Prize-based contests can provide solutions to computational biology problems

To the Editor:

Advances in biotechnology have fueled the generation of unprecedented quantities of data across the life sciences. However, finding analysts who can address such 'big data' problems effectively has become a significant research bottleneck. Historically, prize-based contests have had striking success in attracting unconventional individuals who can overcome difficult challenges. To determine whether this approach could solve a real big-data biologic algorithm problem, we used a complex immunogenomics problem as the basis for a two-week online contest broadcast to participants outside academia and biomedical disciplines. Participants in our contest produced over 600 submissions containing 89 novel computational approaches to the problem. Thirty submissions exceeded the benchmark performance of the US National Institutes of Health's MegaBLAST. The best achieved both greater accuracy and speed (1,000 times greater). Here we show the potential of using online prize-based contests to access individuals without domain-specific backgrounds to address big-data challenges in the life sciences.

The advent of high-throughput biology has resulted in studies that routinely generate gigabytes of data. It has been estimated that genome data available for analysis will grow from petabytes (10^{15}) to exabytes (10^{18}) (ref. 1). Finding people with the necessary training to analyze big data has become a bottleneck in moving from sequencing to discovery[2]. This supply-demand problem extends beyond genomics to other life science areas[3–5] and to many diverse industries. It has been projected that by 2018 there will be a shortage of approximately 200,000 data scientists and 1.5 million other individuals in the US economy with sufficient training and skills to conceptualize and manage big-data analyses[6].

In academia and elsewhere, this bottleneck is more than just a personnel shortage. Available personnel may lack experience with the specific approach or techniques required.

As an alternative to an extensive search to identify and contract with potentially suitable analysts, prize-based contests[7–11] have emerged as a novel approach to find solutions to challenging problems in settings as diverse as industrial R&D, software development and internet commerce. Such contests are one part of a decade-long trend toward solving science problems through large-scale mobilization of individuals by what the popular press refers to as 'crowdsourcing'[12]. In general, crowdsourcing has come to imply a strategy that relies on external, unaffiliated actors to resolve a particular problem. It encompasses a range of approaches intended to accomplish tasks from rote mechanical to highly intellectual problem solving. Strategies can enable cooperation and knowledge sharing among participants or create competitions; may limit entry to specified communities or allow universal participation; and may offer pecuniary and/or non-pecuniary incentives. Lastly, strategies may necessitate creation and use of complex infrastructure such as specialized scientific gaming websites, or they may simply require internet access.

Computational biology has become fertile ground for experimentation with various crowdsourcing approaches. One approach is to transform the problem into a game that nonscientists can play without substantial knowledge of the underlying scientific principles. For example, visual representation of molecular interactions in the Foldit game format has attracted participation by individuals without training in molecular biology or biochemistry to solve protein structure prediction problems that have eluded resolution by systematic, expert research programs[13–15]. Additional tools permit players to generate protein-folding algorithms, the best of which is equivalent to those emerging from academia[16]. Other efforts have focused on soliciting solutions from a larger expert community by moving questions beyond one laboratory or institution. For example, the Critical Assessment of Genome Interpretation project (CAGI; https://genomeinterpretation.org/) focuses on analyses of real data to predict biologically relevant information, and the CLARITY challenge[17] asks scientific teams to sequence three patient genomes (i.e., teams must have access to sequencing infrastructure), identify the variants and develop clinical reporting formats.

Over the last ten years, online prize-based contest platforms have emerged to solve specific scientific and computational problems for the commercial sector. These platforms, with solvers in the range of tens to hundreds of thousands, have achieved considerable success by exposing thousands of problems to larger numbers of heterogeneous problem-solvers and by appealing to a wide range of motivations to exert effort and create innovative solutions[18,19]. The large number of entrants in prize-based contests increases the probability that an 'extreme-value' (or maximally performing) solution can be found through multiple independent trials; this is also known as a parallel-search process[19]. In contrast to traditional approaches, in which experts are predefined and preselected, contest participants self-select to address problems and typically have diverse knowledge, skills and experience that would be virtually impossible to duplicate locally[18]. Thus, the contest sponsor can identify an appropriate solution by allowing many individuals to participate and observing the best performance. This is particularly useful for highly uncertain innovation problems in which prediction of the best solver or approach may be difficult and the best person to solve one problem may be unsuitable for another[19].

The prize-based contest approach used here differs from other academic crowdsourcing efforts in multiple ways: (i) it uses an existing commercial platform with a built-in membership base of computer-science solvers able to immediately attack the problem and able to deliver submissions in the multiple hundreds, as compared to tens for the Dialogue on Reverse Engineering

Assessment and Methods (DREAM)[20] project and CLARITY; (ii) sponsors do not need to develop specialized problem-solving infrastructure (for example, online games interfaces); (iii) it is generalizable to any life sciences problem that can be translated into generic computer-science terms (in contrast, Foldit, for example, is specifically designed to address spatial protein-folding problems); and (iv) it delivers working algorithms rapidly (weeks from launch). Yet it remains to be determined whether academic biomedicine problems are amenable to solving via these ready-made platforms.

Therefore, we experimented with the application of a prize-based contest to solve a data-rich biological problem related to immune repertoire profiling, in which short, recombined and mutated stretches of genetic sequence had to be annotated according to their constituent gene components[5,21–23]. The specificity of immune cells for particular antigens depends on which antibodies B cells secrete and which T cell receptors (TCRs) T cells express, which in turn depends on the sequence of their antibody or TCR genes. Unlike other genes, those for antibodies and TCRs are not encoded as single genes. Instead, they are built up combinatorially in each cell from gene segments; thus each new cell can have a gene with a different DNA sequence[24]. Diversity is increased further by insertion or deletion of nucleotides at the junctions (joins) between segments, by mutation in the resulting gene, and/or by combinatorial pairing of the encoded protein product. Thus a relatively small number of gene segments (<100 for antibody heavy chains in humans) can lead to an extraordinary number of different molecules (~10^{30}). An obligate step toward making sense of this diversity is annotating sequence according to which gene segments contribute to each recombined gene. This exercise is particularly challenging because the segments are often short, and recombined genes routinely have numerous insertions, deletions and substitutions (mutations).

Our goal in the prize contest was the creation of an algorithmic solution with better performance characteristics than sequence-annotation approaches using established methods such as BLAST[25] or IMGT/V-QUEST[26]. As a typical sequencing run for these genes produces on the order of 10^5 sequences[5,21–23], we sought solutions that could annotate this many sequences, with at least comparable accuracy to existing solutions, in much less time, which we defined as ≤30 s on an off-the-shelf desktop computer with ≥1GB memory.

Three crucial steps were taken to transform the highly domain-specific immunogenomics problem into a challenge that would be of interest and attractive to non–life science solvers. First, we rephrased our problem in generic terminology: given 10^5 strings (gene sequences), each generated as the union of three substrings (gene segments; one from each of three sets A, B and C of known substrings), with multi-letter (polynucleotide) insertions and deletions at the junctions and substitutions in the final string, write and implement an algorithm that determines the original three substrings that contributed to each string. This produced a problem statement devoid of biological concepts and presented an information-theory and string-processing task that a computational expert could tackle. The crucial element was removal of all context-specific information and requirements for using existing, preferred approaches so that solvers with heterogeneous backgrounds had the freedom to apply their own diverse perspectives and heuristics to create their solutions[27].

Second, we assembled the required test data for solution generation and scoring. Standard practice in this contest platform is to generate three independent test suites: a public training set for contestants, a private set to enable real-time algorithm evaluation by contestants, and, to prevent over-fitting, a private set for scoring of final submissions by the contest administrators. Our test sets used all known antibody heavy-chain V, D and J gene segments and the nucleotide insertion, deletion and substitution (mutation) frequencies observed in actual antibody heavy-chain sequences[5]. The development of test suites requires considerable care as contestants will carefully examine the data's nuances and characteristics, and artifacts in the data that may be spuriously correlated with results will be discovered and probably exploited for competitive advantage.

Third, we devised a scoring metric that supported our goal of achieving both improved accuracy and computational efficiency (speed), and we disclosed this metric and its components to the contestants. The score was the only metric used to award prizes. All information made available to participants is detailed in **Supplementary Notes**.

We ran our contest on the TopCoder.com online programming competition website, a commercial platform that had the advantage of providing us with an existing community of solvers. Established in 2001, TopCoder currently has a community of over 400,000 software developers who compete regularly to solve programming challenges[13]. Our contest ran for two weeks and offered a $6,000 prize pool, with top-ranking players receiving cash prizes of up to $500 each week. Our challenge drew 733 participants, of whom 122 (17%) submitted software code. This group of submitters, drawn from 69 countries, were roughly half (44%) professionals, with the remainder being students at various levels. Most participants were between 18 and 44 years old. None were academic or industrial computational biologists, and only five described themselves as coming from either R&D or life sciences in any capacity.

Consistent with usual practices in algorithm and software development contests, participants were able to make multiple code submissions to enable testing of solutions and participant learning and improvement. Collectively, participants submitted 654 solutions, averaging to 5.4 submissions per participant. Participants reported spending an average of 22 h each developing solutions, for a total of 2,684 h of development time. Final submissions that received cash awards are available for download under an open source license (see **Supplementary Notes**).

Apart from the use of test suites to rank-order the solutions, we also evaluated the accuracy and speed of each participant's final submission by testing it on an *in silico* set of 10^5 antibody heavy-chain sequences (using the same seed sequences as in the contest test suite) on a desktop computer

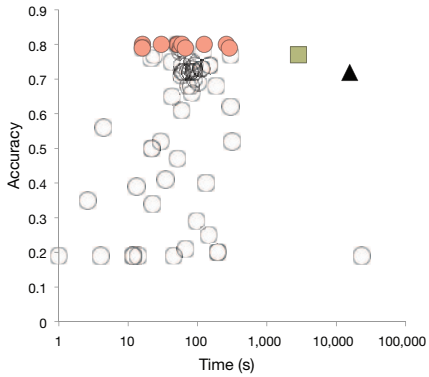

Figure 1 Accuracy score plotted against speed of contest-commissioned immunoglobulin sequence annotation code. Shown are the accuracy and speed (computational time) of 70 final submissions (top ten in red circles; remainder in unfilled circles), MegaBLAST (triangle) and in-house code (square) processing 100,000 *in silico*–generated recombined VDJ sequences. Accuracy score is the fraction of gene segment annotations that matched the corresponding gene segments used in generating the sequence. Note that because the sequence contains mutations, the best annotation might differ from the segment used, resulting in a maximum possible accuracy of <1.0.

Table 1 Elemental techniques used (in combinations) by contestant participants

Method	Description
1	**Filtering by ungapped alignment score (Hamming distance):** Compare the query string against strings from sets A, B and/or C, trying various possible offsets.
2	**Filtering by comparing frequencies of hashed chunks:** For both the query string and strings from A, B and/or C, move a sliding window across the string and make a frequency table of the chunks that appear in the window, optionally after hashing the chunks. Select the best match(es) between the frequency table obtained from the query and those from the corpus.
3	**Dynamic programming:** Compute the actual Levenshtein distance between a portion of the query and strings from sets A, B and/or C.
4	**Dynamic programming extended to more than one section (A, B, C) at once:** Extend the dynamic programming Levenshtein distance computation to find the optimal edit distance between (a portion of) the query and all possible A+B, B+C or A+B+C combinations.
5	**Bit optimizations:** Use bitwise arithmetic to operate on multiple characters at a time.
6	**SSE optimizations:** Use Streaming SIMD Extensions (a CPU instruction set enabling single-instruction multiple-data (SIMD) parallelization) to process up to 16 characters or strings at once.
7	**Refinement of choices after finding initial solution:** As a post-processing step, hold two of the three selections fixed and reoptimize the third.
8	**Fast approximation of edit distance in well-matched regions:** Use restricted dynamic programming, Hamming distance or variants thereof to speed up the computation.
9	**Precomputation of statistics on the string corpus:** Perform offline analysis of the provided sets A, B and C, and use the data obtained for decision making in the algorithm.
10	**Explicitly prefer shorter B strings:** In heuristic approaches, give bonuses to shorter strings from set B (which empirically have greater likelihood of producing high scores).

(2009 iMac Intel Core i5 2.66 GHz with 8GB RAM; **Fig. 1**). We tested these submissions directly against existing industry-standard algorithmic solutions. An approach based on the US National Center for Biotechnology Information's MegaBLAST (**Fig. 1**, black triangle) and our coauthor's (R.A.A.) custom annotation software, idAb (**Fig. 1**, green square), served as benchmarks (**Supplementary Methods**)[5].

Sixteen of the 122 submissions outperformed the accuracy (77%) of the idAb solution, and 30 outperformed the MegaBLAST benchmark for accuracy (72%). Furthermore, eight submissions achieved an 80% accuracy score, which is very near the theoretical maximum for the data set. The remaining error corresponds to sequences that cannot be correctly annotated, owing to either removal of D-segment sequence, truncation or mutation (**Supplementary Methods**).

There was also a notable improvement in speed over both benchmark algorithms. Submissions that were at least as accurate as the benchmarks ran, on average, 30 times faster than idAb on an expanded test suite of 1 million sequences (average of 89 s, compared with 2,845 s for idAB) and 175 times faster than MegaBLAST (15,623 s). The three fastest submissions ran in 16 s—178 times faster than idAb and nearly 1,000 times faster than MegaBLAST. Like idAb and MegaBLAST, algorithms from these top performers have run times that scale linearly with the number of sequences. Thus, these submissions can annotate 10 million sequences in under three minutes and nearly a quarter-billion sequences per hour on a typical desktop machine, demonstrating their potential to scale with constantly improving sequencing technologies.

To investigate the specific technical approaches developed by contestants, we commissioned three independent computer science Ph.D. researchers to review all submissions and determine what techniques were implemented. Their analyses determined that ten distinct elemental methods (**Table 1**) were used in 89 combinations in the 654 submissions. As the number of elemental methods in a submission increased, so did its performance (**Fig. 2** and **Supplementary Methods**), with leaderboard scores increasing by 85.3 points for each additional method employed ($P < 0.01$). Analysis of the benchmark algorithms showed that the methods numbered 2, 3, 5 and 8 were implemented in the MegaBLAST algorithm, and methods 2, 4 and 7 were implemented in the idAb code.

Thus, the results achieved by contest participants in only 14 d improved substantially on the existing solutions available to academic researchers, decreasing processing time by up to three orders of magnitude with accuracy reaching the theoretical maximum. Moreover, 30 different solutions improved upon the state of the art exemplified by both an off-the-shelf, general-purpose tool widely used by the academic community (MegaBLAST) and software developed by a single team addressing the identical problem (idAb), suggesting that prize contests are a robust, reliable approach to efficiently generate desired solutions.

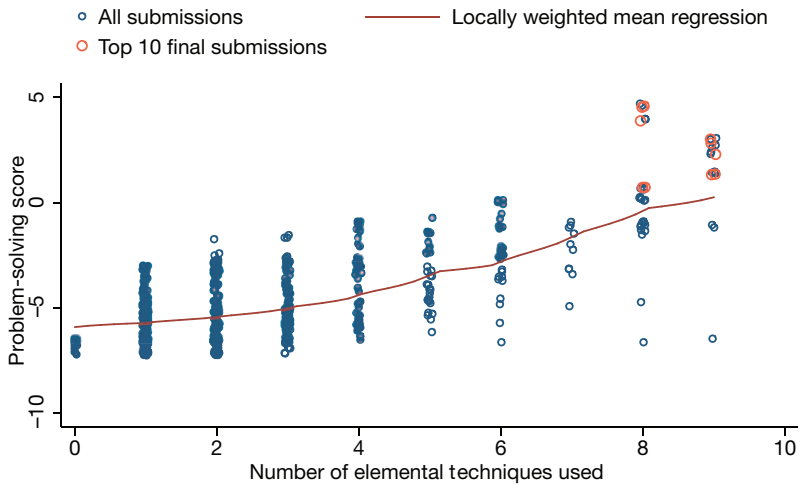

Figure 2 Solution quality plotted against number of techniques. The unit of observation is the individual submission (654 solution submissions by 122 submitters). The problem-solving score is an amalgam of solution quality and time-to-execute used to generate the leaderboard results during the contest (details in **Supplementary Methods**). The curve is a locally weighted polynomial fitted curve. All scores have been monotonically transformed to accentuate small differences among top scores. Positions have been slightly 'jittered' to facilitate viewing of overlapping data points.

Although the solvers were virtually devoid of domain-specific knowledge, abstracting the problem into general algorithmic and mathematical terms allowed a wide range of nondomain experts to address an important, complex problem. These contestants brought to the problem whatever skills and expertise they had or could find, probably yielding a far more diverse toolkit than would be available locally, and generated substantial diversity in technical approaches. Accessing such diversity may be particularly important, as big-data biomedical analytics is a rapidly evolving field in which it is difficult to know a priori the kind, quality and breadth of expertise needed to produce an effective solution.

In summary, we show that a prize-based contest on a commercial platform can effectively recruit skilled individuals to apply their knowledge to a big-data biomedical problem. Deconstruction and transformation of problems for a heterogeneous solver community coupled with adequate data to produce and validate results can support solution diversity and minimize the risk of suboptimal solutions that may arise from limited searches. In addition to the benefits of generating new knowledge, this strategy may be particularly useful in situations where the computational or algorithmic problem, or potentially any science problem, represents a barrier to rapid progress but where finding the solution is not itself the major thrust of the investigator's scientific effort. The America Competes Act passed by the US Congress provides funding agencies with the authority to administer their own prize-based contests and paves the way for establishing how grant recipients might access commercial prize platforms to accelerate their own research.

Note: Supplementary information is available at http://www.nature.com/doifinder/10.1038/nbt.2495.

ACKNOWLEDGMENTS

Harvard Business School's Division of Research and Faculty Development funded the prize money and supported K.R.L., E.L., C.B. and A.M. K.R.L. and K.J.B. were also supported by the NASA Tournament Lab, which is funded by the NASA Human Exploration and Operations Mission Directorate. E.C.G. was supported by the Harvard Clinical and Translational Science Center (NCRR 5UL1RR025758 and 5UL1RR025758-S), and R.A.A. was supported by the Klarman Family Foundation. P.-R.L. was supported by US National Defense Science and Engineeering and US National Science Foundation graduate research fellowships.

AUTHOR CONTRIBUTIONS

K.R.L., K.J.B. and E.C.G. conceived, designed and executed the experiment, analyzed data and wrote the manuscript. C.B. and A.M. contributed to experimental design. E.L. executed the experiment and collected data. R.A.A. identified and codeveloped the immunogenomics problem, tested the submissions and helped write the manuscript. M.L. and L.B. codeveloped the problem statement and test data. P.-R.L. analyzed and categorized all submission data and helped write the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details are available at http://www.nature.com/doifinder/10.1038/nbt.2495.

Karim R Lakhani[1,2,*], Kevin J Boudreau[2,3,*], Po-Ru Loh[4], Lars Backstrom[5], Carliss Baldwin[1], Eric Lonstein[1], Mike Lydon[5], Alan MacCormack[1], Ramy A Arnaout[6,7,*] & Eva C Guinan[7,8,*]

*[1]Harvard Business School, Boston, Massachusetts, USA. [2]Harvard-NASA Tournament Lab, Institute for Quantitative Social Science. [3]London Business School, London, UK. [4]Department of Mathematics and Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. [5]TopCoder.com, Glastonbury, Connecticut, USA. [6]Department of Pathology and Division of Clinical Informatics, Department of Medicine, Beth Israel Deaconess Medical Center, Boston, Massachusetts, USA. [7]Harvard Medical School, Boston, Massachusetts, USA. [8]Department of Radiation Oncology, Dana-Farber Cancer Institute, Boston, Massachusetts, USA. *These authors contributed equally.*
e-mail: eva_guinan@dfci.harvard.edu

1. Zerbino, D.R. *et al. Science* **336**, 179–182 (2012).
2. Tan, T.W. *et al. BMC Genomics* **10** (suppl 3), S36 (2009).
3. Wollman, R. & Stuurman, N. *J. Cell Sci.* **120**, 3715–3722 (2007).
4. Joyce, A. & Palsson, B. *Nat. Rev. Mol. Cell Biol.* **7**, 198–210 (2006).
5. Arnaout, R. *et al. PLoS ONE* **6**, e22365 (2011).
6. Brown, B., Chui, M. & Manyika, J. *McKinsey Q.* **4**, 24–35 (2011).
7. Taylor, C.R. *Am. Econ. Rev.* **85**, 872–890 (1995).
8. Scotchmer, S. *Innovation and Incentives* (MIT Press, Cambridge, MA; 2004).
9. Levine, D.K. *Science* **323**, 1296–1297 (2009).
10. Terwiesch, C. & Xu, Y. *Manage. Sci.* **54**, 1529–1543 (2008).
11. Travis, J. *Science* **319**, 1750–1752 (2008).
12. Howe, J. *Crowdsourcing* (Crown Books, New York; 2008).
13. Eiben, C.B. *et al. Nat. Biotechnol.* **30**, 190–192 (2012).
14. Khatib, F. *et al. Proc. Natl. Acad. Sci. USA* **108**, 18949–18953 (2011).
15. Cooper, S. *et al. Nature* **466**, 756–760 (2010).
16. Khatib, F. *et al. Nat. Struct. Mol. Biol.* **18**, 1175–1177 (2011).
17. Scudellari, M. *Nat. Med.* **18**, 326 (2012).
18. Jeppesen, L. & Lakhani, K.R. *Organ. Sci.* **21**, 1016–1033 (2010).
19. Boudreau, K.J., Lacetera, N. & Lakhani, K.R. *Manage. Sci.* **57**, 843–863 (2011).
20. Marbach, D. *et al. Nat. Methods* **9**, 796–804 (2012).
21. Boyd, S.D. *et al. Sci. Transl. Med.* **1**, 12ra23 (2009).
22. Weinstein, J.A. *et al. Science* **324**, 807–810 (2009).
23. Robins, H.S. *et al. Sci. Transl. Med.* **2**, 47ra64 (2010).
24. Jung, D. *et al. Annu. Rev. Immunol.* **24**, 541–570 (2006).
25. Altschul, S.F. *et al. J. Mol. Biol.* **215**, 403–410 (1990).
26. Brochet, X. *et al. Nucleic Acids Res.* **36**, W503–W508 (2008).
27. Hong, L. & Page, S. *Proc. Natl. Acad. Sci. USA* **101**, 16385–16389 (2004).

Commercialized transgenic traits, maize productivity and yield risk

To the Editor:

Maize expressing different versions of *Bacillus thuringiensis* toxins (*Bt*), 5 enolphyruvylshikimate-3 phosphate synthase (EPSPS) and phosphinothricin acetyl transferase alone or in combination are part of the current wave of agricultural technological change. We analyzed grain yield data from annual field experiments during 1990–2010 in Wisconsin to test hypotheses that maize expressing these transgenic traits alone or in combination (stacked) has greater productivity (as measured by the mean harvested yield) and lower production risk (as measured by the variance, skewness and kurtosis of harvested yield). Compared with conventional hybrids, the impact of transgenic traits (both single and stacked traits) on mean yield ranges from −12.2 to +6.5 bushels per acre. This shows that reducing yield risk is an important source of benefits of transgenic technology, especially for stacked traits. These benefits are estimated to be equivalent to a yield increase of 0.8–4.2 bushels per acre. We found evidence for gene interactions ('yield drag' and 'event lag' effects) that can reduce yield.

The past century has seen marked increases in maize productivity. Average US maize yields increased from 72 to 153 bushels per acre from 1970 to 2010 (ref. 1). Genetic selection has contributed to advances in maize productivity in recent decades[2,3]. Over the past 15 years, productivity gains have been complemented by rapid adoption of transgenic hybrids in the United States (and elsewhere)[2,3]. Rapid adoption of transgenic maize by US farmers suggests that the technology benefits them. Yet documenting the nature and sources of these benefits has been challenging[4,5]. There is some evidence of delayed yield increase due to 'yield lag' and yield drag associated with transgenes[5]. Agricultural production is also subject to substantial risk from unpredictable weather and pest damage. Transgenic crops have been argued to help reduce agricultural production risk, thus motivating insurance companies to