の画面接続統合を行なう。具体的には、DDBJ の SSO を各アプリケーションに組み込み、5 月までの成果を基に、高次解析部の実運用を開始する。また解析結果を目で確認する為の Viewer を組み込む。

(2) 高次解析部として、RNA-seq 解析プログラムの組み込みを行なう。DRA のデータから、発現用タグカウントを定量化する方法を確立する。

5. DDBJ 事業における位置づけ

- 新世代シーケンサ由来のデータは、DRA に向かうとともに、従来の塩基配列データベース DDBJ や、遺伝子発現データベース DOR(従来 CIBEX)にも向かう。そこで、DDBJ では、これらのアーカイブとデータベースへの登録・査定・提供を、利用者の観点も含めて、総合的に扱える情報環境の構築に取り組んでいる。
- これまでの塩基配列データベースDDBJにならぶ事業としてDRAを育成し、長期的に運用するために、予算の面でも人材の面でも計算機資源の面でも、DDBJ 運営費の枠組みと JST BIRD 事業の枠組みとの間で最大限効率的に連携を図ってきている。

以上