3M4-OS-07e-1

# データ・キュレーションの精度評価とタスク設計 ～ゲノム分野の事例～

## Evaluation and Task Planning of Data Curation in Genomic Studies

神沼英里[*1] 藤澤貴智[*1] 中村保一[*1]
Eli Kaminuma Takatomo Fujisawa Yasukazu Nakamura

[*1] 国立遺伝学研究所 生命情報研究センター
Center for Information Biology, National Institute of Genetics

The scale of genome sequencing data has grown rapidly by technological innovation in life science studies. Consequently data curation for genomic annotations by manual operation tends to be omitted due to high cost. We have developed an automatic annotation system of large-scale genomic data and a community annotation/curation system for online journal texts. In this report, we list concrete issues on genomic data curation out of the experiences, and propose "curation evaluation" and "optimal work assignment" in a curation task design.

## 1. はじめに

生命科学分野の技術革新により、ゲノム解読のコストが急激に下がっている。2005 年の米国 454 Life Sciences 社(現 Roche 社)の塩基配列解読の新技術製品のリリースを皮切りに、高速 DNA シークエンサの開発競争がが始まった。これより 2013 年の現時点で、ヒトゲノム解析のコストは数時間 10 万円まで下がっている。この DNA 解読のコスト低減により、ゲノム解析を一部の研究機関が実施する時代は終わり、医療の臨床検査や、農業の生産管理で日常的に利用される様になっている。また屋内から出て屋外現場での DNA 解析という汎化も始まっている。試料を採取した場所で、DNA を簡易解析する製品も既に発売されている。今後は、非専門家が屋外で試料を収集し、これまでに知られていなかった新規の生物を発見する事例も出てくると予想される。

一方、DNA の配列解読コストは下がったものの、配列を解釈する為のの情報解析コストは高止まりと言われている。我々バイオインフォマティクスの研究者には、日常的に高速シークエンサの配列解析の依頼が持ち込まれる。ただし高速 DNA シークエンサの情報解析技術を持つ人数は限られている。この為、我々は 2009 年から国立遺伝学研究所のスーパーコンピュータを使う高速シークエンサ由来配列の解析システム DDBJ Read Annotation Pipeline (http://p.ddbj.nig.ac.jp/)を提供している[Kaminuma 10]。2013 年 4 月時点で、外国人 35 名を含む 136 名が登録しており、2012 年 1 年間では約 1,000 ジョブの利用があった。高速シークエンサ配列の情報解析は、塩基配列に遺伝子領域等の注釈情報を付与する為に自動アノテーションと呼ばれる。

自動アノテーション・ツールの処理後は、手作業による修正が必要な場合が多く、大規模なデータ処理ではボトルネックとなり対処策構築が喫緊の課題になっている。人手による作業はキュレーション、作業者はキュレータと呼ばれている。我々は学術文献か注釈情報をキュレーションする TogoAnnotation と呼ぶ作業支援ツールを開発しており、これを DDBJ Read Annotation Pipeline と繋げて、塩基配列自動注釈後のキュレーションツールとして利用したい。しかし、TogoAnnotation は小規模な専門家コミュニティのキュレーション・プロジェクトを想定しており、高速シークエンサ出力データの様に、大規模な注釈情報をキュレーションする機能はない。TogoAnnotation を大規模データに対応させるには、まずキュレータの人数をこれまでの数十名の規模から大人数へと拡張する必要がある。このとき不足するキュレータ数の増強に、集団(Crowd)の力を借りてコスト低減を図るクラウドソーシング(Crowdsourcing)を想定している。これまでは、高度な専門知識を持つキュレータのみを採用して作業を割当てていたので、キュレータの専門性が問題になる。

本研究では、既存のアノテーションシステム TogoAnnotation にクラウドソーシングの機能を拡張する為に、オントロジを利用したデータ・キュレーションの精度評価法を提案する。特に本稿では、キュレータの評価手法に焦点をあてて、評価手法のパイロット実験を行う。これらキュレーションの評価を検討しながら、将来的に大規模データ・キュレーション向けの機能実装を目指す。

連絡先:神沼英里, 国立遺伝学研究所 生命情報センター 大量遺伝情報研究室, 〒411-8540 三島市谷田 1111, 055-981-6859, ekaminum@nig.ac.jp

## 2. ゲノム研究におけるキュレーション

### 2.1 ゲノム研究におけるキュレーション

ゲノム研究におけるキュレーションでは、「Jamboree」と「Community Annotation」と呼ばれる運用体制が従来とられてきた。Jamboree は、世界中から専門家を一か所に集合させて研究ワークショップを開き、顔を突き合わせて注釈付作業をする形態を指す。Jamboree は世界中の研究者を一か所に集めるコストが問題になる為に、インターネット上でアノテーションの作業ツールにログインして各研究者が作業をする「Community Annotation」という形態が発展した。我々は、Community Annotation ツールとして、複数の研究者が一定期間、非公開環境でゲノムアノテーションを実現する為のプラットフォーム、TogoAnnotation (http://togo.annotation.jp/) を構築し、インターネット上に公開している。

### 2.2 TogoAnnotation: コミュニティ・アノテーションツール

TogoAnnotation は、ゲノムデータベースを特定研究コミュニティ／グループに限定したゲノム情報の提示と、アノテーション編集機能を有しており、モデル微生物を中心にコミュニティアノテーション環境を提供している。Bradyrhizobium sp. S23321 ゲノム解析においては、根粒菌研究コミュニティーによる 31 名のキュレータが 2 泊 3 日の Jamboree 形式でゲノムアノテーションを実施した [Okubo 12]。データ構築では、年平均キュレータ数 8 名の体制で、平成 19 年度から継続してモデル微生物の文献

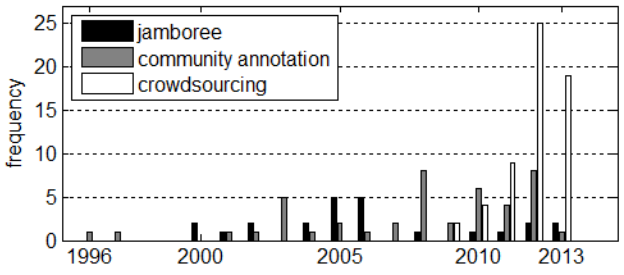
キュレーションによるゲノムアノテーションリファレンスを集積している。2013 年 3 月末時点の総計は、24 生物種、30,698 遺伝子, 文献と遺伝子の関連付けは 221,670 件であった。

また現在、他の遺伝子注釈情報との統合を目的として、セマンティックウェブ技術を導入して RDF 形式のデータを整備中であり、微生物系統間の差異を示す一塩基多型のゲノム注釈データを http://semantic.annotation.jp/sparql/から公開している。

表 1: キュレーション・プロジェクトの運用形態

| Project Type | Curation Type | Curation Style | Total hit# |
|---|---|---|---|
| Annotation Jamboree | Experts | Offline | 24 |
| Community Annotation | Experts | Online Remote | 44 |
| Crowdsourcing | Non-experts | Online Remote | 59 |

図1:プロジェクト形態別の研究報告数の変化(横軸:年)




## 3. ゲノム研究における集合知の利用動向

### 3.1 ライフサイエンス研究における集合知利用

専門家が研究等の作業推進の為に、注釈付(アノテーション/キュレーション)を行うのではなく、非専門家を含む集団(Crowd)の知力を借りる方策は、Crowdsourcing と呼ばれている。注釈付の作業形態は、上述の Jamboree, Community Annotation が作業者を専門家に限定しているのに対し、Crowdsourcing は作業資格を非専門家に拡張する。表 1 に、ワーカのタイプ(専門家 or 非専門家)と作業環境(オフライン or オンライン, リモート)により、Jamboree と Community Annotation と Crowdsourcing の特徴をまとめた。

また、ライフサイエンス分野において Crowdsourcing の研究報告が近年増えている。米国 National Center for Biotechnology Information(NCBI)が構築している、2200 万件収録のライフサイエンス系論文誌等の文献データベース PubMed (http://www.ncbi.nlm.nih.gov/pubmed)を使って、crowdsourcing 等のキーワード検索を行った。図 1 は、3 つのプロジェクト形態別の、検索ヒット件数のヒストグラムである(2013 年分は 4 月時点までの件数)。トータルのヒット数は表1に掲載した。2009 年が初出のキーワード Crowdsourcing は、毎年件数が増えている様子が見られる(2013 年分は 4 月までの件数)。

Crowdsourcing 検索結果のタスク対象は、多岐に亘っている。画像上で白血球等と区別するマラリア タグ付けゲーム[Luengo-Oroz 12]、タンパク質シグナル伝達ネットワークの推定[Prill 11]、自己申告による乾癬リウマチの症例データ収集[Armstrong 13]。などがある。また、Amazon Mechanical Turk の利用評価[Crump 13]の論文や、ゲームの効果についての報告[Khatib 11]などがある。発表に合わせて、国立遺伝学研究所大量遺伝情報研究室のウェブサイト http://charles.genes.nig.ac.jp/curation_eval/に、LINK 等の参考文献のまとめ情報を公開するので参照して頂きたい。

### 3.2 ゲノム分野における集合知利用の先行研究

Crowdsourcing を使ったゲノムアノテーションの先行研究では、wikipedia 上の遺伝子キュレーションについて調査した The Gene Wiki[Huss 08]がある。欧州の代表的な機関 European Bioinformatics Institute(EBI)/Sanger 研究所では長年、塩基配列に Gene Ontology (GO) [GO Consortium 00]のタグ付けをすることで遺伝子の機能注釈を行っている。ゲノム分野での主たるキュレーション作業の一つに、GO のような専門用語の割当作業がある。Huss らは、自動処理で遺伝子ページを生成し、その後のキュレータの編集率を検討している。また、Semantic Wiki Link を導入した検索用アプリケーション GeneWiki+なども発表されている。

### 3.3 TogoAnnotation 機能拡張課題:タスク設計

我々は TogoAnnotation ツールを拡張して、Crowdsourcing に対応するため環境を整備したい。Gene Wiki のように、wikipedia をプラットフォームとして使い、完全に不特定多数のユーザに開放する課題設定もある。しかし、現在の Wikipedia を利用する研究の多くは、ヒトデータが対象である。またキュレーション作業終了までの期限がない。一方 TogoAnnotation では、微生物や植物など多様な生物種を対象としている上に、高速シークエンサによる大量注釈情報を、出来るだけ短時間かつ高精度にキュレーションするプロジェクトを実施したい。

このため、まず機能拡張としてタスクの最適分配方法を考える。専門性により能力差のあるキュレータに、最適なタスクを振り分ける機能を追加する事で、作業時間の短縮を図る。次項では、パイロット実験として、オントロジを利用したキュレータの専門レベルの定量化実験を紹介する。

## 4. オントロジを利用するタスク設計法の提案

### 4.1 提案タスク設計法:キュレータの専門性定量化

ここでは、キュレーションのタスク設計の第一歩として、バイオ分野で利用されているオントロジを利用したキュレータの専門性定量化を提案する。タスクやキュレータの専門レベルを推定する為に、まず専門レベルを定量的に定義する必要がある。本研究では、専門レベルを、(1)その専門分野の用語の熟知度、(2)プロトコルが類似した作業自体の経験数、の 2 つの軸を持つ変数と仮定する。更に、簡便の為プロトコル経験数は考慮せずに、専門用語の熟知度のみを評価尺度として採用する。本研究では、オントロジを構成する統制語(Controlled vocabulary)を対象分野の専門用語とし、各オントロジは専門分野を示すと考える。これよりオントロジを構成する統制語の利用頻度で、定量的な検討が可能になる。

提案法では、キュレータの専門レベルは、キュレータが発表している過去の発表文献におけるオントロジ専門用語の出現率と定義する。またキュレーションデータについては、キュレーション対象の注釈文中での、オントロジ専門用語の出現率で専門レベルを評価する。最終的に、専門性が合致すると判断されたキュレーションタスクのみ、キュレータに割り振るものとする。

### 4.2 実験材料

NCBI PubMed を使い、15 名の研究者の論文アブストラクトデータを収集した。ライフサイエンスのオントロジーは The Open

Biological and Biomedical Ontology Foundry(OBO)[Smith 07]のダウンロードサイト(http://www.obofoundry.org/)から 13 件の OBO オントロジファイルを取得し、”name”項を専門用語として抽出した。

### 4.3 実験手順

実験作業は、専門用語データセットの構築と、キュレータの専門度定量化、に分けられる。専門用語データセットの構築は、13 件の OBO オントロジのファイルから専門用語を、単語単位で抽出した。句読点等は除去するクレンジングを行った。また各オントロジ間で重複する単語は、専門性を反映しないと仮定して除去した。

キュレータの専門度定量化は、ランダムに選択した専門分野の異なる研究者(キュレータ)の発表論文を NCBI PubMed で検索して、そのアブストラクトを収集した。アブストラクト中のユニークな単語のセットを作成し、オントロジ別に専門用語セットにヒットする数を計数した。

## 5. キュレータの専門度定量化実験

13 件の OBO オントロジから、ユニークな単語として 19,368 件の専門用語を抽出した。更に、オントロジ間で重複して出現する単語を削除した後で、83%に相当する 16,017 件の専門用語からデータセットを構築した。

また 15 名のライフサイエンス分野の研究者の論文情報を PubMed から収集したアブストラクトに含まれている単語を、キュレータの既知専門用語セットと定義した。15 名の平均と標準偏差は、1,728±993 単語であった。

各 OBO オントロジへの専門用語の平均ヒット単語数について、15 名のキュレータの平均値と標準偏差を図 2 に示す。縦軸は各オントロジの略名を表す。オントロジが異なると、用語ヒット数にばらつきがある(オントロジを構成する用語数の偏りは、計算に考慮していない)。次に、最大ヒット数により規格化して出現率と定義して、全オントロジの単語分布頻度を視覚化した(図 3)。縦軸はキュレータののべ人数であり、横軸はオントロジ専門用語の出現率を表す。規格化に使ったヒット数最大値のキュレータを外すと、専門用語の出現率が高くなるにつれ、キュレータの対象数が低くなる傾向があるかもしれない。現在は 15 名分のキュレータ数のデータしかない。分布傾向を統計的に検討する為に、今後、キュレータ数を増やして分析する必要がある。

図 2:オントロジ別の専門用語ヒット数平均値と標準偏差

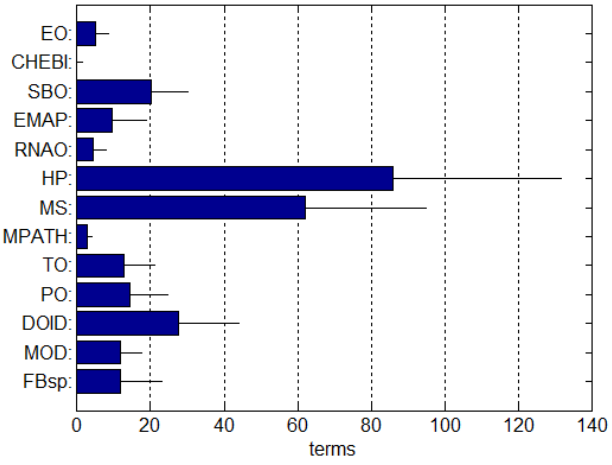
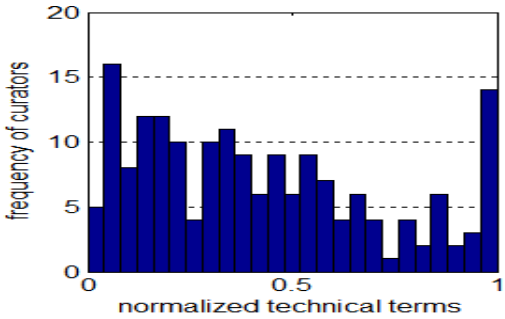



図 3:オントロジ全体の専門用語出現率とのべ人数頻度




## 6. 課題と展望

専門知識に基づいた注釈情報を手作業で付与する、データキュレーションの高コスト問題に対して、オントロジを利用したタスク設計について提案した。タスク設計提案の具体例として、キュレータの専門度の定量化を行った。今後、統計的に検討を行い、キュレータの専門性の判断材料として、閾値設定に利用していく。

課題として、TogoAnnotation に本提案の定量化ツールを実装していきたい。また技術的課題として、Evidence Code の利用拡張がある。Evidence Code は、実験・計算方法・マニュアル文献キュレーションなどの科学的証拠についての統制語彙である。ゲノムアノテーションにおいて、遺伝子注釈 GO の精度評価として利用されている。Inferred by Curator (IC), IEA(Inferred from Electronic Annotation)などのコードがあるが、IC を拡張して定量評価値と共にデータベースに追加していく。

このキュレーション評価ツールの開発に着手した理由は、既発表のデータベース論文の投稿時に、査読者から高度な専門情報について注釈間違いを指摘された事が発端である。このデータベースは全生物種を対象にしたもので、キュレーションタスクとしては困難な部類になる。生物種が網羅される程、キュレーションタスクは困難になり、低精度データが含まれる事になる。しかし当然ながら、最初から低精度の注釈作業をタスクから外せた方が良い。本提案を発展させることで、注釈作業の無駄を省き、注釈精度を向上させていきたい。またライフサイエンスのデータベースを構築するにあたり、全生物種を対象にするケースは、今後も増えていくと考えられる。キュレーション評価法を発展させていく事で、多様な生物種をデータ・キュレーションで取扱うための指針が構築できると考えている。

## 謝辞

データ生成や方法論についてディスカッションを行った国立遺伝学研究所 大量遺伝情報研究室の皆様に感謝致します。

## 参考文献

[Kaminuma 10] Kaminuma EK, Mashima J, et al. : DDBJ launches a new archive database with analytical tools for next-generation sequencing data , Nucleic Acids Res, 38, pp.D33-38 (2010).

[Okubo 12] Okubo T, Tsukui T, et al., Complete genome sequence of Bradyrhizobium sp. S23321: insights into symbiosis evolution in soil oligotrophs., Microbes Environ, 27, pp.306-15 (2012)

[Luengo-Oroz 12] Luengo-Oroz MA, Arranz A, Frean J, Crowdsourcing malaria parasite quantification: an online game for analyzing images of infected thick blood smears, J Med Internet Res.14, pp. e167 (2012)

[Prill 11] Prill RJ, Saez-Rodriguez J, et al, Crowdsourcing network inference : the DRAM predictive signaling network challenge, Sci Signal, 4, pp. mr7 (2011)

[Armstrong 13] Armstrong AW, Wu J, et al, Crowdsourcing for data collection: a pilot study comparing patient-reported experiences and clinical trial data for the treatment of seborrheic dermatitis, Skin Res Technol, pp. 55-57 (2013)

[Crump 13] Crump MJ, McDonnell JV, Gureckis TM, Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research, PLoS One, 8 , pp.e57410 (2013)

[Khatib 11] Khatib F, DiMaio F, et al., Crystal structure of a monometic retroviral protease solved by protein folding game player, Nature Struc & Mol Biol, 18, pp.1175-1177 (2011)

[Huss 08] Huss JW, Orozco C, et al. A Gene Wiki for Community annotation of gene function, PloS Biol, 6, pp.e175 (2008)

[GO Consortium 00] The Gene Ontology Consortium, Gene ontology: tool for the unification of biology, Nat Genet, 25, pp.25-29 (2000)

[Smith 07] Smith B, Ashburner M, et al., The OBO Fundry: coordinated evolution of ontologies to support biomedical data integration, Nat Biotech, 25, pp.1251-5 (2007)