Citizens unite for computational immunology!

Orrin S. Belden[*], **Sarah Catherine Baker**[*], and **Brian M. Baker**[**]

Department of Chemistry and Biochemistry and the Harper Cancer Research Institute, 251 Nieuwland Science Hall, University of Notre Dame, Notre Dame, IN 46556

Abstract

Crowdsourcing by recruiting volunteers who can provide computational time, programming expertise, or puzzle-solving talent has emerged as a powerful tool for biomedical research. Recent projects demonstrate the potential for crowdsourcing in immunology. Tools for developing applications, new funding, and an eager public make crowdsourcing a serious option for creative solutions for computationally-challenging immunological problems. Expanded uses of crowdsourcing in immunology will allow for more efficient large-scale data collection and analysis. It will also involve, inspire, educate, and engage the public in a variety of meaningful ways. The benefits are real – it's time to jump in!

Keywords

Crowdsourcing; computational biology; distributed computing; immunology

Crowdsourcing – the process of enlisting large groups of online volunteers to address computational problems – has emerged as a powerful tool for creative problem solving in science, engineering, and technology. Its usefulness is evident by the fact that it has no geographic boundaries, is inexpensive, and can quickly accelerate research. Early efforts in crowdsourcing began in the 1990s, focusing on problems in mathematics and astronomy. Since that time, advances in computing and the almost universal adoption of the internet as a means of networking and communicating have resulted in dozens if not hundreds of crowdsourcing efforts, many of which are biomedically-focused and have direct applications in immunology.

Crowdsourcing often utilizes idle, internet-connected personal computers (PCs) in a process known as distributed computing [1]. In these projects, volunteers download software – often a screensaver – which connects to a central server that assigns the PC a small "chunk" of a larger computational project. One of the first and most widely recognized crowdsourcing projects, SETI@home, uses a screensaver running on idle PCs to examine radio telescope data for possible signals from extraterrestrials. Advances in hardware and connectivity have allowed SETI@home and other distributed computing projects to evolve beyond applications such as screensavers for PCs, and projects now include applications for gaming consoles and even mobile phones.

[**]Corresponding author: brian-baker@nd.edu, (574) 631-9810.
[*]Equal author contribution

One of the first biomedical projects to take advantage of the power of distributed computing is Folding@home from the laboratory of Vijay Pande at Stanford. Folding@home aims to understand the pathways and mechanisms of protein folding and mis-folding, and has been used to study diseases such as Alzheimer's and Huntington's, as well as cancer, malaria, and Parkinson's. With applications for all common computer operating systems and Android-powered phones, Folding@home now claims over 180,000 participating machines and a computing power of over 46 quadrillion floating-point calculations per second, making it competitive with the world's most powerful supercomputers (for those keeping score, the Folding@home project is in the petaflop range!).

A more interactive approach for crowdsourcing relies on adapting scientific problems into downloadable games. The most successful of these is Foldit from David Baker's laboratory at the University of Washington. Players of Foldit, who are typically non-scientists, compete to predict and optimize protein structures [2, 3]. The ranking system has identified an elite cadre of "super folders" who are able to intuitively grasp the principles of protein structure and folding. The efforts of Foldit players have led to not only successful protein structure prediction, but also the design of enzymes with enhanced activity [4]. The broad appeal of this platform is undeniable: currently over 240,000 Foldit players are attempting to surmount the roadblocks of molecular design so familiar to "traditional" structural biologists and biophysicists. Beyond the direct applicability of the game, analyses of the "out of the box" strategies adopted by successful Foldit players and their incorporation into automated algorithms may lead to wholly new breakthroughs in molecular design. In addition to Folding@home and Foldit, numerous other biomedically-oriented crowdsourcing projects have arisen, such as efforts directed at DNA and RNA origami, protein docking, multiple sequence alignments, and genome annotation.

How do these efforts impact immunology? Fundamental issues such as protein folding and complex bioinformatics tasks have an obvious immunological connection, yet many direct applications in immunology are both ongoing and emerging. Arguably, the most successful biomedical crowdsourcing efforts have addressed problems rooted in structural biology. Structural biology in turn has revolutionized immunology, with the structures of major histocompatibility complex proteins and their complexes with T cell receptors perhaps the most commonly recognized examples. Structural biology is now routinely applied to efforts in the design of vaccines and other immunologically-based therapeutics. The structural simulation methods incorporated into Folding@home were used recently by Chris Garcia and Vijai Pande to help engineer a high affinity variant of interleukin-2 with potential antitumor activity [5]. The Rosetta design platform [6] incorporated into Foldit has been used by our own group to engineer high affinity T cell receptors for possible use in cancer immunotherapy [7]. Pultz and colleagues used Foldit to redesign an acid-stable peptidase to target the proline-glutamine enriched peptides that underlie the inflammatory responses in Celiac disease [8]. IBM's World Community Grid has hosted several projects that utilize virtual screening or structure-based drug design to target pathogens, including the long-running FightAIDS@home project, which seeks to identify HIV protease inhibitors.

As structural biology crowdsourcing efforts broaden and increase in their complexity, opportunities to directly address immunological problems rooted in structural biology also

expand. Possible examples include simulations of large molecular assemblies such as viruses or receptors in their membrane environments, wide-scale assessments of immunoglobulin cross-reactivity, structural visualization of signaling cascades, and the structure-based design of novel therapeutics. Another potential application is the examination of how molecular motion impacts immune recognition and signaling, as discussed in the recent *Trends in Immunology* review from Natalie Borg and colleagues [9]. An exciting use of structural-based crowdsourcing could be the use of modeling and simulation to predict conformational epitopes and tumor neo-antigens [10, 11], with the aesthetically pleasing outcome that the public at large would be participants in the development of personalized medicine. Although Folding@home and Foldit remain connected with individual laboratories and are not fully open source, the availability of platforms such as the Berkeley Open Infrastructure for Network Computing and IBM's World Computing Grid provide a means for generating new distributed computing tools.

Another example of a crowdsourcing approach applied to immunology did not use idle computers or games, but instead challenged programmers at the online community Topcoder. Topcoder members are "programmers for hire" who compete for prize money by improving or designing new computational algorithms. Eva Guinan and colleagues challenged Topcoders to generate a faster approach for identifying variable gene segments in massive sequences libraries of antibody and T cell receptor genes [12]. A key innovation was recasting the problem in terms typically encountered in generic pattern matching problems, devoid of references to genetics or immunobiology. One hundred and twenty two submissions from 69 countries were received over 14 days. Many of the submissions showed improved accuracy and speed relative to both the public BLAST-based tool hosted by the NCBI as well as custom software developed by the authors' laboratory; remarkably, speed enhancements for the most successful submissions were as much as three orders of magnitude faster than the BLAST-based tool.

A slightly different take on leveraging outside expertise to address immunological problems is the Rheumatoid Arthritis Responder Challenge, which aims to improve the analysis of data from genome-wide association studies (GWAS). In an effort to improve predictions of how rheumatoid arthritis patients respond to anti-TNF therapy, Robert Plenge and colleagues designed an open competition in which participants created genetic models of the response to anti-TNF therapy from GWAS data on thousands of rheumatoid arthritis patients [13]. The competition was hosted by Synapse, a platform for collaborative, open biomedical data analysis developed by Sage Bionetworks. The difference here is that the competition was pitched to experts, but nonetheless recruited numerous participants from academia, private foundations, and for-profit companies. This example highlights the potential for crowdsourcing to catalyze the formation of diverse, interdisciplinary research teams that are becoming ever more critical for addressing today's complex problems.

With the growth of crowdsourcing in biomedical research, some cautionary notes have emerged. Graber and Graber emphasized that recruitment of participants is subject to ethical concerns [14], particularly in projects which reward players with scores and online "bragging rights." Such incentives could have impacts on the social, economic, or even health status of participants. Further, each project is in itself an experiment with human

subjects who may be minimally aware of the scientific goals and methods. For instance, the Foldit project described above is essentially an experiment that asks whether "human" strategies can be found and incorporated into novel design algorithms. Additionally, some crowdsourcing strategies may marginalize the effort of anonymous participants, although there is evidence to suggest this is not the case [3]. Although the ethical ramifications require further study, a proposed solution to these concerns is to obtain institutional review board approval for crowdsourcing projects.

These cautionary notes should of course be considered alongside the broader benefits, which are many. As mentioned previously, crowdsourcing has the potential to gather large amounts of data quickly and at relatively low costs. There are limited geographic constraints, which reduces the barriers for forming meaningful collaborations and diverse research teams. Finally, and perhaps most importantly, crowdsourcing can be exploited as a platform for immunologists to educate and communicate at a broad scale, as the process hinges on public involvement. With these projects come the education, involvement, and inspiration of scientist and non-scientist alike; participants are not only made aware of the goals and basic principles behind a project, but are inducted into the very process of scientific discovery. Public awareness is critical to the overall health of science, particularly given the dependence on taxpayer funds, and the potential impact on future trainees cannot be underestimated ("Mommy, I folded a protein!"). More soberly, every day brings another reason to fight for improved scientific literacy, and crowdsourcing can be an effective means to contribute.

In conclusion, as the computing power and the complexity of addressable problems expands, so does the potential for immunological applications in crowdsourcing. Recognizing this potential, the US National Institutes of Health recently requested grant applications for development of "interactive digital media that engages the public, experts or non-experts" in the analysis or interpretation of biomedical data. The results of this legitimizing (if somewhat skinny) "dip" by NIH into the waters of crowdsourcing will be telling. If judged successful, such efforts should undoubtedly expand, as crowdsourcing is not only capable of efficiently producing publishable data on both smaller time scales and smaller budgets, but could also be responsible for a new generation of scientific literacy and appreciation among non-scientists.

References

1. Shirts M, Pande VS. Screen Savers of the World Unite! Science. 2000; 290:1903–1904. [PubMed: 17742054]

2. Cooper S, Khatib F, Treuille A, Barbero J, Lee J, Beenen M, players, F. Predicting protein structures with a multiplayer online game. Nature. 2010; 466:756–760. [PubMed: 20686574]

3. Cooper S, Khatib F, Baker D. Increasing Public Involvement in Structural Biology. Structure. 2013; 21:1482–1484. [PubMed: 24010706]

4. Eiben CB, Siegel JB, Bale JB, Cooper S, Khatib F, Shen BW, Baker D. Increased Diels-Alderase activity through backbone remodeling guided by Foldit players. Nat Biotech. 2012; 30:190–192.

5. Levin AM, Bates DL, Ring AM, Krieg C, Lin JT, Su L, Garcia KC. Exploiting a natural conformational switch to engineer an interleukin-2/'superkine/'. Nature. 2012; 484:529–533. [PubMed: 22446627]

6. Kaufmann KW, Lemmon GH, DeLuca SL, Sheehan JH, Meiler J. Practically Useful: What the Rosetta Protein Modeling Suite Can Do for You. Biochemistry. 2010; 49:2987–2998. [PubMed: 20235548]

7. Pierce BG, Hellman LM, Hossain M, Singh NK, Vander Kooi CW, Weng Z, Baker BM. Computational Design of the Affinity and Specificity of a Therapeutic T Cell Receptor. PLoS Comput Biol. 2014; 10:e1003478. [PubMed: 24550723]

8. Gordon SR, Stanley EJ, Wolf S, Toland A, Wu SJ, Hadidi D, Siegel JB. Computational Design of an α-Gliadin Peptidase. Journal of the American Chemical Society. 2012; 134:20513–20520. [PubMed: 23153249]

9. Kass I, Buckle AM, Borg NA. Understanding the structural dynamics of TCR-pMHC interactions. Trends in Immunology. 2014; 35:604–612. [PubMed: 25466310]

10. Duan F, Duitama J, Al Seesi S, Ayres CM, Corcelli SA, Pawashe AP, Srivastava PK. Genomic and bioinformatic profiling of mutational neoepitopes reveals new rules to predict anticancer immunogenicity. The Journal of Experimental Medicine. 2014; 211:2231–2248. [PubMed: 25245761]

11. Yadav M, Jhunjhunwala S, Phung QT, Lupardus P, Tanguay J, Bumbaca S, Delamarre L. Predicting immunogenic tumour mutations by combining mass spectrometry and exome sequencing. Nature. 2014; 515:572–576. [PubMed: 25428506]

12. Lakhani KR, Boudreau KJ, Loh PR, Backstrom L, Baldwin C, Lonstein E, Guinan EC. Prize-based contests can provide solutions to computational biology problems. Nat Biotech. 2013; 31:108–111.

13. Plenge RM, Greenberg JD, Mangravite LM, Derry JMJ, Stahl EA, Coenen MJH, Stolovitzky G. Crowdsourcing genetic prediction of clinical utility in the Rheumatoid Arthritis Responder Challenge. Nat Genet. 2013; 45:468–469. [PubMed: 23619782]

14. Graber MA, Graber A. Internet-based crowdsourcing and research ethics: the case for IRB review. Journal of Medical Ethics. 2012

Box 1

Web links of interest

Folding@home

A distributed computing project developed by the Pande Laboratory at Stanford University that utilizes idle personal computers for disease research. Folding@home software is available for download for anyone that with an internet-connected computer (Windows, Mac, or Linux) or an Android tablet phone. http://folding.stanford.edu

Fold It

An online video game with protein folding puzzles developed by University of Washington researchers, with goals of solving immediate problems as well as identifying out of the box human solutions. The Foldit video game is available for download for Windows, Mac, and Linux platforms. http://fold.it

Rosetta@home

A distributed computing project developed by the Baker Laboratory at the University of Washington that utilizes idle personal computers for protein structure prediction. http://boinc.bakerlab.org/

Berkeley Open Infrastructure for Network Computing

An open-sourced grid computing software system developed that employs central servers to distribute work units to idle computers and can be used to build distributed computing tools. http://boinc.berkeley.edu/

IBM World Community Grid

A distributed computing system that hosts projects for several research teams, allowing participants to get involved via computers or Android tablets or phones. World Community Grid also solicits proposals for new research projects. http://worldcommunitygrid.org

Topcoder

A company that holds fortnightly computer programming contests emphasizing the development of new and more efficient algorithms. Topcoder contests are open to everyone who registers. http://topcoder.com

Sage Bionetworks

A nonprofit organization dedicated to collaboration data analysis. http://sagebase.org/

NIH RFA for advancing biomedical science via crowdsourcing

A recent NIH request for applications for developing crowdsourcing applications to address biomedical problems. http://grants.nih.gov/grants/guide/rfa-files/RFA-CA-15-006.html