1J5-OS-18b-3

遺伝子構造キュレーションのクラウドソーシング・タスク設計

Task Designs for Crowdsourced Data Curation in Genomic Studies

神沼英里 *1 Eli Kaminuma　馬場雪乃 *2*3 Yukino Baba　藤澤貴智 *1 Takatomo Fujisawa　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国立情報学研究所 ビッグデータ数理国際研究センター
Global Research Center for Big Data Mathematics, National Institute of Informatics

*3JST, ERATO, 河原林巨大グラフプロジェクト
JST, ERATO, Kawarabayashi Large Graph Project

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

In lifescience studies, data scale of generated genomic sequences has grown rapidly by technological innovation. 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 introduce several crowdsourced tasks such as image-based annotation of gene structure, discrimination of professional words to investigate difference of precision between expert curators and crowd workers. Further, towards curation task design, we propose a quantitative measure of professional/non-professional tasks as the degree of difficulty annotation task.

1. はじめに

2005 年頃から登場した高速 DNA シークエンサの出現により、ゲノム配列の解読コストが急激に下がった。現在では、ゲノム配列が大量に生成されるようになり、配列に遺伝子領域等の注釈情報を付与するアノテーションのニーズが増大している。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って高速 DNA シークエンサ由来の自動配列注釈システム DDBJ Read Annotation Pipeline (http://p.ddbj.nig.ac.jp/) を提供している [Kaminuma 10]。2014 年 3 月時点で、外国人を含む 404 名が登録しており、2013 年 1 年間では約 4,000 ジョブの利用があった。

機械による自動注釈処理の後は、手作業により配列の注釈情報を修正する（キュレーション）が、大規模データの場合はコストが高いために、最近はキュレーションをせずに低精度のまま論文発表するケースが増えている。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 [Fujisawa 14]、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。TogoAnnotation は、これまで小規模な専門家コミュニティのキュレーション・プロジェクトや特定のモデル微生物の文献キュレーションを想定しており、高速シークエンサ出力データに対して、大規模な注釈情報をキュレーションするには適さない。TogoAnnotation を活用し、データの規模拡大を行なう上での解決策の 1 つとしては、キュレータを少人数体制から大人数体制へと拡張する必要があるが、高度なゲノム分野の知識を持ち合わせる大人数の専門家集団によるキュレーション体制を継続的に維持するのは困難である。そこで、不足するキュレータの増強の為に、群衆 (クラウド) の力を借りてコスト低減を図る。

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

本稿では、タスクを簡素化したゲノム科学分野のキュレーション作業を、実際に商用クラウドソーシング・プラットフォームからクラウドワーカに発注して精度を検証する。対象タスクは、画像アノテーションとテキストアノテーションの 2 種類に絞った。最初の画像アノテーションタスクは、未公開配列の遺伝子構造キュレーション作業として、クラウドワーカが遺伝子領域部分を画像から判定する作業である。クラウドワーカの精度を調査する目的の為に、非専門家のクラウドワーカが容易に実行できる画像注釈タスクにまとめた。2 つめのテキスト・アノテーションタスクは、提示された英文テキストの中から、分子生物学の専門用語単語の識別作業を行う。専門的知識を持つキュレータと、非専門家であるクラウドワーカのタスク精度の比較検討を行う。3 つめに、タスク自体のテキスト・データを使ってタスク専門度の定量化を行う。ゲノム分野のタスクは、単語が専門用語で記載されている為に、慣れていないと内容を理解できない。タスク毎に専門難易度を定量的に評価できれば、高度に専門的なタスクは専門家へ割当てて、それ以外の比較的容易なタスクはクラウドワーカへ割当る事が可能になる。タスク専門度を定量化する為に、自然言語処理 (NLP) の統計言語モデルを利用して、テキスト難易度を基に専門度定量尺度を定義した。以下に、2 つのアノテーション・タスク事例とテキスト専門語定量尺度の報告を、3 項に分けて報告する。

2. 画像注釈タスクのクラウドソーシング精度 (遺伝子構造領域キュレーション)

現在、ゲノム分野での主たるキュレーションタスクは、自動注釈プログラムの注釈誤りの修正である。自動注釈には、遺伝子や一塩基多型といったゲノム構造を、配列から領域推定するタスクと、構造の機能推定を行うタスクの 2 種類がある。機能推定の場合はテキスト注釈となり、専門知識が関係するタスクなので、次項以降に説明する。

ここでは専門知識が不要のタスクとして、遺伝子の構造領域推定に焦点をあてて、画像情報を基にキュレーション作業を

図 1: 画像注釈タスク (遺伝子構造領域キュレーション) の 1 例

行うタスクを紹介する。キュレーションを実際に行なう時には、ゲノムブラウザという配列領域の可視化ツールを使って、拡大・縮小操作を行ないながら複数のゲノム領域注釈情報を元に修正して行く。通常は注釈作業に専門用語が入るが、説明文から専門用語を排除して画像注釈タスクとしてデザインする事で、非専門ワーカの学習コストを下げる事が出来る。またゲノムブラウザは、拡大作業で配列情報を取得できるが、画像タスクにまとめる事で、配列情報も非提示になる。ここでは、画像注釈タスクを実施する時のクラウドワーカの精度を検討したい。簡易タスクとする為に、2 つの自動注釈結果の差分領域と、参照情報として遺伝子発現情報を提示する。横軸で両オブジェクトが重なるかを問う画像注釈タスクとして設計した。

2.1 実験条件：遺伝子構造領域キュレーションの画像注釈タスク化、被験者

キュレーション対象として、遺伝子構造領域の画像データを 38 枚生成した。オンラインで公開されている、モデル植物シロイヌナズナ (Arabidopsis thaliana) の低温条件での遺伝子発現データ [*1] を使用した。自動注釈ツールは Augustus server[Hoff 13] を利用してゲノム配列から領域注釈した結果と、遺伝子発現データを cufflinks v2.1.1[Trapnell 10] 注釈ツールを使って領域注釈した結果の差分情報を用いた。

参照情報として遺伝子発現タグの積算領域画像を使うが、描画にゲノムブラウザ IGV v2.3.26[Robinson 11] を用いた。積算画像に差分領域画像を追加して、タスク画像とした。AT4G00200 など 4 番染色体の 10 遺伝子から、38 部位の画像注釈タスクを生成した。注釈結果の正解データは、シロイヌナズナの注釈データベース TAIR[Lamesch 12] を参考に、手作業で構築した。

商用クラウドソーシング・プラットフォームのランサーズ [*2] を用いて、19 個の領域判断を 1 件にまとめて 95 円とし、2 件で各 20 名のクラウドワーカに画像注釈タスクを依頼した。クラウドワーカを集め易くする目的で、タスクの文面には専門的な表現は記載しなかった。タイトルは「グラフの重なり検出」、画像注釈手順として「それぞれの画像において、画像下部の赤色で示された領域が、灰色のグラフの領域にかかっているかどうかを判断してください。」とのみ説明した。図 1 にタスクの例を示す。参考情報として、注釈結果の事例を 3 件提示した。

2.2 実験結果：画像注釈タスクにおけるクラウドワーカ精度

クラウドワーカは、片方のみ稼働したワーカ数は 12 で、全員で 26 名である。1 件 20 名のクラウドワーカ正答率は、0.85 ± 0.12 であった。最も低い正答率は 0.53 である。判断条件の説明不足により、全ワーカが不正解のケースも見られた。判断条件のタスク明文化程度は、精度に影響する。クラウドワーカからの複数回のフィードバックを得て高精度を目指すか、低精度のままで通すか、明文化コストと精度のトレードオフは大きな問題である。

今回は少量のタスクでテストを行ったが、実際の遺伝子構造注釈はタスク数が多くなる (遺伝子数はシロイヌナズナで約 27,000)。必要なワーカ数も多くなるので、高精度ワーカを選別してタスクを発注する方式は現実的ではない。低精度のワーカも集めて、クラウドワーカの一致率を閾値として、高精度結果を選別する方式が考えられる。これには、先述の明文化コストを払って一致率と正答率の相関を高めておく必要がある。

補足情報として、画像提示条件とクラウドワーカ一致率について Spearman 順位相関検定を行った。画像の参照オブジェクトと判断用オブジェクト間の距離 (pixel) と一致率の相関係数は r=0.09(P=0.71)、オブジェクトサイズと一致率の相関係数は r=-0.13(P=0.60) と有意差はなかった。

3. テキスト注釈タスクのクラウドソーシング精度 (分子生物学単語の判別)

3.1 ゲノム科学分野でのテキスト注釈タスクの問題点

ゲノム科学分野でのキュレーションタスクとして、オンラインの雑誌論文テキストから注釈情報を抽出してデータベースを構築する場合が多い。筆者らも、一塩基多型の形質遺伝率注釈 DB(H2DB)[Kaminuma 13] や論文中の遺伝子共起関係 DB[Fujisawa 14] といった文献由来のキュレーション・データベースを公開している。文献からの知識抽出は、専門用語の知識がないクラウドワーカが実施するのはタスク完遂自体が困難であると考えられる。この為に、まずクラウドワーカと専門家であるキュレータとの精度比較実験を行う。テキスト・アノテーションタスクとして、提示された英文テキストの中から、分子生物学の専門用語単語の識別作業を行った。専門的知識を持つキュレータと、非専門家であるクラウドワーカのタスク精度について比較検討を行う。

3.2 実験条件：テキスト注釈タスクと被験者

キュレーション対象の英文テキストとして、分子生物学固有表現抽出の共通タスク [*3] で使われた、英語論文アブストラクトのデータを採用した。このデータでは、アブストラクト中の各単語が分子生物学の専門用語であるか否かの注釈が正解として与えられている。正解の専門用語は、“DNA”, “RNA”, “protein”, “cell-type”, “cell-line” のいずれかを表す語で構成されている。本実験では、ランダムに選択した 5 件の論文アブストラクトを使用した。5 件のアブストラクトの平均単語数は 187 語であった。被験者には各アブストラクトを提示し、専門用語を選択するよう依頼した。また、キュレーション作業用の画面を開いている時間を作業時間として取得した。

専門的知識をもつ 3 名のキュレータと、非専門家である 17 名のクラウドワーカが被験者として実験に参加した。キュレータは全員、分子生物学分野の文献キュレーション作業への 2 年以上の従事経験がある。2 名が博士号取得者、1 名が修士号取得者である。クラウドワーカはランサーズで雇いキュレーション作業を依頼した。キュレータには無償での作業を依頼し、クラウドワーカには 5 件のアブストラクトへの注釈作業を 1,000 円で依頼した。クラウドワーカには注釈作業の他に、生命科学の知識レベルに関するアンケートへの回答を依頼した。具体的には、生命科学に関する授業の履修経験、大学・大学院での専攻状況、生命科学に関する仕事への従事経験を回答させた。

*1 http://bioviz.org/quickload/A_thaliana_Jun_2009/cold_stress/cold_treatment.sm.bam
*2 http://www.lancers.jp
*3 http://www.nactem.ac.uk/tsujii/GENIA/ERtask/report.html

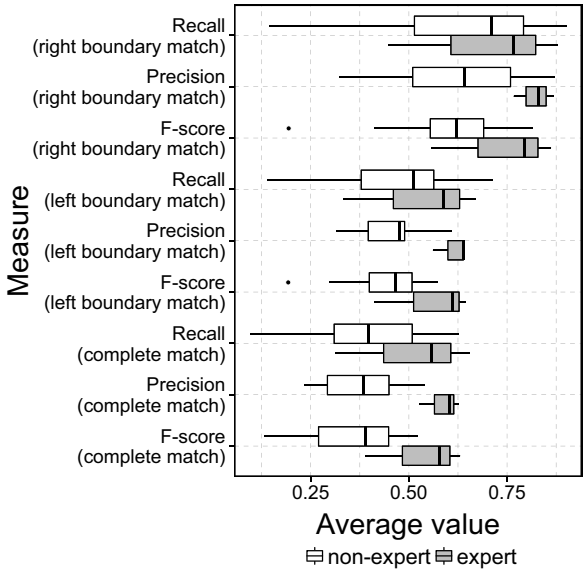

図 2: 専門家（キュレータ）と非専門家（クラウドワーカ）の回答性能の比較．5 件の作業における各性能指標の平均値の分布を示している。

また、英語力を調べるために TOEIC の点数も回答させた。生命科学の大学院での専攻経験または関連する仕事への従事経験があるのは 17 名中 3 名だった。また、TOEIC の平均点数は 675 点であった。

3.3 実験結果 : クラウドワーカとキュレータの精度比較

各被験者の回答を正解データと比較し性能を調べた。正答判定基準として「右側一致」「左側一致」「完全一致」をそれぞれ用いた。右側一致は、回答中の専門用語区間が正解データ中のいずれかの専門用語区間と終端が一致する場合に正答とみなす。左側一致も同様である。完全一致は、始端・終端の両方が一致する場合だけ正答とみなす。回答性能を測る指標として再現率、適合率、F 値をそれぞれ用いた。

図 2 にクラウドワーカとキュレータの性能比較結果を示す。いずれの指標においてもキュレータがクラウドワーカよりも高い性能を示している。一方、クラウドワーカは性能のばらつきが大きく、キュレータと同程度の性能を示すワーカが存在することが確認できた。表 1 に、全キュレータと、高性能を示したクラウドワーカ 2 名の性能を示す。どちらのワーカもキュレータ C を上回る F 値を示しているが、両者とも生命科学の知識レベルは高校で授業を受けた程度であり、また TOEIC の点数の回答は共に 550 点で英語力も決して高くはない。この結果から、専門知識をもたないクラウドワーカであっても分子生物学の専門用語判別においてキュレータ並の性能を発揮できる可能性があることが確認できた。クラウドワーカの一件辺りの平均作業時間は 493.77 秒であるがワーカ A は約 3 倍の作業時間を掛けており、このことから、専門知識がなくても時間を掛けることで精度良く作業できる可能性が示唆される。

また、適合率よりも再現率において、キュレータとクラウドワーカの性能差が小さいという傾向が観測された。この結果から、テキスト注釈タスクの二段階実施がクラウドソーシングの活用方法として有効だと考えられる。クラウドワーカにまずは再現率が高くなるように専門用語単語を選択させ、次にキュレータに、ワーカが選んだ中から不適切なものを削る作業をさせることで再現率・適合率を維持しながらキュレータの作業時間を減らす、という運用が考えられる。

表 1: キュレータと高性能を示したクラウドワーカとの性能比較。各値は 5 件の作業の平均値。再現率・適合率・F 値は完全一致を正答基準とした場合の値である。

被験者	作業時間（秒）	再現率	適合率	F 値
ワーカ A	1514.37	0.5085	0.5379	0.5219
ワーカ B	505.61	0.5736	0.4509	0.4999
キュレータ C	240.21	0.3139	0.5278	0.3888
キュレータ D	405.34	0.5574	0.6024	0.5783
キュレータ E	804.80	0.6560	0.6262	0.6306

今回の実験結果により、専門知識をもたないクラウドワーカの中にも専門家に劣らない性能を示す人がいることを確認できた。今後は、高い性能のクラウドワーカを効率的に見つけ出す方法を検討する。

4. タスク専門度定量化：NLP によるテキスト難易度推定

4.1 タスク専門度定量尺度の提案

ゲノム分野に限らずライフサイエンス全体の専門性が高いタスクの、専門度を定量化できれば、専門性の高いタスクを省いたクラウドワーカへの発注が可能になる。ここでは、タスクの専門度定量尺度化を目指す。まず専門的難易度を、タスクの提示文章の「可読難易度 (Readability)」とタスク自体の「実行難易度 (Compexity)」に分けて定義する。本稿では、タスクの可読難易度 (Text Readability/Reading Difficulty) に着目して、自然言語処理 (NLP: Natural Language Processing) の統計言語モデルを利用した専門度定量尺度を提案する。可読難易度は、N-gram 統計言語モデルを使った定量化が提案されている [Collins-Thompson 04]。本稿での専門度定量尺度 (可読難易度) と定義は、専門論文雑誌のテキストによる専門 N-gram モデルと一般向けテキストによる非専門 N-gram モデルを構築して、クエリテキスト長で規格化した対数尤度比とする。

4.2 実験条件：専門モデルと非専門モデルの構築

本稿では簡便の為、Word 単位の Unigram モデルを構築して、提案専門度尺度である対数尤度比を計算する。Unigram の場合のテキスト尤度は、単語毎の頻度から計算する。コーパスは、専門モデル構築用に米国 NCBI Pubmed データベース [NCBI 14] から論文アブストラクト・テキストを 2014 年 3 月 1 日以降出版の検索で抽出した。非専門モデル構築用には、無料コーパスの Project Gutenberg *4 から A Brief History of the United States by Joel Dorman Steele のデジタルテキストを利用した。Pubmed のデータ量は、Project Gutenberg の量に合わせて調整した。句読点・記号などは削除して単語を抽出した後で、Stop words*5 を除外した。単語頻度 (tf) を使って、差異係数は $(tf_{\text{専門}} - tf_{\text{非専門}})/(tf_{\text{専門}} + tf_{\text{非専門}})$ で計算した。提案尺度のテスト用データは、日本語テキストの場合は Google 翻訳ツール *6 で英文に変換した。

*4 http://www.gutenberg.org/
*5 http://jmlr.org/papers/volume5/lewis04a/a11-smart-stop-list/english.stop
*6 http://translate.google.co.jp/

表 2: 専門, 非専門の出現頻度の差異係数上位 5 位

ランク	専門単語	差異係数	非専門単語	差異係数
1	data	0.9836	union	−0.9906
2	properties	0.9831	british	−0.9881
3	bond	0.9780	river	−0.9875
4	analysis	0.9770	american	−0.9839
5	target	0.9706	french	−0.9839

4.3 提案専門度尺度のタスク別定量結果

専門コーパスと非専門コーパスで各々Unigram 統計言語モデルを構築した。出現単語数は、各モデルで 12,999 語と 10,479 語となった。両モデルに存在する 2,283 単語で、差異係数の大きな上位 5 単語を表 2 に示す。χ^2 検定で出現頻度に有意差 (P< 0.05) があった単語の割合は、専門 18%、非専門 19%だった。

提案専門度尺度のタスク別計算結果を、Lancers(画像注釈タスク), CrowdSolving[*7](予測モデル構築タスク), H2DB(専門文献の注釈タスク), CROWD4U[*8][Morishima 12](テキスト・画像注釈タスク), KOKOPIN[*9](テキスト・画像注釈タスク), に分けて図 3 に示す。縦軸は専門/非専門の対数尤度比を表し、値が高いタスク程、専門性が高いと評価される。H2DB タスクの提案尺度計算結果を見ると、全てのタスクで専門評価尺度値が高い訳ではなく、専門度低値のタスク切出に役立つ。また H2DB タスクより、CrowdSolving の予測タスクは評価値が高かった。一方、KOKOPIN プラットフォームでは専門的タスクと非専門的タスクが混在している様子が伺える。

今後、タスクのテキストの専門用語出現割合を客観的データとして収集し、提案尺度の有効性を検証していく。また提案尺度は可読難易度のみを反映している。タスク実行難易度のデータも収集していき、注釈時間予測ツールを構築していきたい。

まとめと今後の展望

本稿では、ゲノム科学分野のキュレーション作業を簡素化した、遺伝子構造領域の画像注釈と専門用語同定のテキスト注釈の 2 種類のタスクを商用クラウドソーシング・プラットフォームを通して実行し、精度を検証した。遺伝子構造領域の注釈タスクの方は、専門性の高いクラウドワーカを必要としない。一方、テキスト注釈タスクは専門性が問題になる。非専門家に割振るための専門性の低いタスクを選別する為に、タスク提示文章の「可読難易度 (Readability)」を Unigram 統計言語モデルを使って推定する専門度の客観的評価尺度を提案した。

今後は、遺伝子領域注釈や文献情報抽出の実 DB 構築用のデータを、クラウドソーシング・タスクにまとめて精度を検証していく予定である。更にキュレーション・タスクを Semantic Web 技術によって統合する、RDF データを整備中である。クラウドソーシング用のプロトコルを構築して RDF データを用いた効率的なタスク生成ワークフローを構築していく。

*7 http://crowdsolving.jp/
*8 http://crowd4u.org/
*9 http://www.kokopin.com/

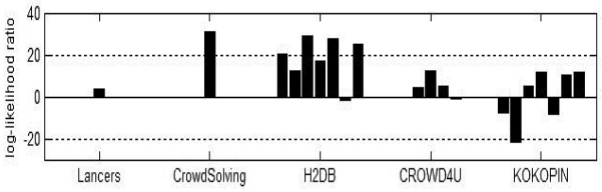

図 3: 専門度提案尺度のタスク別計算結果。縦軸は対数尤度比 (専門/非専門)。

Acknowledgements

We are thankful to researchers constructing crowdsourcing platforms in section 4 : Atsuyuki Morishima(Tsukuba University), Osamu Matsuda (Kyushu University), Shigeru Saito (OPT Inc.), and members of National Institute of Genetics : Kazuo Hara, Takako Mochizuki, Yasuhiro Tanizawa, Hideki Nagasaki.

参考文献

[Kaminuma 10] Kaminuma, E., 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).

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

[Kaminuma 13] Kaminuma, E., Fujisawa, T., et al. : a heritability database across multiple species by annotating trait-associated genomic loci, Nucleic Acids Res, 41, pp.D880-4 (2013). http://tga.nig.ac.jp/h2db/.

[Robinson 11] Robinson, J.T., Thorvaldsdttir, H., et al. : Integrative genomics viewer, Nat Biotech, 29, pp.24-26 (2011).

[Trapnell 10] Trapnell., C., Williams, B.A. et al. : Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation, Nat Biotech, 28, pp.511-515 (2010).

[Hoff 13] Hoff, K.J. and Stanke, M. : WebAUGUSTUS – a web service for training AUGUSTUS and predicting genes in eukaryotes, Nucleic Acids Research, 41, W123-W128 (2013).

[Lamesch 12] Lamesch, P., Berardini, T.Z., et al. : The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools, Nucleic Acids Research, 40, D1202-10 (2012).

[Collins-Thompson 04] Collins-Thompson, K. and Callan, J. : A language modeling approach to predicting reading difficulty, Proceedings of HLT / NAACL 2004, (2004).

[NCBI 14] NCBI Resource Coordinators : Database resources of the National Center for Biotechnology Information, , Nucleic Acids Res, 42, pp.D7-17 (2014).

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