資料１
科学技術・学術審議会学術分科会
学術情報委員会（第３回）
平成２７年６月２４日（水）

ライフサイエンス分野における研究データの共有について

東京大学大学院理学系研究科
科学技術振興機構バイオサイエンスデータベースセンター
国立遺伝学研究所DDBJセンター

高木利久

ライフサイエンスに係るデータの実態（1）

- データベースの数
  - 世界全体:10,000から20,000
  - メジャーなもの（NAR誌のDB特集収録）:約1,600
  - 我が国のDB数(NBDC Integbioカタログ）:約1,000
- データの種類
  - NAR誌での分類：15カテゴリ、40サブカテゴリ
  - Integbioカタログ:生物種、対象、データ種類で分類
    - 生物種：動物、植物、原生生物、菌類、真正細菌、ウイルス
    - 対象：ゲノム、遺伝子、cDNA、多型、タンパク質、酵素、細胞
    - データ種類：配列、構造、発現、相互作用、画像、オントロジー
- ゲノムプロジェクトの数(GOLD)：約64,000
- データDB開発国(NAR誌）：約50

ライフサイエンスに係るデータの実態（2）

- データのサイズ
  - 米国NCBI: SRA 4PB, dbGaP 2PB
  - 年率 1.5倍程度の伸び（ムーアの法則と同じ程度）
  - 文献(PuBMed)は2,500万件
- 主要なDBセンター：NCBI, EBI, DDBJ、等
  - ゲノムのデータが中心
  - 欧米のセンターは数百人規模
  - 10～30peta程度のストレージ保有
  - 2020年にはゲノムだけで2EB程度必要との試算

主要なDBセンター

	日本			米国	欧州	中国
	ROIS		NBDC/JST	NCBI	EBI	BGI
	DBCLS	DDBJ				
組織形態	ライフサイエンス分野におけるデータベースの利便性や付加価値の向上に関する研究開発を担う我が国唯一の機関	機構傘下の国立遺伝学研究所の付属施設 「生命情報学」の我が国における研究拠点 我が国を代表するDNAデータベースを運営	DB基盤技術と分野別統合化の委託機関を公募し、ライフサイエンスデータベース統合推進事業を推進 研究部門と事務局で構成	NIH傘下のNLMの付属機関 分子生物学分野を支援するソフトの提供と計算機を利用した基礎研究機関	EMBL傘下の非営利学術機関 バイオインフォマティクスの研究とサービスの中心機関	ヒトや動植物、微生物のゲノム解析研究を手がけるDNA解析研究機関
組織の永続性	予算の9割近くをNBDCからの時限付委託費により運営	国立遺伝学研究所の運営費交付金により運営	JST の運営費交付金（ライフサイエンスデータベース統合推進事業）により運営	根拠法：Public Law 100-607	費用の半分は20か国の公的研究資金で運営されるEMBLから提供 残りは、ウェルカム財団、 NIH 、UK Research Councilsの資金等	中国科学院より施設及び設立資金を提供

統合の考え方、意義、効果

- 少数の数式や法則で表現できない、データはインフラかつフロンティア
- 小規模データからビッグデータへ、ビッグデータから知識へ
- 統計解析のパワーアップ、データの価値の最大化
- 他の観点からのデータ活用、イノベーション促進
- 研究成果の再現性や検証、研究不正への対応
- データ収集の重複の排除、失敗データ活用、研究（資金）の効率化

- 文献データの共有は1960年代より
- 研究データの共有は1970年代より
- ヒトゲノムプロジェクトのバミューダ原則（研究コミュニティ）
- アルツハイマー研究ADNIの全データ共有ルール（研究コミュニティ）
  - Sharing of Data Leads to Progress on Alzheimer's　NY Times Aug12, 2010
- 米国NIH NCBIのヒトの制限アクセスDB(dbGaP)
  - 2007年以降約30,000件の利用申請（許可されたもの約20,000件）
  - ４1カ国からの利用申請
  - データの2次利用によって約920報の論文出版

米国NIHの国立医学図書館(NLM)改革ビジョン

（1）アクセシブルかつ信頼性の高い生物医学に関する研究結果及び信用性のある保健衛生情報を収集し、一般大衆、医療専門家、世界中の研究者に普及させる上での主導的立場を取り、常に進化し続けなければならない

（2）生物医学の情報と透明性をもった分析は公共財であるという概念を普及するよう努め、オープンサイエンス、データ共有、研究の再生産等を支援する取組みを先導して行うべきである

（3）NIHにおけるデータ科学に関する知識やプログラムの中心であるべきであり、生物医学に関する研究等を通じて、NIHの振興を図るべきである

（4）継続的かつ集中的な研修を通じ、生物医学情報学やデータ科学、図書館学その他の関連分野における次世代の研究者を育成し、NLMの役割を強めるべきである

（5）長期利用を可能とするために、米国の歴史を通じた、生物医学の研究や医学の向上に関する業績を維持管理、保存し、それらをアクセシブルにするべきである

（6）NLMの首脳部は、NLMがその使命を十分に果たし、その資源を最善の形で配置するために、どのような才能や資源、組織構造が必要であるか評価する必要がある

国立国会図書館の情報ポータルより
http://current.ndl.go.jp/node/28682

NIH approves strategic vision to transform National Library of Medicine
http://www.nih.gov/news/health/jun2015/od-11.htm

データの利用に関する障害

- 自分の専門外のDBを使う必要性あり
- DBや解析ツールの数が多すぎて使い方分からない
  - 生体内相互作用DBだけでも500以上のDB
- 注釈が信頼性のあるものとないものが混在
- フォーマットや用語がバラバラ
  - 遺伝子の概念さえDBによって違う
  - 同じ遺伝子にも多数の名前あり
- データの文脈依存性、曖昧性、冗長性、複雑性、誤差
- 単にレポジトリするだけでは再利用性低い

我が国の事情

- 資金配分機関からの共有の義務化ルールなし
- データの囲い込み（公開されないデータも多数）
- データの権利関係不明
- バイオインフォマティシャンの不足→競争に負ける
- 受け皿となる中核DBセンターがない
- プロジェクト終了するとデータの維持管理更新されない

我が国のライフサイエンスDB統合推進事業

- データの共有、公共財化を促進し、その価値を最大化
- 内閣府CSTP主導の統合データベースプロジェクト（H18～）
  - 文科省、経産省、農水省、厚労省で実施
  - H23年12月に四省連携のポータルサイト
- 文科省の統合データベースプロジェクト(H18～）
  - 中核センターの設立
    - H19～情報・システム研究機構ライフサイエンス統合DBセンターDBCLS
    - H23～科学技術振興機構バイオサイエンスDBセンターNBDC
  - クリエイティブコモンズ(CC)ライセンスによるデータの共有
  - フォーマット、辞書、統合技術、動画教材などの開発
  - カタログ、横断検索、アーカイブの構築など種々のサービス提供
  - 研究分野ごとのデータベース統合化進行中（ファンディングによる）
  - ヒト由来データの共有・セキュリティガイドラインの作成
  - ヒトDB（オープン、制限アクセス）の構築、受入れ(DDBJと連携して）

収集・共有すべきデータの範囲（1）

機関名	対象データ
OSTP （米国大統領行政府　科学技術政策局）	政府機関が全部あるいは部分的に助成する機密扱いに区分されない研究から得られるデジタル形式の科学データ。 デジタルデータの定義： ・学術論文の裏付けとなるデータセットなど研究結果を立証するのに必要な科学界で共通に受け入れられるデジタル的に記録された事実に基づくデータ ・実験ノート、準備分析、論文の原案、将来の計画、査読報告書、同僚との通信、物理的対象物は含まない
NIH （米国国立衛生研究所）	・研究を目的とする最終研究データ（科学コミュニティが研究結果を文書化しサポートするのに共通に必要とする記録された事実。サマリー表などではなくそれの元となるデータ） ・NIHが助成する基礎研究、臨床研究、調査等。特に複製が不可能なユニークデータ（莫大な費用がかかるため複製できない大規模調査、自然災害、稀な人口群に関する調査など）については特に重要 ・助成期間中、年＄500,000 以上の直接経費を要求する申請 ・2003 年 10 月以降の申請
NSF （米国立科学財団）	・NSF 助成により得られたすべての重要な結果（findings） ・NSF 助成により得られる一次データ、サンプル、物理的な収集物、その他作成、収集したサポート材料。 ・NSF 助成により作成されるソフトウェアや発明の共有、それらあるいはそれらにより得られる製品が広く利用できるようにすることを奨励する

JST科学技術情報委員会「わが国におけるデータシェアリングのあり方に関する提言」（平成27年4月）
別添資料1「ステークホルダーにおけるデータシェアリングの動向」より

収集・共有すべきデータの範囲（2）

DOE （米国エネルギー省）	・機密扱いに分類されない、制限のない、助成研究成果を立証するのに必要なデジタル・データ。 ・デジタル研究データの定義：実験、観測、シミュレーションデータ；コード、ソフトウェア、アルゴリズム；テキスト；数字情報；画像；ビデオ；音声；関係するメタデータ。ローデータ、処理されたデータ、分析データ、公開データ、アーカイブデータなど多様な形式の情報を含む。 *データの収集が実際に行われていない要素については、連邦諮問委員会や公示等の方法により、DOE としてどのようなデータを収集するのが適しているかを判断する。
BBSRC （英国バイオテクノロジー・生物科学研究会議）	・データが論文で使用されているか否かに関わらず助成研究から発生するすべてのデータ ・特に重要と考えられる分野： ①多量的実験から発生するデータ ②長期的なまたは累積的アプローチにより発生する低情報量データ ③システムアプローチから発生するモデル
EC HORIZON2020	研究データの定義：「研究データ」は、検討・考察のために収集され、推論、議論、または計算のベースとなる、特に事実や数字を意味する。研究のコンテキストでは、データの例として統計、実験結果、測定値、フィールドワークからの観察、調査結果、インタビューの録音や画像が挙げられる。焦点はデジタル形式で提供されている研究データである。 Open Research Data Pilot（※）で適用されるデータ （1）論文公開時の結果を検証するのに必要なデータ（メタデータを含む）。 （2）"データ管理計画"に記載するその他データ（メタデータを含む）。

国際連携の枠組みと活動主体

- INSDC(国際塩基配列DB)
  - 米国NCBI, 欧州EBI,日本DDBJ
- dbGaP/EGA/JGA(ヒトゲノムDB)
  - 同上
- wwPDB(タンパク質立体構造)
  - 米国BMRB, RCSB PDB, 欧州PDBe, 日本PDBj
- HUPO(ヒトプロテオーム機構)
  - 約20カ国
- ADNI(Alzheimer's Disease Neuroimaging Initiative)
  - 米国発祥、その後、欧州、日本、オーストラリア、台湾、なども
- GA4GH(Global Alliance for Genomics and Health)
  - 33カ国、330機関
  - アマゾン、グーグル、マイクロソフト、イルミナなども

データの質や利活用を保証するためのルール

- データの質保証
  - おもにメタデータのチェック
  - なかにはデータそのもののチェックも
    - 例えば、INSDCでは、データの整合性確認、補正、データ重複の確認。データに記載されている生物名と配列自体から予測される系統学的位置との矛盾を検証、など
    - dbGaP では性別などの表現型と遺伝型間の整合性を検証。別々のプロジェクトに登録されている参加者が同一人物かどうかゲノムデータから判定、など
    - 遺伝子発現DB(GEO) ではアレイシグナル強度のばらつきを検証、など
    - PDBはタンパク質としての構造上の制約を満たすか、などの検証

- 利活用のためのライセンス
  - CC-BYやCC0のライセンス付与
    - ChEMBL（RDFをCC-SA-BY)、Uniprot（CC-BY-ND）
    - CC0：BioMed Central、Nature Publishing Group linked data

INSDCにおけるデータ利用のルール

GenBank Data Usage

The GenBank database is designed to provide and encourage access within the scientific community to the most up to date and comprehensive DNA sequence information. Therefore, NCBI places no restrictions on the use or distribution of the GenBank data. However, some submitters may claim patent, copyright, or other intellectual property rights in all or a portion of the data they have submitted. NCBI is not in a position to assess the validity of such claims, and therefore cannot provide comment or unrestricted permission concerning the use, copying, or distribution of the information contained in GenBank.

DDBJの記載

DDBJデータ内の配列データには特許の認められている配列は含まれますが，利用や再配布を制限する著作権はありません。しかし それぞれのエントリ内には著作に近い表現があります。（中略）INSDCデータは（中略）個別の登録者から著作権譲渡を受けたものではありません。DDBJ は GenBank や EMBL-Bank 同様なんらの利用制限も付加しませんが，それぞれのエントリの中に登録者の著作が含まれる可能性があり，全体のコピー改変再配布について DDBJ として明言することができませんでした。これについて 2002 年に国際諮問委員会により利用制限について助言をいただき，その後は登録される際に著者の方々に助言された以下の方針を一読いただいています。「INSD は，公開データにその利用を制限するような記述ならびに，このデータを利用した出版物を禁止するような制限事項は付記しない。特に，公開されたいかなる配列データにも利用制限や利用許可取得義務を設けず，公開データの二次公開や公開データベースの利用についても利用制限や利用許可取得義務を課さない。」

データ提供者の義務とインセンティブ付与

- 出版社
  - 論文投稿前の公的DB登録の義務化
- 資金配分機関
  - 研究申請時にデータマネージメントプラン提出
  - 論文公開の義務化
  - データ共有の義務化
- インセンティブ
  - INSDC では登録者の氏名と所属組織をデータに記載しているので、評価対象として検索可能
  - データにDOI付与の動き加速

学術論文の公開とエビデンスデータの公開

- PubMed Central(PMC)
  - 米国NIHが運営、2008年より義務化
- BioMed Central(BMC)
  - シュプリンガーが運営、投稿者費用負担モデル
- Public Library of Science(PLOS)
  - Non profit publisher、投稿者費用負担モデル

- 資金配分機関および出版社が義務化
  - 公的DBやコミュニティDBに登録公開を義務化

公開の意義と公開すべきエビデンスデータの範囲

- 資金配分機関：研究の成果を最大限活用するため
  - NIH：資金援助した研究の結果および成果を、研究者コミュニにティや一般社会が最大限利用できるようにする
  - Wellcome Trust：資金援助した研究から出る成果の価値を最大化するため
- 出版社：発表論文に関するエビデンスの確保
  - PLOS系列：論文作成に関係する全データを制限なしで公開
  - Nature系列：論文Submit時にはEditorsやReviewersが見られるように
  - Science系列：論文中のデータを作成したデータを自由に見ることができるように

Natureの規定

Mandates for specific datasets

For the following types of data set, submission to a community-endorsed, public repository is mandatory. Accession numbers must be provided in the paper. Examples of appropriate public repositories are listed below.

Mandatory deposition	Suitable repositories
Protein sequences	Uniprot
DNA and RNA sequences	Genbank
	DNA DataBank of Japan (DDBJ)
	EMBL Nucleotide Sequence Database (ENA)
DNA and RNA sequencing data	NCBI Trace Archive
	NCBI Sequence Read Archive (SRA)
Genetic polymorphisms	dbSNP
	dbVar
	European Variation Archive (EVA)
Linked genotype and phenotype data	dbGAP
	The European Genome-phenome Archive (EGA)
Macromolecular structure	Worldwide Protein Data Bank (wwPDB)
	Biological Magnetic Resonance Data Bank (BMRB)
	Electron Microscopy Data Bank (EMDB)
Microarray data (must be MIAME compliant)	Gene Expression Omnibus (GEO)
	ArrayExpress
Crystallographic data for small molecules	Cambridge Structural Database

http://www.nature.com/authors/policies/availability.htmlより

SCIENTIFIC DATA

Home | Archive | About ▼ | For Authors ▼ | For Referees | Data Policies ▼ | Collections ▼

Home ▸ Data Policies ▸ Recommended Repositories

Recommended Data Repositories

Scientific Data mandates the release of datasets accompanying our manuscripts, but we do not ourselves host data. Instead, we encourage submission of datasets to community-recognized repositories where possible, or to general-science repositories if no community resource is available. Repositories included on this page have been evaluated to ensure that they meet our requirements for data access, preservation and stability. Please be aware, however, that some repositories on this page may only accept data depositions from those with specific funding, or may charge for deposition of data. Please ensure you are aware of any deposition policies for your chosen repository. If your repository of choice is not listed please see our guidelines for suggesting additional repositories.

Authors must deposit their data in an approved data repository as part of the manuscript submission process; manuscripts will not otherwise be sent to review. We may recommend temporary

http://www.nature.com/sdata/data-policies/repositoriesより

公開、非公開の判断基準など

- 制限アクセスデータと非制限アクセスデータ
  - パーソナルゲノムなど個人同定可能なものは制限アクセス
  - データアクセス委員会で利用申請を判断
- 公開の時期
  - NIH GDS(genomic data sharing) policyではヒトデータは 品質管理が済み次第dbGaPに登録。非公開期間は最大6か月。エンバーゴは設定できない。
  - 非ヒトデータは論文公開まで非公開が認められている。
- 臨床試験データや臨床データなど新たな共有に向けた動きも
  - そのためのデータアクセス委員会の機能強化の動き

データの保管・公開にかかわる役割分担

- データの保管、整備、運用
  - DBセンター：米国NCBI,欧州 EBI, 日本DDBJ,NBDC
  - 研究機関：分野毎のDB構築、公開
    - 米国NCI、JGI、他
  - 出版社：レポジトリ提供
  - 学会：自前のDBや推奨DBもつものあり
- データ共有、公開のルール作り
  - 資金配分機関
    - 米国NIH、英国Wellcome Trustなどガイドライン策定
    - 制限アクセスデータのアクセス委員会（研究機関がもつ場合も）
  - 出版社
    - エビデンスデータの公的DBやコミュニティDBへの登録義務化

公募要領にデータ提供協力依頼記載

- 文科省ライフ課委託プロジェクト(H20～）
- JST戦略事業（CREST、さきがけ）(H23～）
- 厚労科研費(H24～）
- 文科省科研費(H25～）
- AMED-CREST, PRIME(H27～）
- 医療分野研究成果展開事業
  産学連携医療イノベーション創出プログラム(H27～）
- ナショナルバイオリソースプロジェクト
  「ゲノム情報等整備プログラム」(H27～）

6.8　バイオサイエンスデータベースセンターへの協力

ライフサイエンス分野の本事業実施者は、論文発表等で公表された成果に関わる生データの複製物、又は構築した公開用データベースの複製物を、バイオサイエンスデータベースセンター(※)に提供くださるようご協力をお願いします。提供された複製物は、非独占的に複製・改変その他必要な形で利用できるものとします。複製物の提供を受けた機関の求めに応じ、複製物を利用するに当たって必要となる情報の提供にもご協力をお願いすることがあります。