1L4-OS-09a-6

CrowdR&D: クラウド協働評価のための参加型R&Dプロジェクト情報統合基盤

CrowdR&D: An Integrated Information Platform of Participatory R&D Projects for Crowd Collaboration and Performance Evaluation

神沼英里 *1 Eli Kaminuma　望月芳樹 *2 Yoshiki Mochizuki　藤澤貴智 *1 Takatomo Fujisawa　馬場雪乃 *4 Yukino Baba　藤山秋佐夫 *1*3 Asao Fujiyama　鹿島久嗣 *4 Hisashi Kashima

中村保一 *1 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

*2理化学研究所 統合生命医科学研究センター
RIKEN Center for Integrative Medical Sciences

*3国立情報学研究所 情報学プリンシパル研究系
Principles of Informatics Res Div, National Institute of Informatics

*4京都大学 情報学研究科
Grad School of Informatics, Kyoto University

High-performance next-generation sequencing (NGS) technologies are advancing genomics and molecular biological research. However, massive amounts of NGS sequence data have created a bottleneck at human curation with manual tasks. To resolve the problem, we investigated crowdsourcing approach to accomplish curation tasks. In this report, we propose an integrated information platform of participatory R&D projects, named by CrowdR&D, for the purpose of crowd collaboration and performance evaluation. Researchers can use the CrowdR&D website as a portal to access individual crowdsourcing websites. It provides quantitative evaluation of individual tasks and crowd performance. Finally, we describe the information of ethical review process for protecting crowds.

1. はじめに

近年、ライフサイエンス分野ではビッグデータが生成される様になり、研究工程でクラウド (群衆) の活用が始まっている。高速 DNA 解読装置の普及と共に、大規模な DNA 配列データが生成される様になり、我々は 2009 年から DNA 配列自動注釈システム DDBJ Read Annotation Pipeline[Kaminuma 10] を提供している。DDBJ Pipeline は国立遺伝学研究所スーパーコンピュータ [Ogasawara 13] 上に実装されており、2015 年 1 年間で 253 名の新規ユーザ登録があり約 9,000 ジョブが実行された。

自動注釈処理後は注釈エラーを修正する為に、手作業が必要になる。我々はオンラインでの手作業注釈支援ツール TogoAnnotation を開発している [Fujisawa 14]。TogoAnnotation はこれまで専門家作業者を対象に運営してきた。しかし専門家は人数が少なく大規模データを扱う場合には、全体工程中で手作業部がボトルネックになっている。そこで 2014 年に Pilot 研究として非専門家に専門作業を実施してもらい、専門家との性能を比較した。結果として、非専門家の平均精度は専門家より低いが分散が大きく、非専門家の中には専門家より高いパフォーマンスを示すケースがある事が判明した [Kaminuma 14]。専門タスクへの非専門家の貢献可能性が示唆された為、非専門家と専門家の協働参加体制を構築できれば、研究開発推進に繋がる可能性がある。

また手作業データを蓄積すれば、機械学習モデルの精度向上に繋がる [Von Ahn 06]。蓄積した手作業データを機械学習モデルの訓練に用いる工程が確立できれば、注釈間違いデータの削減も期待できる。この作業データ蓄積から機械学習モデル性能向上までのワークフロー化を目指したい。Data Tamr[Palmer 13] は手作業から機械処理化までの枠組みを利用者に提供している。しかし個々タスクに対応した分析評価法がワークフローの構成要素として提供されていない。オープンなデータ評価分析情報をワークフローと共に提供できれば、クラウドの参加判断材料に使えるだろう。

そこで本研究では、参加型研究促進の為のクラウド協働体制と評価分析機能付きワークフローの構築を目指して、参加型 R&D プロジェクトの情報統合基盤「CrowdR&D」を提案する。CrowdR&D 情報統合基盤は、タスクのカタログ・ポータル機能、参加投稿データの評価分析機能、クラウド参加実績評価機能を持つ。参加型プロジェクトの情報を集めたポータルサイトとして、クラウドに R&D プロジェクトへアクセスする為の支援情報を提供する。

CrowdR&D の対象タスクは、クラウドの参加促進の為に R&D に関係する内容であれば、専門分野は問わない。専門分野を限定しないので、幅広いバックグラウンドのクラウドが参加対象になる。ある分野の専門家クラウドは、別分野では非専門家クラウドとして参加貢献が可能になる。また分野横断的に研究開発タスクを俯瞰できるので、タスク発行者側には異分野参入情報も期待できる。

また CrowdR&D 情報基盤はクラウド（＝参加者) の参加実績を評価する為に、被験者保護の観点から研究倫理審査の対象となる。本稿の最後で、研究倫理審査の情報を紹介する。

2. 情報統合基盤 CrowdR&D の機能

CrowdR&D サイトは参加型 R&D プロジェクトの情報統合基盤であり、クラウド協働の仕組化と参加評価分析の情報提供を目標としている。CrowdR&D の主機能は次 3 項目になる。

[1] 参加型タスクのカタログ機能

[2] 投稿データの評価分析機能

[3] クラウド参加実績評価機能

1 番目のカタログ機能は、個別のウェブサイトで提供されている参加型研究開発タスクの情報を CrowdR&D サイトに集約

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

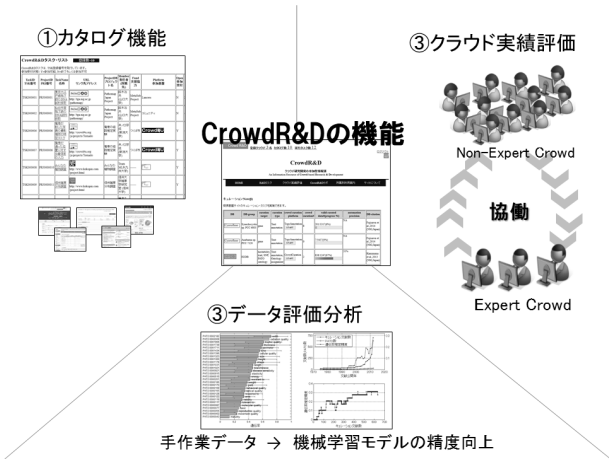

図 1: 参加型 R&D 情報統合基盤 CrowdR&D

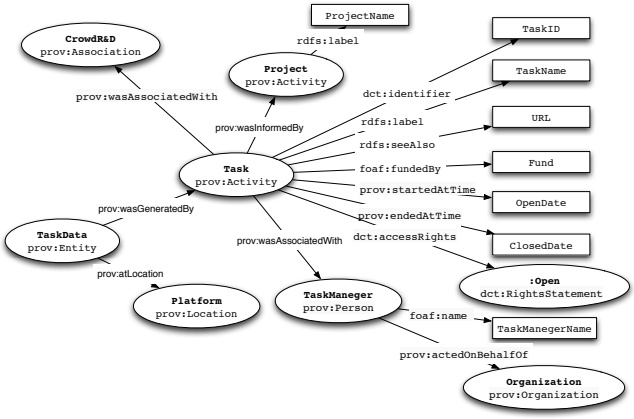

図 2: CrowdR&D カタログ機能:タスクのデータモデル

して統合データベースとして提供する。2 番と 3 番の機能は、参加協働の評価機能である。参加協働評価は、参加者 (クラウド) 対象と作業データ対象の 2 種類が存在する。クラウド対象は参加実績評価機能で、作業データ対象は投稿データ評価分析機能を示す。3 機能を以下順番に紹介する。

2.1 カタログ機能：参加型 R&D タスク情報の統合

1 番目のカタログ機能は、各ウェブサイトで公開されている参加型タスクの情報を集約して参加型タスクの統合カタログとしてユーザに情報を提供する。カタログ機能の目的は、各研究者が提供する (CrowdR&D 外部の) タスク実行サイトへのユーザの誘導である。

カタログの掲載タスクは、参加型と研究開発の特徴で選別しており、承認制で掲載内容の質を担保している。カタログは、クラウド参加型のタスク中で研究開発 (R&D) に特化したデータを対象とする。幅広い研究開発タスクをサポートする為に、専門分野を限定しない。また登録時に CrowdR&D のボードメンバで内容を審査する事で、タスクの質を担保する。基本的に研究開発に携わる者がサポートするタスクを、タスク内容と研究倫理の観点から審査している。参加型研究は新分野で従来助成枠に該当しにくい為に、研究者・開発技術者らがプライベートで運営しているケースも多い。この理由から仕事・プライベートの区別なく審査を行っている。

カタログのメタデータは、「TASK」「PROJECT」「SUPPORT PLATFORM」の 3 項目を中心に設計されている。TASK は作業タスク自体に割当てる。PROJECT は、各タスクを発行する R&D プロジェクトを表し、1 つの PROJECT に複数の TASK を割り当て可能とする。一人のタスク管理者が、複数の PROJECT を保有して、各 PROJECT は複数 TASK を構成する。また SUPPORT PLATFORM は、タスクを作業する実行基盤サイトの情報を記載する。品質管理の為に、表 1 の様に、接頭部に 6 桁数字を含む識別番号を項目毎に割当てる。

表 1: CrowdR&D カタログ機能の識別番号

Catalog Data Type	Identifier	Description
TASK	CRT000001～	作業タスク
PROJECT	CRP000001～	プロジェクト
SUPPORT PLATFORM	CRS000001～	作業実行基盤

メタデータは、RDF 表現のデータモデルを定義した。図 2 に TASK データモデル (Provisional 版) を示す。属性には、参加の判断材料となり得る情報を採用した。2016 年 3 月時点での TASK データは、Crowd4U[Morishima 12]、ビッグデータ大学 [Kashima 14]、ここピン [Matsuda 12] 等を掲載している。本カタログ機能により、各ウェブサイトへ訪問する前に、クラウドが R&D 協働募集状況を確認する事が可能になる。

CrowdR&D での参加者（クラウド）は、非専門家と専門家の両方を対象とする。専門的な参加型タスクの場合は非専門家と専門家の両方を対象にデータを収取し、非専門家データの採用精度の閾値を設定する場合が多い。この為に、特定の専門家が所属するコミュニティにのみに開かれた研究開発タスク (＝専門家のみ募集の限定参加型タスク) の場合もカタログに掲載される。

2.2 投稿データの評価分析機能

2 番目の投稿データ評価分析機能は、参加実行結果で生成される「データを対象」に評価分析を行う機能である。この機能は、開発目標の評価分析機能付きワークフローの構成要素との位置付けである。ワークフロー開発には、手作業データ収集→機械学習用データセット構造化→モデル精度向上 (機械学習モデリング) →人手処理と機械処理の役割分担最適化、と段階的な発展を想定している。この為に、本評価分析機能は構成要素ツールを個々に検討していく。

今回は第一段階の手作業データ収集に着目して、タスク投稿数・タスク目標進捗率・タスク投稿数予測・タスク貢献格差分析の検討を行った。評価分析用データには、提携サイトから次の 5 タスクを用いた。データ注釈タスクの光合成細菌 Synechocystis sp. PCC 6803 系統の遺伝子機能キュレーション (PCC6803 注釈 [Fujisawa 14] タスク=A1), 遺伝率キュレーション (H2DB[Kaminuma 13] タスク=A2), データモデリング・コンペティション基盤ビッグデータ大学での公開 3 タスク (テキスト分類問題その 2=M1, オンラインマーケットでの購買予測=M2, 周辺地点の気象情報からの気温予測=M3) の投稿データについて評価分析を行った。

(1) タスク投稿数とタスク目標進捗率

提携サイトのタスク投稿数とタスク進捗率を、CrowdR&D サイトの評価分析ページに表示する。タスク投稿数は、投稿日を横軸に、のべ投稿総数を縦軸としたグラフで表す。タスク

目標進捗率は、タスク毎に設定されるタスク目標値に対する現時点での投稿数で定義する。例えば PCC6803 注釈タスクの場合は、タスク目標値は PCC6803 系統の遺伝子総数 3,317 である。現時点の投稿済遺伝子数 1,086 で、タスク進捗率は 0.33 となる。1 遺伝子に対して複数の参加者が注釈を投稿するので、タスク投稿数と投稿済遺伝子数が異なる点に注意が必要である。

(2) タスク投稿数の予測

タスク投稿数は、最終投稿締切日での予測値があれば、タスク発行者にとって有用な情報となる。参加型タスクの募集形式では、締切日を設定する閉鎖型 (Closed-ended) と、締切日を設定しない開放型 (Open-ended) がある。閉鎖型の場合は締切日で、開放型の場合はマイルストーンで、タスク投稿数を予測する。

今回は試験的に、投稿数予測とタスク日数 (Duration) の関係を調査した。締切日までの日数はタスク毎に異なるので、正規化時系列に変換する。教師有学習アルゴリズムにはサポートベクトル回帰モデル (SVR) を用いた。 Duration を 10%から

表 2: CrowdR&D 評価分析機能: 投稿総数予測

タスク名	ID *1	投稿総数	日数	MAPE
PCC6803	A1	1149	58	0.056
H2DB	A2	913	120	0.061
テキスト分類	M1	191	38	0.310
購買予測	M2	187	33	0.065
気温予測	M3	481	43	0.120

*1 A: ANNOTATION, M: MODELING

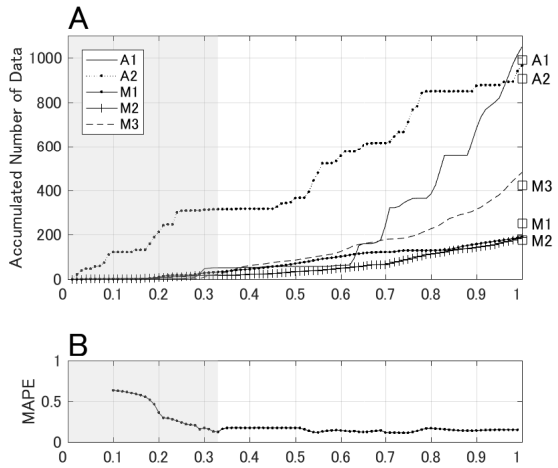

図 3: A. 締切日の投稿数予測とタスク投稿数, B. 機械学習による予測誤差, 共に横軸は正規化タスク日数

99%まで変化させて投稿数と日数をモデルに学習して、締切日投稿数の予測誤差を Mean Absolute Percent Error(MAPE) で計算した (表 2)。図 3A 横軸にタスク日数、縦軸にタスク別の投稿数をグラフ化した。Duration 33%を訓練に用いた時の各タスク予測投稿数を、図 3A の右に□マークで示した。図 3B 縦軸は MAPE である。表 2 に Duration 33%でのタスク別 MAPE 値を示した。

参加型データ分析の過去研究では、米国大手クラウドファンディングサイト Kickstarter[Kickstarter 09] のデータを用いたプロジェクト成功予測、最終資金額予測の報告がある ([Greenberg 13], [Chung 15]). Chung らの報告ではプロジェクトの属性情報、SNS 情報、資金の時系列初期情報を入力として、AdaBoost モデルでプロジェクト成否を予測している。Kickstarter 分析では 1 万プロジェクトを超えるデータセットが訓練に使われており、CrowdR&D でも予測精度向上に繋がるタスクデータの蓄積が今後の課題である。

(3) 参加貢献格差の定量分析

提携サイトのデータには投稿日時情報と共にクラウド ID が割り当てられており、参加者単位で投稿数を計算出来る。参加型プロジェクトの貢献量には、クラウド間で偏りが生じる事が知られている [Ortega 08][Koch 11]。

参加貢献の不均衡定量分析に、Lorenz 曲線と Gini 係数が用いられる。Gini 係数は、参加者間の貢献格差が大きい程 1 に近付き (maximum inequality)、格差が無いと 0 になる (perfect equality)。OSS(Open-Source Software) 開発で調査された Gini 係数は 0.75 である [Singh 07]。オンライン注釈の Gini 係数は、市民科学サイト Zooniverse の 7 プロジェクトで 0.77〜0.91[Sauermanna 15]、参加型百科事典構築の Wikipedia は 0.92 以上 [Ortega 08]、参加型地図作成の OpenStreetMap は 0.95 以上 [Yang 15] である。

タスク別の Lorenz 曲線を図 4 に、Gini 係数を表 3 に示す (H2DB タスクはキュレータ数 2 名と少ない為に省いた)。 平

表 3: CrowdR&D 評価分析機能: 貢献格差分析

ID	参加者数	平均投稿数	最大貢献率	Gini 係数
A1	13	89	0.27	0.58
M1	16	12	0.20	0.43
M2	22	8.5	0.16	0.54
M3	45	11	0.07	0.50

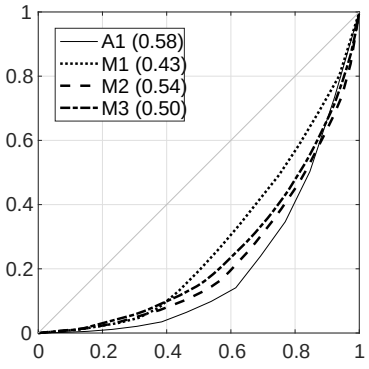

図 4: タスク単位参加貢献分析: Lorenz 曲線と Gini 係数

均投稿数は投稿総数/参加者で、最大貢献率は一番貢献したクラウドの投稿数/投稿総数で計算した。ゲノム研究分野で、専門家のみを集めた遺伝子注釈作業はコミュニティアノテーションと呼ぶ。注釈タスクで、参加者に等しい作業量を割当てる場合は Gini 係数は 0 に近づく。一方 PCC6803 注釈タスクでは、13 人の参加者に均等割当をせず自由投稿形式を採用している。結果の Gini 係数は 0.58 で偏りがあり、一番投稿数の多いクラウドは、全体の 27%の貢献をしていた。ビッグデータ大学の 3 タスクの平均 Gini 係数は 0.49 であり、PCC6803 注釈タスクより貢献格差は少ない。今後、分析対象タスク数を増やして、参加貢献格差への影響因子を調査していく。

2.3 クラウド参加実績評価機能

3 番目のクラウド参加実績評価機能は、「参加者 (クラウド) を対象」として参加の評価分析を行う機能である。参加実績評価の指標として「参加回数」と「タスク実行の成績・性能」が考えられる。まずは、実績評価項目としてクラウドの参加回数に着目する。現行版では、CrowdR&D と提携するクラウドソーシング・サイトが、クラウドの参加回数を公開している場合に限り CrowdR&D で実績評価を行う。

提携するクラウドソーシング・サイトとは SNS アカウントを介して、実績参加回数を管理している。現状では、複数の SNS アカウントを利用している場合は、アカウント別の実績管理となる。

複数のクラウドソーシング・サイトでの参加実績は、ユーザのポートフォリオとして CrowdR&D 上で管理される。現版では、ユーザの参加回数は非公開である。今後、ユーザから要望があれば、CrowdR&D 上での参加実績情報の公開を検討していく。

3. クラウド保護と研究倫理審査

クラウドソーシング研究は、参加者に害を及ぼす可能性がある場合は、研究倫理審査 (Institutional Review Board:IRB) の承認が必要である [Graber 13]。CrowdR&D は、参加者であるクラウドの実績評価機能を持つので、クラウドは被験者扱いとなる。この為、被験者保護の観点から、研究倫理審査での研究内容承認が必要となる。CrowdR&D のサイト公開にあたり、国立遺伝学研究所の IRB に審査を申請した。

IRB 要件に、クラウドの研究同意書 (Informed Consent:IC) 取得がある。IC の項目は下記である。

1) 研究目的・ 協力方法・ 実施体制・ 研究期間について
2) 本研究が国立遺伝学所の倫理審査委員会で承認された上、開始されること
3) 本研究成果の公表について
4) 利益・不利益について
5) 本研究のデータの個人情報保護および 匿名化について
6) 本研究のデータの保管と廃棄について

CrowdR&D サイトへの最初のアカウント認証時にユーザに IC を提示する。参加希望ユーザで IC に同意しなかった場合でも、実績評価機能以外の CrowdR&D 機能を利用出来る。IRB から承認が得られれば、http://crowdrnd.jp/から CrowdR&D のウェブサービスを公開する予定である。

4. まとめと今後の展望

本稿では参加型 R&D 推進のための情報統合基盤システム CrowdR&D を紹介した。特に CrowdR&D の主要 3 機能として、カタログ機能、データ評価分析機能、クラウド参加実績評価機能の説明を行った。

今後の展望として、投稿データ精度評価とクラウド性能評価に焦点をあてて CrowdR&D の機能を追加していく。特に専門家クラウドと非専門家クラウドでの性能比較、蓄積データからの機械学習モデリングの為の、機能ツールの実装を進めていく。将来的には、一連の構成要素機能をまとめてワークフロー化を行い、作業工程期間の短縮を図る。クラウドの実績評価については、登録ユーザアカウントの名寄せ機能を作成していく。現状では 1 名が複数のユーザアカウントを保持出来る。名寄せ機能があれば、複数のアカウントの参加実績の一元管理が可能になる。

参加型 R&D 情報統合基盤として CrowdR&D を発展させる為に、皆様からコメントを頂ければ幸いである。

Acknowledgements

We are thankful to Atsuyuki Morishima (Tsukuba University), Nori Kurata (National Insitute of Genetics), for advices to the proposed platform. We are also grateful to Kousaku Okubo, Isao Katsura, Naruya Saito (National Institute of Genetics) for Ethical Review Procedures.

参考文献

[Ogasawara 13] Ogasawara, O. et al.: DDBJ new system and service refactoring, NAR, 41, pp.D25-29 (2013)

[Kaminuma 10] Kaminuma, E. et al.: DDBJ launches a new archive database with analytical tools for next-generation sequencing data, NAR, 38, pp.D33-38 (2010). http://p.ddbj.nig.ac.jp/

[Fujisawa 14] Fujisawa, T. et al.: CyanoBase and RhizoBase: databases of manually curated annotations for cyanobacterial and rhizobial genomes, Nucleic Acids Res, 42, pp.D666-70 (2014). http://annotation.jp/

[Kaminuma 14] 人工知能学会全国大会 1J5-OS-18b-3 (2014)

[Von Ahn 06] Von Ahn, L., Games with a Purpose, Computer, 39, pp.92-94 (2006)

[Palmer 13] http://www.tamr.com/ (2013)

[Graber 13] Graber, MA1 et al. : Internet-based crowdsourcing and research ethics, J. of Med Ethics, 39, pp.115-8 (2013)

[Kashima 14] http://universityofbigdata.net/

[Morishima 12] CyLog/Crowd4U: a declarative platform for complex data-centric crowdsourcing, The Proceedings of the VLDB Endowment, 5, pp.1918-1921 (2012)

[Matsuda 12] http://www.kokopin.com/

[Kaminuma 13] http://tga.nig.ac.jp/h2db/.

[Kickstarter 09] https://www.kickstarter.com/

[Greenberg 13] Greenberg, MD. et al. : Crowdfunding Support Tools, CHI EA '13, pp.1815-1820 (2013)

[Chung 15] Chung, J. et al.: A Long-Term Study of a Crowdfunding Platform, ACM HT'15, pp.211 (2015)

[Singh 07] http://citeseerx.ist.psu.edu/viewdoc/summary?doi= 10.1.1.145.4204 (2007)

[Sauermanna 15] Sauermanna, H. and Franzonib, C., Crowd science user contribution patterns and their implications, PNAS, 112 pp.679-684 (2015)

[Ortega 08] Ortega F, et al., On the inequality of contributions to Wikipedia. Hawaii Int. Conf. on System Sciences, pp.304 (2008)

[Koch 11] Koch,S., Organisation of work in Open Source Projects: expended effort and efficiency, Revue d'economie industrielle, pp.17-38, (2011)

[Yang 15] Yang, A., Fan, H. et al., Temporal Analysis on Contribution Inequality in OpenStreetMap ISPRS Int. J. Geo-Inf., 5,p.5 (2016)