統合データベースプロジェクト研究運営委員会作業部会分科会
【遺伝研・JST/中核の参画機関(基盤技術関係機関)】議事要旨

【日　時】　平成２２年１月２９日（金）１０：００～１３：００
【場　所】　ライフサイエンス統合データベースセンター大会議室
【出席者】　五條堀孝(遺伝研)、菅原秀明(遺伝研)、黒田雅子(JST)、中村保一(遺伝研兼かずさ DNA 研)、藤澤貴智(かずさ DNA 研)、岡本忍（DBCLS)、浅井潔(CBRC)、野口保(CBRC)、福井一彦(CBRC)、松本裕治(奈良先端大)、堀田凱樹（ROIS)、大久保公策（遺伝研)、高木利久(作業部会主査)
永井啓一、西川哲夫、川本祥子、坊農秀雅、畠中秀樹、吉羽洋周、河野信、高祖歩美、八塚茂、箕輪真理(以上、DBCLS)

（敬称略・順不同）

【議　事】

１．平成 22 年度予算配分について(文科省資料を高木主査より説明)

◇資料説明◇　資料 1-1～4、資料 2

本来ならライフ課が説明する予定だったが、本日は担当者の日程が合わなかったので、代わりに説明する。資料 1－1 にあるように、目的は適正な委託費の配分のためで、そのための評価を作業部会が行うという方針。実施については、1 月 25 日に開催された運営委員会でも承認された。本来なら外部へ評価を依頼したいが、内容が複雑なので、ご判断いただける作業部会のメンバーに評価を依頼し、その結果を持って、文科省が配分を判断する。資料 1－2 にあるように、本 PJ が最終年度にあたるので、ユーザーが利用する価値のあるものかどうか、判断してほしい。22 年度の計画については、開発は上期(9 月末まで)に終了し、後はブラッシュアップに注力してほしい。内容についても JST 内に設立予定の新センターへ引き継ぐべき内容かどうかという観点で判断してほしい。22 年度予算としては、DBCLS 以外は 4000 万円、DBCLS の予算も 5000 万円くらいの減額が見込まれ、限られた予算の中での精査のための評価をお願いしたい。評価の単位は機関ごと、ただし、医科歯科 G と東大 G についてはグループ単位とする。自分の利害の関わらない範囲でお願いしたい。資料 1－3 はスケジュール、1－4 は参考資料として各機関の 21 年度の委託費の内訳(概要)の一覧。まとめて評価が難しいという意見が運営委員会でも出たが、細かい意見はコメント欄に記載し、全体評価をお願いしたい。

資料 2 は 21 年度の計画と 12 月末の時点での達成状況をまとめたものである。これも参考にしながら、議論し、評価をお願いしたい。

◆質疑応答◆

○JST の立場としてはむずかしいので、評価を辞退したいが可能か。

→文科省とご相談いただきたい。

○DBCLS も評価するのか？

→文科省からの説明では、「利害関係がない範囲で」ということらしいが。例えば菅原先生は中核機関の作業もやられているので、その場合はふさわしくないので、という意見があった。しかし、できる範囲でなるべく多くの機関に対して評価をしていただきたい。忌憚のない意見をお願いする。

２．中核機関より

プロジェクトの最終年度の取りまとめを前にして、中核機関の活動内容をご理解いただくことが肝要と考え、中核機関の達成状況、22 年度の目標について、川本特任准教授から紹介された。

◇資料説明◇　資料 3

これまで DBCLS は他の機関のように発表する機会が無かったので、今回の一連の作業部会で発表させていただくことになった。

取り組みとしては、統合ホームページの概要（LSDB Lab を新設したので、ご利用いただきたい)、アクセス

統計、カタログ系サービス、横断検索、アーカイブサービス、統合 TV、共通基盤開発(その情報流通における位置づけ)、これらの年次計画上での進捗を紹介した。次年度の計画としては、統合化の全体像とステップを提示した上で、中核（DBCLS）と各機関の連携状況、横断検索の機能向上から統合検索への展開、そのためのツールとしての辞書シソーラス、統合検索のプロトとしての TogoProt を紹介した。また、ユーザー評価について、各機関の協力に感謝するとともに、結果の公開について紹介した。

◆質疑応答◆

特になし

３．プロジェクトの平成 21 年度進捗状況および平成 22 年度業務計画について

- 遺伝研

  ◇資料説明◇　資料 4-1、4-2

  当初 DRA Trace Archive という名称でスタートしたが、次世代シークエンサー対応のため ShortRead をつけた名称に一時変更し、ShortRead もそぐわなくなったので今年度からは DDBJ Read Archive という名称にした。現在登録件数は 113 件で 70%圧縮後のデータサイズが 2TB。（参考：NCBI の SRA は 1 万件）昨年の 5 月くらいから、ツール・システムが揃ってきた。登録機関数 13 機関、公開件数は 18 件（Submission ベース）。来年度の目標（案）として、①受付システムの高度化と運用、②公開システムの高度化と運用、③管理システムの高度化と運用、④国際連携（3 極会議@EBI、2010.5 予定）を検討中。今後のデータ量の見積もりとして、国内のシークエンサー設置数の状況をみると、総数 110 台程度。設置はされているがまだ DRA への登録の無いところもあるので、今年の 6 月以降位にデータが出てくるのではないかと思う。引き続き、DRA のホームページの紹介。

  ◆質疑応答◆

  ○補完課題としては Read Archive から始まっているが、Omics やパイプライン等、DDBJ 全体の連携についても補完課題の中でも検討していただけないか。成果として見やすくなってくると思う。

  →志としては可能。予算次第で盛り込むよう検討する。

- JST

  ◇資料説明◇　資料 5

  ROIS の開発分を分担するという形で､外部への発注を行った｡発注内容の一覧については資料参照のこと。これらについての成果報告書をどのように ROIS と分担するのかなど、文科省と相談しながら今後進める。それ以外の JST としての活動については、中核との内容の重複に見える部分もあるので、今後要調整であるが、ポータルサイトとして初年度から開発してきた WINGpro についてはユーザーからの要望に従って、リンクなどを追加した部分もあるが、基本的には大きな開発はしていない。ただ、書き込みがあれば表に反映される仕組みになっている。事業サイトとしては、プロジェクトの報告書などの掲載のほか、ライフ課関係の情報サイト、初年度開発されたサイトも継続して運用している。また、19 年度に開発した DB の標準化のためのメタデータの情報サイトについても継続運用しているが、時間の関係もあり、新たな情報の追加はあまりできていない。次年度は、中核機関の予算執行の手続きが作業としては大きな割合になると思うので、それを主としてあとは紹介した既開発分のサイトの運用などを行う。昨年、サーバーへの不正アクセスがあった関係で、PJ 用の独自システムは無くなった。JST 内で他の用途に使っているシステム（セキュリティがかなりしっかりしているもの）の下に組み込んだ。

  ◆質疑応答◆

  ○（補足として）H19 年度から H20 年度にかけて大幅な PJ 予算縮減があったため、H20 年度からは JST は予算配分無しにプロジェクト関連の作業を実施することになっており、今回の発表内容の後半については

予算ゼロで実施していただいたもの。さらに、H21 年度は、文科省からの委託費が減り、JST の運営費交付金から一部 PJ へ回してもらっていて、PJ 全体としては合計 11 億で変わらずとなっているが、そのうちの 2.5 億円については、JST の予算から捻出されており、資料の 1 ページの開発項目についてはその 2.5 億円が充てられている。H22 年度は文科省からの予算は 3.8 億円となるので、JST 予算からの捻出が 6 億円程度になる予定。H23 年度以降何らかの形で本 PJ の後継があるとすれば、すべて JST 負担になるのではないかという予想。

→本年度 2.5 億円のうち、3 千万円については、遺伝研菅原先生の開発費用に充てられている。従来の BIRD の活動内容についても精査しながら予算をやりくりしている。

○セキュリティをしっかりするということが、Openness を担保することに悪い方向に働かないといいが、そのバランスが難しい。

→リンク先のサーバーの都合で自動遷移されていたりすると、それだけでセキュリティ上問題になるケースもあった。所属機関のセキュリティポリシーによっては、見えないところで様々な制限がかかっている場合もある。どこの機関もいろいろな対策を取っているようで、費用もかかることなので検討は継続中。今回のケースについては、公的なサイトでもあったので、高度なセキュリティを設定して対応した。

- かずさ

◇資料説明◇　資料 6　配布資料に一部追加されたパワーポイントを投影しながら説明

ゲノムアノテーションを新しい Reference を追加しながら維持するモデルケースとして、ユーザーによる情報追加が可能な仕組みを構築。蓄積された情報は即時公開。情報がかなり高密度に集積されている（5613 文献分、145,890 件の遺伝子名出現情報を人手で入力、さらに計算処理によってその情報が 249,082 件まで増えている）ので、その情報を閲覧するための方法も DASviewer を利用して開発。入力時のタグ情報を標準化し選択できるようにして、入力支援機能を追加。遠隔アノテーター10 名、常に 5-6 名が活動、週 1 回でネット上会議(47 回、USTREAM や、SKYPE を利用)、1 回だけセンターで実際に会議。議事録等も共有。

次年度は、システムについては維持運用、情報については蓄積を中心に推進。

◆質疑応答◆

○リードアーカイブのデータが増えていくに従って、アノテーションニーズが高くなるが、今後どう進めるべきと考えるか。

→すべての生物種にこのシステムを対応させるのは無理だが、すでに精査されたアノテーションがついている類似の生物ゲノムをベースにそれとの関連を見ていくのが速いと思う。それをサポートするためのリソースの整備が必要。このモデルがうまく動いているのは、対象の生物種にもともと詳しい研究者が参加していることが大きい。新たな生物種に対応するためにはチーム構成を新たに検討する必要あり。

○かずさの成果としては、１．対象とする生物種のアノテーション DB が構築できた、２．この生物種と類似の生物種についてモデルとなりうるデータができた、の２点と考えていいか。

→そう思う。

○遠隔地キュレーターへの支払いについて。

→遠隔地雇用者へはかずさからは直接払えないので、業者を仲介して雇用している。

○論文内にもオルソログ情報などを利用した予測にすぎない情報も含まれている。そのような情報への依存度も高まっていくと思うが、一方で悪影響を及ぼしている状況もある。メタゲノムの情報に、オルソログ関係等から得られた情報を付与していくときに、どの情報が実験に裏打ちされているかという注釈が必要になってくると思うが、そのような情報を付けていくという計画はあるのか。

→遺伝研の業務にもかかわるとは思うが、今後は新型シークエンサーから出てきた配列情報の処理などにも、そのような情報が必要になってくると思う。

○来年度の着地点としては、モデルとして使える DB の確立とその応用ということでいいか。その先はどのような考え方で進めるべきか。
→このプロジェクトとしては、モデルの確認をし、主要な光合成生物について応用を実施して一区切りと考えている。その先の展開としては、同じモデルを別の生物に展開するほか、同じ仕組みを用いて、集積する情報の検索対象を変えたり、DB に格納する情報をリッチにするなどが考えられる。
○次世代シークエンサーが多くの施設に入っているようだが、データの処理をどうするのかを各施設ではきちんと考えているのか。
→何も考えていないと思う。遺伝研ではそのためにデータを持ってしまった人たち向けのパイプライン(基本アノテーション)を構築しており、これは基本的な処理を終えたデータの登録を促すためのもの。
○機械的な処理では付与できるアノテーションはわずかになってしまうのでは。
○きっちりしたアノテーションがついた情報があれば、ラフに読んだ情報へのアノテーションとして応用できるという考え方が主流になっている。問題は表現のしかたで、今はまだ読みにくいが、これを簡単に読めるくらいまでにしたい。
→表現の仕方を工夫するのと、情報の使い方についての提案をしていく必要があると思う。
○シアノバクテリアのウェブ上の本となるくらいの詳細な情報をまとめる努力をしてほしい。
→論文を読みこんで初めて得られるような情報もあるので、そのような情報が失われないようにしたい。
○信頼できるアノテーションを付けるのも本 PJ のミッションではあるが、同時に大量データが出てくるようになっているので、そのような時代に対応するにはどうすべきかをこの 1 年間でプロジェクトから提案できるとうれしい。
○シークエンサーが入るのはいいが、それを使いこなすところで一苦労、またデータが出てきて一苦労、となるのが目に見えている。
○アノテーションは機械と同じように予算を付ければデータが出てくるというものでは無い。工夫が必要。

- 産業技術総合研究所生命情報工学研究センター（CBRC）

◇資料説明◇　資料 7
H21 年度の実績として、１．タンパク質立体構造モデリングワークフロー（WF）［ウェブ環境で使える利便性の高い WF］、２．アクティブ WF［パワーユーザー向け、いくつかの環境で使える柔軟性の高い WF、処理過程が見える］に向けた環境開発、３．CBRC 統合情報基盤サイト HP の更新を行った。２．については、CBRC のサーバーを用いる環境、負荷の高い計算処理だけ CBRC サーバーを用いる環境、ユーザーの環境で Local に実行する環境をそれぞれ開発した。提供プラットフォームとしては KNIME を使用している。最終年度はユーザーが考えるであろう WF についての要望を念頭に、アクティブ WF 向け要素技術の開発［要素技術のノード化］、WF によるプラットフォーム開発［利用者の行いたい解析フローを CBRC 内外のリソースを用いて構築］を行う。
◆質疑応答◆
○このモデリング WF はある程度相同性の低いものでも解析できるようにするものか。また、すでにあるモデリング結果の DB を参照するというやり方もあると思うが、それについてはいかがか。
→プログラムについては BIRD で開発された WF を参照している。これはヒトゲノムを主な対象としており、あらかじめ狭められた領域について予測を行う Swiss モデルや Rosetta などとは異なり、まずどこかドメインかということを見ながら検出するというのが特徴で比較的長いものを対象とできる。遠い種についての精度はあまりでていない。また、他の予測データは参照していない。
→これまでの方針では CBRC で開発された予測ツールを使うことが前提であり、作り手の観点でできている。すでにスタンダードとなっているものや、既存データの参照を含めて解析しようというユーザー視点で

はない。現段階では、まだツールや外部データとの組み合わせの試験的な部分もある。今後の計画の中で既存情報への検索結果も併せて表示するといった観点も重要だと思う。
○資料 7 の 13p にあるユーザーの問題意識の解決になっているか。
→実際には、直接の答えとなるものを提供しているわけではなく、そのための要素を開発している段階。結局、やって見せるしかないと思っている。次世代シークエンサーのデータを考えてみても、データを外に出したくない（出せない）とか、膨大な計算機パワーを要するなど、いろいろな問題がある。それらの解決に向けてヒントになるものをあと 1 年で出したい。
○CBRC で開発したものを使うよりも、すでに確立したものを使いたいという気持ちがユーザーには大きいのではないか。外部で確立したものを取りこむほうが、ユーザーの理解を得られやすいのではないか。
→この PJ の始まりが開発だったという経緯があるが、確立したものをモジュール化して組み込むことはやっていきたい。
○国際開発会議 BioHackathon（BH）との関係はどう考えるか。
→BH のみならず、外部機関との連携については、DBCLS と相談しながら進めたい。特に配列系の WF については、CBRC 内部には 1 次 DB が無いので、連携の必要性が大きい。
→KNIME の中には確立された手法のコンポーネントもあるので、CBRC ツールに限定しているということではない。資料の中では CBRC ツールが特に紹介されているが、限定するものではない。

- 奈良先端大

◇資料説明◇
今年度の成果として、目標として設定していた、専門用語辞書システムの開発［内部構造・同義語による検索、意味情報の管理］、専門用語解析技術の開発［内部構造解析データを追加、自動解析の開発、高性能化］、専門用語抽出ツールの設計と開発［意味クラス分類手法、シソーラス登録システム UI 設計］を実施した。最終年度は、専門用語辞書システムの開発［用語の意味関係や類似度を表示する機能、辞書システムの汎用化］、専門用語解析技術の開発［データの拡大、解析精度の向上］、専門用語抽出ツールの設計と開発［分類手法の高度化、シソーラス登録機能の実装他］を予定している。
◆質疑応答◆
○内部構造がわかると何ができるか。
→１つは用語分類精度が上がる。また、用語の関係性がわかる。
○内部構造の種類として対象 DB はどのくらいの単語数のものを用いているのか。
→ライフサイエンス辞書は 90000 語、病名マスターは 23000 語で、両方に共通しているのは 5500 語位。
○PJ の中での位置づけは、技術開発なのか。
→技術開発を進めつつも、精度を高めるために範囲を病気に絞って、実際にデータを貯めている。
○英単語との連携は。
→ライフサイエンス辞書では２万数千の用語について MESH との対応付けがあるので、それらについては関連付けられる。
○他の DB のアノテーションと関連付けられないか。
→重ねることはできるが、データの齟齬があった時にどちらを信じるかなどの判断が必要にある。ライフサイエンス辞書については MESH との対応付けで齟齬が無いようになっているので、問題無いと思う。
○類義語の抽出は従来の方法とどこが違うか。
→構文解析をして、直接の関係がある語を見ている
○PNE はすべて構文解析されているのか。
→そうである。解析自体にはそれほど時間がかからない（数時間程度）

〇例として提示されているアルブミンの類義語として挙げられているものがおかしい。
→似たような文脈で使われているという機械的な抽出しかしていないので、まだ意味までは解釈していない。これは内部構造を検討していないので、そのような処理をすれば、精度が上がってくる可能性はある。
〇まだ、コーパスが足りないということもあるのではないか。
〇横断検索は MeCab を使って英語日本語連携をしている。MeCab の拡張（MESH との連携）を使えないか検討したい。

- 総合討論
  特になし

（13：00 終了）