

【日 時】 平成21年5月20日(水) 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】 林健志(九大)、五斗進(京大)、田中博(東京医科歯科大)、下川和郎(東京医科歯科大)、井戸敬介(東京医科歯科大)、古崎晃司(阪大)、徳永勝士(東大)、小池麻子(日立製作所)、新聞陽一(産総研)、鹿内俊秀(産総研)、田中康博(文科省)、大久保公策、高木利久(作業部会主査)  
永井啓一、西川哲夫、川本祥子、畠中秀樹、箕輪真理(以上、DBCLS)

(敬称略・順不同)

## 【議 事】

### 平成21年度業務計画について

議事に入る前に、高木主査から会議の進め方について、各グループ30分の割り当てのうち、発表を10分程度とし議論に多くの時間を割いていつ何が公開されるかに重点を置いてご発表お願いしたい旨、依頼があった。

#### ▶ 東京医科歯科大学グループ

##### ◇資料説明◇

【医科歯科大】20年度成果としては2つあり、①疾患・臨床医科学データベースに特化した統合技術開発に関する研究については、昨年は検索を含めプロトタイプ構築は引き続き完成させるが、そのほかに、症例の入ったDBについての倫理規程案を検討、論点を整理して、倫理規定として予定される項目を洗い出した。オントロジーについては2015年に改定予定のICD11(単なる分類ではなく、オントロジー概念を導入)に田中らが提唱するIntegrated Biomedical DBのモデルがたたき台として採用された。もう一つの成果②疾患データベースの高度化と実証的統合に関する研究について、当初の目標から数を絞ってモデル的にやるよという方針変換があり、そのモデルとして医科歯科・阪大のDBとGeMDBJとを統合検索できるようにした。インターフェース(=IF)では検索条件がみえるようにし、分子情報のクラスタリングまでできるようにした。平成21年度の目標としては、20年度と同じ項目についてそれぞれ進める。肝臓がんについてはすでに公開済み。英語バージョンも作成したので、国際的にも公開予定。一部ユーザ登録が必要ではあるが一部は完全公開。447例症例について登録されている。一般公開に際し、使いやすいものにしていきたい。また、中核機関との連携を進めていく。

【阪大】医科歯科の要件定義に合わせた形で阪大のデータを整形し、提供中。ユーザの意見を取り入れ、IFをグラフ化した。21年度は、技術的な開発に加え、倫理的な検討についても医科歯科と調整しながら、データベースの公開を進めたい。前回の作業部会の意見も取り入れて、具体的な項目を立てた。DBの相互連携、使い勝手の良いDB、オープンアクセスに向けたポリシー変更(統計値はPW臨床なしで公開)を、阪大としては検討している。

##### ◆質疑応答◆

○倫理問題について、2次利用に当たるのではと思うが、その際には匿名化されていれば問題はほとんどクリアされるのではないのか?統合化によって新たな問題が出ているのか?

→症例DBなので、匿名化であれば一応使えることになっているが、今の状態では個人の情報まで見ることが可能になっており、他のDBを参照することである程度調べられるので、その部分が問題。どのくらいの情報に曖昧化すればいいかを検討する必要あり。DBの使用目的としては統計的な使い方がほとんどと思うが、絞り込める可能性があるデータを公開してしまう可能性があるのでは、気をつけたいといけな。

○何らかのルールに基づいて考えなければいけないのでは。他のデータと組み合わせるとわかる個人情報であれば、公開した人には責任はないのでは、という理解である。現状のルールを明らかにして、この制限のもとではこんなことができる、という見解を出してほしい。また、公開しようとしているデータについ

ては、他の人に公開するとどんなメリットがあるのか？

→了解。メリットについては、今までは肝臓がんについては分子のデータのみであったのが、今回のDBでは、発現パターン、コピー数変異(=CNV)の情報もあり、画像なども含まれている。

○公開データを用いて研究ができるか？他のデータと合わせることも可能か？

→ある程度の研究は可能。すでに肝臓がんだけでも200例位入っているが、別にあるデータと合わせることも可能になる。現在は画像中心だが、将来的には項目も増やしていく。カルテから個人情報にかかわる部分をのぞいたものすべてを公開していきたいと考えている。

○退院時のサマリデータの所有者は誰か？診療科か？先生がそれを利用することは大丈夫か？

→診療科ではないかと思う。半分は患者のものではないか。完全に医師の著作権とすると話がおかしくなる。

○同意書に患者のものではないということが書かれているのではないか？

→医師のものであるということはないと思う。医師の持ち物(=作品)であるという考えだとおかしくなる。

○パブリックドメイン(=PD)に置いておくことは問題ないのか？

→(医科歯科では)IDや個人にまつわる情報を削除すれば、個人を特定できないので問題なしと考えている。

→阪大の医師の間では電子カルテなどは病院外に公開するという事は考えられない(機関によっても理解に差があるのではないか)。

○今まで病院内でしか使っていなかったから、問題になっていなかったが、公開するとなると問題？

→CT画像などを見ること、出版することはこれまでも問題なく行われている。それを使って、研究することについては例がないが、集合情報であれば問題ないと考えている。GEOなどでも症例情報と一緒に出しているものを集合情報として扱っている。

○先生がしようと思っっていることについて、問題がないかを確認しながらやってもらえると、その情報が役立つので、それをお願いしたい。PDに置くということは誰でも見えるということなので。

→登録申請なしでも見える情報もあるようにしたい。NCBIのようにOpenにできるかはわからないが。

○PDに置くということは誰でもデータを追加したりできるようにデータを提供するということですね？

→かなり要約した情報であれば、問題ないと思っっているが。たとえばCT画像を全部出す必要はなくて、特徴的な数枚を出せば足りるのではと思う。

○どの程度の情報を出すか、という基準はあるか？量の問題か？

→すべて出すと問題があるかもしれないので、特徴的な画像だけ数枚出すといった対処をする。

○何を出していいのか、何がいけないのかをはっきりしないと、出そうとしても皆がわからないので、そのような情報が必要である。

○匿名化されれば病院の情報を2次利用できるという、臨床規定の改訂(規制緩和)が4月にあったはず。匿名化データについては病院での倫理委員会を通さなくても、利用する側の倫理審査があれば使ってよいとする改訂内容だと思う。

○何が公開できるかできないか、について規定の有無や世界的な状況をもう少しまとめてほしい。公開して価値があるのか、という点について、利用者の想定も含め、説明してほしい。

→自分の関心がある症例について、遺伝子発現やゲノム情報が付随した形で得られるというのがメリット。

○大阪大学についても公開の意義についてはおなじか？東大との連携については？

→パーキンソン病での意義についてはもう一度確認してご連絡する。東大での公開についてはまだ内容を確認していないので、一度確認してから検討したい。

○(東大グループから)症例についての記述が共通にできれば連携可能。

○大阪大学の症例についてもゲノム解析を行うということか。

→ゲノム解析についてはやるかどうかも含め、検討中。

○ユーザ登録については何名くらい？

→200名くらい。もっと広報が必要であると考えている。ユーザ登録がなくても見える部分についても充実したい。

○公開の際にはルールについても明確にして情報提供してほしい。

○複雑な仕組みを作って解析ができるようにするよりは、生データの公開をしたほうがいい。クレジットを確保したうえで、加工や編集やアプリケーション開発は自由にしていい、としたほうが、使われる可能性が高くなる。

▶ 東京大学グループ

◇資料説明◇

資料3の1ページにサマリアがある。それぞれの開発項目について進捗を記載。CNVについては計算方法がいろいろあるので信頼性検討中。リシークエンシングについては倫理検討委員会の決定を待って公開する。倫理（公開方針）検討作業を実施し、6月の検討委員会を待っている状況。（以下、詳細説明）標準DBとしてはすでに900検体分のデータを蓄積。Case Control DBのほうは、8疾患を公開中、専門外の人でも分かりやすいIFで解析結果の有効活用を目指す。CNVについてはアルゴリズムが何種類かあるので、その結果を併記。リシークエンシングデータについては、ALSとパーキンソン病が対象。統合DBとしての共有方針の決定（6月初旬に検討開始）を待って公開を進める。今年度の計画としては、海外との連携については公開共有方針決定待ち。資料には無いが、医科歯科・阪大との連携をこの夏位をめどに作業したい。

◆質疑応答◆

○GAINでは関連解析はデータ生産者とは切り離され、信頼性を高めている。

→外部に出せるデータが統計情報だけなので、元データを外部に出すことができないため、第3者による解析が難しい。

○米国の仕組みでは、生データの提出がタスクとなっている契約。その後その解析をするという別の研究契約があるという形なので。日本ではそうになっていないので、第3者が解析できない点については問題の所在が異なる。

○外部の人には透明性がわからない。透明性の担保のために外部を利用してはどうかという提案をしている。本DBではデータの生産者とは異なる計算結果も出しているが、独自の計算方法を使っているのでは？GAINでは、どのようなツールを使っているかを開示している。

→独自の解析だけではなく、一般的に使われている解析方法も盛り込んでおり、その情報を付ける予定。一部はGenotypeをもらえないデータもあるので、再計算できていない。dbGaPにおいても、データ生産者が計算した結果を載せているものと、dbGaPが解析を委託されたものがある。

○dbGaPは独自に全部を解析しているのでは？

→（独自に解析しているのは）一部だけであって、全部がそうではない。

○論文に出ている結果であれば信頼できるが、論文以前のもを出しては信頼性を損なうのではないか？

→データを出したグループが計算して論文発表した結果を載せている。中にはその元データを提供してくれるグループもあるが、倫理検討委員会での方針の検討を待っている段階でその決定内容に従って公開する。その準備はしている。

○CNVについては公開可能ということだが、手法についてはすでに決定しているのか？手法の実装の際にバグが入る可能性は？いずれにせよ手法についてすべて公開すべき。

→2種類程度の手法を考えている。一部は実装もしているが、アルゴリズムとしては標準的なものを採用。

「標準的な」というのは、よく使われているという意味。どういう手法を使ったのかを記載する。

○客観性とか透明性はDBのところにきちんと方法論について記載すれば良いのか？

○方法論として論文等がアクセプトされてから使うほうが良いのでは？

→アクセプト後に計算結果を公開する。

○方法論についてプロジェクトの中で議論する必要はない。Add Value の部分についてはプロジェクトではないところでやればよい。プロジェクトの成果としては DB を作ってみんなに使ってもらうことで評価の対象としてほしい。

○目指しているところは同じだと思うので、いただいた意見を受けて改善すべきところは改善するというところでお願いする。

➤ 九州大学

◇資料説明◇

GWAS データ解析についてのいろいろな問題点を背景として、昨年度まではコミュニティに客観性を示すためのパイプライン化を実施。今年度は GAINQC より信頼性の高い PLINKQC を採用し、実データでパイプラインを動かす。CNV についても信頼性の高いと思われる Birdsuite への対応を検討する。また、計算の基盤となる日本人のハプロタイプ情報を整備する。

◆質疑応答◆

○提案されている手法については、東大グループでもいろいろ試しているが、別途やることの意義は何か？  
→すでにやられているかどうか記載されていないので、わからない。確立された方法でも複数箇所ですることにより意義がある場合もある。

○GAINQC や PLINKQC はすべてパブリックライセンスか？(東大グループへの質問として)東大で開発した手法も公開していくか？第3者が比較して、東大の解析方法が良ければ評価されるのだから、どんどん公開すべき。

○(東大グループ)用いているプログラムはすべて GNU ライセンス。東大での解析方法も公開する。いろいろな計算結果を出しておき、信頼性の高いほうをユーザが選べるようにするべきと考えている。実際にはユーザが計算方法も選べるようになっている。現在はアルゴリズムやプログラムについての記述が足りないと思うので、追加する。また、ポリシーを明確にしていく。一方で、九大の日本人ハプロタイプの結果も統合 DB に盛り込めればユニークになっていくので、是非使わせてもらいたい。

○昨年度の成果として、パイプラインの構築と DB の開示があったと思うが、具体的なユーザは？

→限定公開の場所は作ったがまだユーザがいない。パイプライン自体は公開していないが、パイプラインしかないのでは、(サンプル以外の) データは入っていない。

○成果の公開の観点から、公開されているものやどういうことをやったのかを記述してほしい。どういうものかの記述を示してほしい。

→東大グループからのサンプルデータをつかって計算した結果だけだが、見られるようにすることは可能。一般に公開してもいい。

○公開の方法やアナウンスのやり方等も含め、今後ご相談させていただく。

➤ 京都大学

◇資料説明◇

共通基盤開発と統合 DB 開発・運用を推進中。今年度開発としては、統合 DB 開発：ゲノムネット医薬品 DB の高機能化としてキーワード検索の改良、新規 DB の追加、統合 DB 内の他の DB との連携を予定。共通基盤技術開発：構造情報、光学異性を考慮した検索、ネットワーク予測の取り込み等を予定。化合物関係の DB や技術を統合して、「ゲノムネット科学情報データベース」の構築を目指す。

◆質疑応答◆

○利用統計に関して、20000 名くらいというのはこの分野のユーザはこのくらいの規模か？内訳は？

→規模的にはこのくらいだと思う。内訳はアカデミアより民間のほうが多い。製薬会社のドメインも多い。

○いろいろな情報源を持ってきて整形して同じように見せる仕組みということか？

→検索だけできるようになっており、元データは元の DB に飛んでいく。インデックスを持っているだけ。

LinkDB のしくみも ID のみの関連付けなので、基本はリンク先に飛んでいくだけ。

○構造まで検索できるのは、そのような情報を提供している DB があるからか？

→KNApSAcK については情報が提供されており、PubChem は API で取得できる。

○案内サービスであるので、利用規約などの問題は生じないということか？

→そうである。

○インデックス更新は？

→定期的に元サイトを見に行き、更新があれば更新するしくみになっている。

○LinkDB はかなり人手を入れて作っているのか？

→人手のかかるところと機械的なところがある。LipidBank など構造情報を入れ直している。

○Semantic の勉強会で、人手をかけずに RDF(Resource Description Framework)を書けば自動的に LinkDB ができるといった話もあるが。

→もとの DB が RDF 対応になっていないと可能ではないと思う。

○化合物、遺伝子と副作用をつなぎたいという人も多いが、この DB の中で遺伝子情報はどのくらいつながっているか？

→ダイレクトではないが、化合物→反応→遺伝子というつながり方はできる。KEGG のほうでもターゲットについての情報を付けている。情報ソースは、教科書、論文（総説）、また、JAPIC から一部引用している。

○高度なキーワード検索とはどのようなものか？

→たとえば、DB 毎に検索するのではなく LinkDB を使って関連 DB を一括検索するといった仕組みのこと。

## ➤ 産業総合研究所糖鎖医工学研究センター

### ◇資料説明◇

日本に散在する糖鎖関連のデータベースをなるべく多く集めての統合するために、まずは産総研 DB を中心に基盤整備を実施しながら横断検索できるようにした。検索にはキーワード、構造(一部)を用いることができる。

開発段階として3つのフェーズを考えているが、残る2年で統合検索のプロトタイプを作りながら、国内のほかの DB についての API も整備していく。具体的には、新しい協力機関への呼びかけ、糖鎖 DB の普及、他機関が持つ実験データとの連携、オンラインプロトコール整備と糖鎖合成のための DB(方法論の収集など)構築、(糖鎖関連)疾患と遺伝子の関連 DB 整備、糖鎖機能の情報発信、技術的にはインデックス更新システム構築を行っていく。また、糖鎖に関する用語の整備を行い、統合検索プロトを開発する。

### ◆質疑応答◆

○構造検索プログラムは自前で作ったのか？標準的なものを使用したのか？ソフトウェアの信頼性は？

→糖鎖構造の表記は産総研で XML 記述を定義されたものを採用し、論文化されている検索手法をもちいている。今回は高速化と国際的な標準への適用を実施した。

○オンラインプロトコールは何か本から転載するのか？

→書き下ろしで外部への依頼も含めて作成する。利用転用も自由にしたい。ウェット実験のためのプロトコール。糖鎖の扱いの入門版という位置づけ。

○蛋白質科学会が蛋白質の実験プロトコールで同じような試みをしている。

○それは転載等自由か？確認してほしい。

○(糖鎖関連)疾患と遺伝子の関連 DB について、OMIM とはどこが違うのか？

→国内で情報源となっている個別のサイトなどと連携して、情報源を明らかにした形で作るという点。

○産総研に KO マウスの DB はあるのか？

→入れ物を産総研で作り、外部の先生が入力したもの。産総研の DB とリンクしている。

○今入っていない糖鎖関連の DB というのはあるか？ほとんど網羅したと置いていいか？

→北大の DB が入っていないくらいで、他の公開 DB はかなり網羅。

○ユーザ数は？

→独立 IP で 2000 位（ポータルのみ、詳細は報告書にあるので参照されたい）、年間で 20000 位だと思う。

○スペクトル情報を含めた DB がすでにあると思うが、それらとの相互参照はできないか？

→一部、MS スペクトルの DB は参照できるようになっている。NMR などはない。

○経産省と文科省の区別は？

→経産省側ではヒトを対象とした DB が主。経産省側で作成した API は利用している。

○経産省の作業については何らかの指示が産総研グループに来るのか？

→特に作業としてやっているわけではなく、先方の外注先に直接指示が出ているようだ。

○産総研では Anonymous FTP サイトを作れないと聞いたが。

→その点については判らないが、産総研内の LAN を使うとやりにくい部分があり、JCGGDB についても外のドメインを使っている。おそらく産総研内のポリシーの問題だと思う。

#### ➤ 総合討論

すべて個別の発表に続く議論の中で終了したので、特になし。

最後に高木主査より、1)前回の議事録確認依頼、2)シンポジウム(6月12日)への協力御礼と依頼があり、会を終了した。

(13:00 終了)