CROWDSOURCING AND MINING CROWD DATA

ROBERT LEAMAN†

National Center for Biotechnology Information (NCBI)
8600 Rockville Pike, Bethesda, MD 20894, USA
Email: robert.leaman@nih.gov

BENJAMIN M. GOOD‡

Department of Molecular and Experimental Medicine, The Scripps Research Institute
10550 North Torrey Pines Road, La Jolla, CA 92037, USA
Email: bgood@scripps.edu

ANDREW I. SU‡

Department of Molecular and Experimental Medicine, The Scripps Research Institute
10550 North Torrey Pines Road, La Jolla, CA 92037, USA
Email: asu@scripps.edu

ZHIYONG LU†

National Center for Biotechnology Information (NCBI)
8600 Rockville Pike, Bethesda, MD 20894, USA
Email: zhiyong.lu@nih.gov

1. Introduction

The "crowd" is that body of people that will either respond to an open call for participation ("crowdsourcing") or who, through the actions they take in public forums, leave behind a trail of information that can be mined to identify new knowledge ("crowd data"). This session considers a variety of approaches utilizing the crowd as a resource to enable biomedical discovery.

While the family of crowdsourcing methodologies is still being actively created, the existing approaches are already diverse [1]. Well-known examples include microtask environments where workers are paid to perform discrete tasks, including Amazon Mechanical Turk [2], games with a purpose such as FoldIt [3], collaborative content creation frameworks like Wikipedia [4], and systems like Twitter that produce repositories of crowd data [5]. The advantages of crowd-driven

† Work supported by NIH Intramural Research Program, National Library of Medicine.

‡ Work supported by National Institute of General Medical Sciences of the National Institutes of Health (R01GM089820 and R01GM083924) and by the National Center for Advancing Translational Sciences (UL1TR001114).

approaches include reduced cost, increased data sizes, and environments closer to those in the real world. These characteristics may ultimately enable research not possible via traditional methods.

Crowdsourcing remains challenging for several reasons, however. The overall problem being addressed must be decomposed into tasks appropriate for a heterogeneous population, typically with minimal training. Suitable incentive strategies need to be devised and implemented. The responses from multiple members of the crowd must be aggregated to produce a high-quality signal. As these are all new challenges, the emerging protocols and resulting data must be validated using robust analyses and evaluation.

2. Session articles

While the six articles accepted to this session address a wide variety of computational tasks within biomedicine, their use of crowdsourcing falls within three primary themes.

The first pair of articles evaluate the ability of novice workers in microtask environments. Irshad et al. use the CrowdFlower microtask platform to annotate images to detect and segment the nuclei of cells. They compare crowdsourced annotations against those performed by pathologists, reporting f-measures as high as 0.885 at a fraction of the time and cost. They report that performance degrades significantly with larger images. Good et al. use Amazon Mechanical Turk to create an annotated text corpus of disease mentions in PubMed abstracts. Their methodology can alternately emphasize precision or recall, or balance cost and quality. They demonstrate an f-measure of 0.872 against gold-standard data, again at a fraction of the cost and time. Both articles conclude that crowdsourcing in a microtask environment can be an effective way to generate annotated datasets.

The second pair of articles stretches the concept of the crowd beyond the more typical lay-public to include focused groups of experts. Both articles involve crowds of experts, but they differ substantially in their approach and problem area. Binder et al. build a collaborative reputation-based system for describing the biology of protein networks, and apply it to curate pathways relevant to chronic obstructive pulmonary disease. Their crowdsourcing experiment resulted in the submission of 885 pieces of evidence, with a validity rate of 77%. Tastan et al. query multiple experts to determine whether protein-protein interactions described in the published literature on HIV represent a physical or indirect interaction. They use a probabilistic latent variable model to jointly estimate the accuracy of each annotation and the probability of each interaction being correct. They evaluate their method with synthetic data, demonstrating significant improvements over the consensus baseline.

The third pair of articles uses the crowd for exploratory analysis. Waldispühl et al. create an online game for structural alignment of non-coding RNA, a computationally challenging problem. Players explore potential alignments via tile-matching, gaining points for finding better alignments, resulting in a casual game similar to Phylo [6]. Odgers et al. note the inadequacy of relying on spontaneous reports of adverse drug events and evaluate whether the search logs of healthcare professionals could be a useful source of signal. Their method provides an AUC of 0.85

for well-known adverse reactions and an AUC of 0.68 for adverse reactions described recently, suggesting the possibility of also detecting novel adverse reactions.

Taken together, the articles presented in this session provide a rich sample of the many emerging efforts to harness the wisdom of the crowd for biomedical research.

Acknowledgments

The authors thank the many anonymous reviewers for their generous assistance and insight as well as the thousands of members of the crowd that collectively made this session possible.

References

1. Good BM, Su AI: **Crowdsourcing for Bioinformatics**. *Bioinformatics* 2013, **29**(16):1925-1933.
2. Burger JD, Doughty E, Khare R, Wei CH, Mishra R, Aberdeen J, Tresner-Kirsch D, Wellner B, Kann MG, Lu Z *et al*: **Hybrid curation of gene-mutation relations combining automated extraction and crowdsourcing**. *Database (Oxford)* 2014, **2014**.
3. Cooper S, Khatib F, Treuille A, Barbero J, Lee J, Beenen M, Leaver-Fay A, Baker D, Popovic Z, Players F: **Predicting protein structures with a multiplayer online game**. *Nature* 2010, **466**(7307):756-760.
4. Finn RD, Gardner PP, Bateman A: **Making your database available through Wikipedia: the pros and cons**. *Nucleic Acids Res* 2012, **40**(Database issue):D9-12.
5. Eysenbach G: **Infodemiology and infoveillance: framework for an emerging set of public health informatics methods to analyze search, communication and publication behavior on the Internet**. *Journal of medical Internet research* 2009, **11**(1):e11.
6. Kawrykow A, Roumanis G, Kam A, Kwak D, Leung C, Wu C, Zarour E, Sarmenta L, Blanchette M, Waldispuhl J: **Phylo: a citizen science approach for improving multiple sequence alignment**. *PloS one* 2012, **7**(3):e31362.