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（医学･化合物･糖鎖関係）議事要旨

【日　時】　平成２１年１月２１日（水）９：００～１３：００
【場　所】　ライフサイエンス統合データベースセンター会議室
【出席者】　浅井委員、五斗委員、田中委員、下川氏(東京医科歯科大)、古崎氏(大阪大)、徳永委員、小池氏(日立製作所)、成松委員、新間氏(産総研)、鹿内氏(産総研)、林氏(九州大学)、山中氏(文科省)、高木委員
堀田機構長(ROIS)、永井、西川、川本、箕輪(以上、DBCLS)

【議　事】
平成 21 年度業務計画・平成 20 年度プロジェクト進捗状況について

議事内容に入る前に高木委員から、作業部会の構成の変更について、「これまでの作業部会では、参画機関、分担機関、補完課題の代表者が全員集まって議論してきたが、議題も多く消化不良となりがちであった。その解消のために、数日にわたって全員で実施という案もあったが、なかなか実現できないそうにないので、関係するテーマを集めて分科会とした。」との説明があった。
初顔あわせのメンバーもいるので、引き続き各自簡単に自己紹介した。

- 予算配分額について

◇資料説明◇
資料 1-2 は中間評価の評価内容の一部として、費用配分が示唆されたもの。これを尊重して総計 11 億円の予算を作成した（資料 1-1）→11 億円の内訳として、プロジェクト自体の予算は 8.5 億円であり、2.5 億円については、当プロジェクトの継続性を考慮した将来的な JST との一体化に向けて、JST からの予算配分となる。2.2 億円が情報・システム研究機構(→DBCLS)に、0.3 億円は遺伝研へ直接配分される。予算に関する資料を提出いただいたので、資料 1-3 のスケジュールにしたがって、今後微調整を経て、確定していく予定。
これまでの経緯については、資料 1-4(中間評価に対する各機関の対応)、資料 1-5（20 年度の目標と進捗）を参照していただきながら、21 年度の計画についてご発表いただきたい。
◆質疑応答◆
とくになし。

- 東京医科歯科大学グループ

◇資料説明◇
《東京医科歯科大学》
20 年度は中間評価の指摘によりプロトタイプの作成に注力。医学関係の統合 DB のあるべき姿を検討することと、そのために全国の DB の状況を把握することに努めた。それに必要な要素技術として、ターミノロジー、倫理規定の基本調査、検索エンジンの要素化等を実施。臨床オントロジーは、国際的な検討の場に提案。また、DB の機能向上のため、現在専門家向け公開中の症例 DB を一部加工して一般向けに公開予定であり、GeMDBJ との連携を一部実施、条件入力が容易な検索 GUI を開発中。21 年度の目標として、オントロジー整備を進め、要素技術の問題点を整理し、倫理規定についても中核機関と相談しながら具体化する。少数 DB の統合を実際に行いつつ、プロトタイプの

改良、具体的な DB 連携方式の開発、網羅的疾患情報の組織化を推進する。
《大阪大学》
医科歯科大と議論しながら、共同で作業している。具体的にはパーキンソン病について医科歯科大と阪大の DB 統合のための要件を定義した。また、19 年度開発の DB の症例を追加し、高度化した。21 年度においては、要件定義に基づきプロトを試作し、DB 公開に向けて倫理問題を解決する予定。
◆質疑応答◆
○倫理規定については東大 Gr でも GWAS について実施しているが？
→来年度は一緒に議論したい
→臨床とゲノム遺伝子解析という点で異なる（関係する法律も違う）もので、現時点ではオーバーラップが少ないが、将来的には包括的なものを作れる可能性はある。
→臨床情報にもゲノム情報などが入ってきている。
○ゲノムを読むことが一般的になる可能性がある将来的なことを考えて、包括的なものが初めからあったほうがいい。
○まずは医科歯科大 Gr と東大 Gr で考えてもらい、それからプロジェクトの中で確認していくことにしたい。
→それでいいと思う。その後文科省や厚労省にも諮る必要あり。国内にガイドラインがないので、米国を参考に 18 項目（連邦法 The Health Insurance Portability and Accountability Act［HIPAA: 健康保険に関する携行性および説明責任に関する法律］に基づく「患者プライバシー保護の基準」に記載された容易に個人を特定できる情報）を除く形で現在は対応している。たたき台は出すので中核機関を中心に検討を進めて欲しい。
○まずは医科歯科大 Gr 案を示してもらい、東大 Gr 案と比較検討し、プロジェクトとしての案を作ることにしたい。その後文科省での検討も必要。厚労省の動きは?
→関係学会などで検討することが必要ではないか。ただ医学団体まで巻き込むと大変になる可能性も。
○検討委員会に代表者が入ってもらうことは必要。
○DB の公開について、匿名化以外にどのような問題があるか？
→匿名化だけでは足らない場合がある。他の情報との組み合わせで特定に近くなる(10 人以下に絞れる)のはダメ。18 項目を削除してもだめなケースも考えられる。米国では 18 項目を削ればデータの売却さえも可能な一方、個人の同意書だけではだめ。
○個人の権利を守るというシステムが別途きちんとしているから、医療情報については緩くできるという状況があるのでは？
→国によって扱いがいろいろ(英国では教育目的ならかなり緩い)。一般に公開する情報は絞り込みできないものにし、教育目的などについては手続きを取ったうえで共有を許すという公開・共有の２つの段階がある。特に共有についてはある種の契約が必要。
○研究のほうはすでにある枠組みで共有が OK なのではないか？
→２つの基準が混同されている。実際は細かいルールができていない。
○学会等では公開や統合ということは議論されていないのか？
→医療情報を扱う学会では議論されているが、個別の学会では議論は無い。

〇がんセンターの情報との統合について、
→発現データなどは研究者自身が自由に解釈できるように生データを置く。
〇検索のためのターミノロジーの統一が重要だと思う。
→検索のためには、項目の一致とターミノロジーの一致の２つの方式が取りうる。
〇がんセンターとの連携は中身のレベルの連携もあるということか？パーキンソンについては、入れ物を共通化する、ということが主目的か？
→癌とパーキンソンでは違うところが多いので、全面的な連携は難しい。データ(のオントロジー)を何段階かさかのぼる必要がある。
〇分子から臨床生活情報までとはどういうことか？
→すでにこういった項目のデータは入っていて、別途解析にも利用している。DB 上で分子から症状までの情報がつながるようなものを想定。
〇医師にとってこの DB に参加することのインセンティブは何か？
→昨年の調査によれば、PJ 終了後に継続してデータメンテしている DB 管理者に聞くと、参加することに結構積極的。メンテのサポートを期待している面もあると思う。
〇医師にとって最終的に患者に役に立つことが重要。DB の大きさと精度が高くなればそれにつながると考えるのでは。

- 東京大学グループ

◇資料説明◇
順調に進んでいる。標準 DB の構築については、既搭載の 500+200+200 検体に加え、1000 検体の解析も進行中、来年度搭載できそう。Case-control DB については、データを産生、あるいは利用する立場のメンバーも参画しており、利用者の意見を盛り込みながら構築。現在 8 疾患公開中。今後も機能追加予定。20 年度新規分として、Copy Number Variation（CNV） DB を構築中。データの受け入れと再配布については、倫理面の検討を実施。特定ゲノムの共有方針を参考に、手続きに必要な文書等も整備中、検討委員会も準備中。ReSequence DB は一部構築完了、一部構築中。倫理面の検討を慎重に進めている。21 年度は CNV 関連の開発、DB 全体の拡張、預け入れ･再配布サービスの開始、海外機関との連携、ReSequence DB 拡張を予定。
◆質疑応答◆
〇GeMDBJ データはどのような使われ方をしているか?
→GWAS データがあるが、計算済みの情報だけを実験情報と共に提供。表示がなされないので研究者としては見にくいものであるため、こちらで見やすい形で提供。倫理規定の関係で、頻度情報から計算できるものを表示している。
〇5 人ぐらいまでしか特定できない項目、さらに個人を特定できる項目、というのは具体的に何か？
→Case By Case。一人の患者しか持っていない特定の遺伝子変異は出せない。研究の場合は申請に基づき利用可能な場合も。
〇付随情報や実験計画などの情報を合わせて５人くらいに絞り込めるかどうかの判断。
〇絞込みができるのは研究者ですね。
〇データの受け取りには、特定はしないということを誓約書などで規定する必要あり

○海外との連携の場合、やりとりするのは頻度情報？
→これも Case By Case。相手が必要とするものを、手続きを踏んでということ。
○学会での動きは？
→学会レベルでは動きはないが、学会誌の投稿規定として、公的な DB にデポジットすることを要求している。海外でも同様の動き。DB に入れなくてもリクエストされたら提供することは必要。本 DB の構築についても論文として発表し、アピールしたい。
○パーキンソンのリシークエンスは将来的には阪大との連携は可能か？
→可能ではないか？臨床情報などでつなげるのでは。臨床情報はあまり多くないが。
○どのくらいの遺伝子についてどのくらい DNA データ量があるのか？
→遺伝子数にしたら 30 くらい。一般的にはもっと少ない。人種によっても異なる。
○CNV についてのデータ量は？
→現在、200 検体。今後の 1,000 検体の結果も入れていける。アルゴリズムによって結果が大きく違うので、その表現の仕方については要検討。比較検討が必要。
○まだ確立していない段階でコントロールデータを作るというのはまだ早いのでは？
→そうかもしれない。ただ、ノイズ・エラーが少なく、あとから解析しても問題が少ないだろうという自信の持てる方式を採用すれば、一応構築できる。他の DB と解析手法では共通している部分もある。
○新型シークエンサーのデータについては将来対応を検討しているのか?
→GWAS データは、数があることが必要。新型シークエンサーはまだ多数の検体を一度に解析できるという段階ではない。少数のリシークエンスのデータは拡張の範囲で対応できると思うが、遺伝病を扱うことから倫理的な問題のほうが大きい。
○現在は神経系の疾患がメインのターゲット？他の疾患への広がりは？
→広がっていくと思う。あとはコストと解析パワーの問題。

- 九州大学

◇資料説明◇
GWAS の世界的な DB である dbGaP では方法論が明示されている。解析について様々な手法がある現状では、このようにどの手法を用いた結果であるかということを明確にすることが重要。東大 Gr の DB を国際標準とするためには手法等を明確にする必要があると考え、データの信頼性を高めるための検証を促進する QC パイプラインを提唱(提案 1)。手法としては、世界的に使われているものを使う。（先行研究の GAIN Project での解析フローを紹介。QC プロトコールとしては GAINQC と PLINKQC を紹介。）QC の部分をデータの生産者/DB 構築者とは異なる第三者に実施させる。その際、第三者は、自身の研究にそれらのデータを用いない事を明記しておくなどルールを明確に。モデルデータを用いて、構築した QC パイプラインの結果･出力等を提示。提案２として、日本人のハプロタイプ DB を構築したい。これは HapMap で提供されているアジア人のデータが他の人種のデータに比べてエラーが多いと予想されるためで、それを解消するためのデータを取得するもの。
◆質疑応答◆
○QC 担当者は QC 以外にはデータを使わないということか？

→Yes。QC に出すか出さないかもデータ生産者の自由。結果は公表前に生産者に通知。
○現状では QC のためにデータを外部に出すことについて倫理審査を通すことは困難。Log は出していないが PLINK は既に用いている。そのほかの QC 自体は必要だが、むやみにデータの足切りをしてしまうと２次解析に支障が出る可能性があるのでは？
→データがどういったものかを明示することが重要であると考えている。理想としては、データ生産者と DB 構築者は切り離すべき。
○東大 Gr では PLINK を含めて QC はしているが、実施していることについて明示していないので、その点は改善したい。
○PLINK はどのレベルのデータに効いてくるのか？
→Genotype を出すところの QC でプラットフォームに依存しない部分。
○生データの QC も必要ではないか？
→そこまでやる必要はないと考える。
○東大 Gr でも QC を行っているとすると、さらに外部でやる必要がどれだけあるのか?
→データ生産者以外の研究者がチェックすることが重要だと提案している。
○提案２については新規提案か？東大 G との相乗効果は？
○東大 Gr のサイトからデータを呼び出すことができるので、有効。
○３０万人 PJ ではこのようなことは？
→理研は理研の中で独自に DB 作成。それとの連携はちょっと先の話。東大 Gr は理研以外の部分をほぼすべて担当している。
○QC に関する提案について本日は議論を詰められないので、継続審議。

- 京都大学

◇資料説明◇
検索機能については、中核で実施している内容とは重ならないようにするため、化合物に特徴的な情報についての検索機能を実装。DB 内容の拡充については、無料で使用できる化合物 DB を利用。医薬品については、別途拡充。そのほか化合物関連の DB(DailyMed, LipidBank, LigandBox)についても拡充。医薬品 DB については公開以来、順調にユーザー数・アクセス数ともに増加。検索に必要なデータについては、オリジナルデータサイトから取得できるものと、こちらで関連性を抽出して作成しているものがある。キーワードで一緒に検索できるだけでなく、データベース間であらかじめ同じものを示すエントリに関連性をつけておいて、一緒に見ることができるように開発している（LinkDB)。構造からの検索について、類似・部分検索から光学異性を考慮する検索へ改良予定。酵素反応関連では、反応ネットワーク予測を実装中だが、今後は、ゲノム情報との関連検索などを盛り込む予定。今後は、開発してきたいろいろな機能をまとめ、ゲノムネット科学情報データベースという形で公開を目指す。
◆質疑応答◆
○カテゴリー3 データについてはキーワードをクローリング等で取得しているのか?
→ftp があるものはそこから。化合物に関して現在クローリングは実施していない。
○ユーザーが順調に増えているが、その分析は？

→国内の薬学系の大学が多い。製薬メーカーも国内外から来ているが、アカデミックが主。使用目的については不明。中核での解析のためにデータを渡すことも可能。
○具体的なユーザーからの反響については？
→ゲノムネット全体のフィードバックは取れるが、化合物に特化したものは無い。内容への感想は取れていないので、個別に取ったほうがいいかもしれない。
○LipidBank は糖鎖とも共通だが？
→7000 件くらいデータがあるが、更新は止まっているので、特に手間をかけていない。
○糖鎖 DB においては、Lipid には糖鎖が付いているのでリンクをつけている。
→化合物 DB としては反応産物などに Lipid が多いので取り込んでいる。
○キーワード検索については今後拡張？
→新たに盛り込む予定の DB については盛り込む。
○ほかに候補となる DB はあるのか？
→小さいものはあるが、DB 拡張の有効性を考えてどこまで拡張するか、の問題。有機系化合物 DB はあるが、有料のものも多いので。
○キーワードの更新は？
→ftp しているので、相手先が更新されたら実施。
○厚労省などでも化合物の毒性などについてまとめられているものもあると思うが？
→環境物質については拡張の可能性はある。他にも可能性のあるものもあると思う。
○化合物の中に糖鎖は入っているか？
→構造データが入っている。KEGG Glycan は単糖が中心。糖鎖 Gr でやられているものとの連携ができればそのほうがいいと思う。
○外部から連携するためのしくみ(API 等)は？
→公開はしていないが、仕組みとしてはすでにある。

- 産業技術総合研究所糖鎖医工学研究センター

◇資料説明◇
糖鎖に関しては構造解析技術や合成技術についての世界標準が無く、それぞれの研究者･機関がデータを独自に取っている。その意味で、糖鎖 Gr の構築している DB は独自性が強い。公開している主な DB は以下のとおり。
GlycoProtDB(糖タンパク質の糖鎖結合部位 DB。現在は線虫のみ公開、今後ヒトも）
GGdb（糖鎖関連遺伝子についてのサマリーDB）
LfDB（レクチンと糖鎖の結合情報ＤＢ。ヒト未同定遺伝子についても発現･検証中）
GMDB（糖鎖構造と Mass データの関係 DB）
これらの DB についてすでに横断検索を実装済み。他に複数の関連 DB も導入済み。
糖鎖に関しては 8 割くらいのデータを産総研で保持しているので、これらを中心にして、今後も協力してもらえるところから糖鎖関連の情報を集めて完成を目指す。
今後 2 年については、1）糖鎖の研究者を支援する DB の構築→新しい DB のためのデータの生産、2）糖鎖研究者以外が利用できる DB の構築　を目指す。

21 年度は、糖鎖コンフォメーション、糖鎖合成、ノックダウンによるフェノタイプについての DB をすでに搭載できる予定。このほかにも打診中。横断検索をさらに進めた統合検索を考え、20 年度もそのための準備を進めている。外部からの登録も可能にしたい。糖鎖研究者以外の研究者向けとしては、読み物 DB からはじめる。DB 全体の仕組みとしては、各機関の DB はそれぞれの機関で維持しつつも、糖鎖 Gr で統合検索を可能にするものを検討し、そのためのインフラや辞書などのツールを整えていく。全体としては 2 年かけてやる内容を紹介した。

◆質疑応答◆

〇読み物とかプロトコールについてはオリジナルに作成するのか？何処かにあるものをキュレーションして集めるのか?

→両方。一部は製品メーカーのウェブサイトにある(要交渉) 。そうでないものは専門家による書き起こし。半分以上書き起こしになると思う。

〇GlycoForum については？

→読み物があるので、集めたい。遺伝子名等の書き方の違い等をまとめる必要あり。

〇統合化や、テキストマイニングの手法等については DBCLS にもノウハウがあるので、分担などを今後相談・検討しよう。

〇提案内容が予算配分を超えていると思うが、予算に収めるとすればどの範囲？

→参加機関の数で調整。打診中の相手が不参加の場合、産総研で対応する部分もある。

→人件費が主で、連携のための作業などに必要。計算機費用の部分が変わってくる。

〇経産省の統合 DB との関連は？

→経産省側の API については外注費として協力してもらっている。

〇統合化やテキストマイニングの内容をセンターで対応すれば予算の調整が可能か？

→可能。

〇統合検索のイメージは？検索結果を取得して何らかの作業をする？

→Yes。インタフェースを統一して一つの DB のように見えるものを想定。

〇一つの DB にして見せることが可能なくらいに糖鎖関連のデータはコンパクト？

→分野は広いが、ある程度コンパクトな部分もある。たとえば、糖タンパク質の糖鎖構造をキーにまとめることができる。

〇目的によっていろいろな統合のスタイルがあるはずだが、それを産総研で FIX してメニューにするのか、ユーザーの希望する形を作れるようにするのか？

→オリジナリティを出すなら前者だが、面白いのは後者かもしれない。いろいろなパーツを独自に組み合わせることも可能。

→糖鎖の構造は通常切り離して解析されるが、糖鎖結合部位と組み合わせれば分子全体の結晶構造解析の検討もできるなど、研究者としては面白い情報を引き出せる。

〇API の整備については、統合化を目指してサポートする場合、構築は結構難しいのでは？糖鎖という限定があるからできそうなのか？実現性についてはどの程度？

→特に難しいのは構造情報。機関によっては画像情報しか持っていない場合もあるので、使えるデータに変換することが必要。

〇マス（質量分析）データについてはデータ全体をダウンロードするやり方で持つのか？

→企業の権利との兼ね合いで、ピークのイメージデータのみ使える。
○経産省関連では、いろいろな機関の DB や、資金源の異なるものなどあると思うが、そのほかの制約についてはどのように対応するのか？
→制約のあるものについては権利関係の整理をして、クリエイティブコモンズのラベルをつけている。
→NEDO 関係のデータが多い。国内の産業育成が主目的なので、企業の意見が大きく影響する。文科と経産で壁があるので破ってほしいが、経産省としては企業化の可能性のあるもので、すでに企業向けにライセンスしている場合契約上どうしようもない。
○予算についてはほかの機関との打ち合わせも今後あるので、今後協議する。

- まとめ　他

・堀田機構長から、JST との関係、当プロジェクト終了後の体制について、これまでの検討経緯および今後予定されている対応について説明あり。JST との一体化が示唆され、それに向けて準備と初期対応が進んでいる。
・《追加資料配布》平成 21 年 1 月 13 日付けライフサイエンス情報基盤整備作業部会報告書「ライフサイエンスデータベースの統合・維持・運用の在り方」
・高木委員から内閣府タスクフォースについて説明あり。4 省連携、永続化の体制などが論点。
○JST へ移った場合は、継続か否か JST の判断になるのか?
→JST の判断もあるだろうが、文科省の方針もあり、それに沿ったものになるのでは。
○JST の中に新たにセンターができるという理解か?
→そのように報告書には書かれている。しかし具体的な体制は整っているわけではない。
○BIRD の予算はどれくらい？
→16 億くらいだが、大半は DB 構築へのファンドとして使われている。当 PJ でもそのような DB があることが前提なので、これを完全に無くすということは本末転倒。
○国際的に興味をもたれる DB にしていかないと、評価されないのでは?
→当 PJ ではまだ国際的な情報発信が足りないので、その点は進めて認知度をあげたい。

・2009 年 6 月 12 日(金) に成果発表を兼ねたプロジェクトのシンポジウムを予定しているので、ご協力願いたい。

(13:00 終了)