DDBJ Data Analysis Challenge: a machine learning competition to predict *Arabidopsis* chromatin feature annotations from DNA sequences

Eli Kaminuma^{1*}, Yukino Baba², Masahiro Mochizuki³, Hirotaka Matsumoto⁴,
Haruka Ozaki⁴, Toshitsugu Okayama⁵, Takuya Kato⁶, Shinya Oki⁷,
Takatomo Fujisawa¹, Yasukazu Nakamura¹, Masanori Arita¹,
Osamu Ogasawara¹, Hisashi Kashima² and Toshihisa Takagi^{1,8}

¹Center for Information Biology, National Institute of Genetics, 1111 Yata, Mishima, Shizuoka 411-8540, Japan

²Graduate School of Informatics, Kyoto University, 36-1 Yoshida-Honmachi, Sakyo, Kyoto 606-8501, Japan

³IMSBIO Co., Ltd., 4-21-1, Higashi-ikebukuro, Toshima, Tokyo 170-0013, Japan

⁴Laboratory for Bioinformatics Research, RIKEN Center for Biosystems Dynamics Research,
2-1 Hirosawa, Wako, Saitama 351-0198, Japan

⁵BITS Co., Ltd., 1-5-5 Surugadai, Kanda, Chiyoda, Tokyo 101-0062, Japan

⁶Graduate School of Information Science and Technology, The University of Tokyo,
7-3-1 Hongo, Bunkyo, Tokyo 113-8656, Japan

⁷Graduate School of Medical Sciences, Kyushu University, 3-1-1 Maidashi,
Higashi, Fukuoka 812-8582, Japan

⁸Graduate School of Science, The University of Tokyo, 7-3-1 Hongo,
Bunkyo, Tokyo 113-0033, Japan

(Received 22 July 2019, accepted 1 November 2019; J-STAGE Advance published date: 26 March 2020)

Recently, the prospect of applying machine learning tools for automating the process of annotation analysis of large-scale sequences from next-generation sequencers has raised the interest of researchers. However, finding research collaborators with knowledge of machine learning techniques is difficult for many experimental life scientists. One solution to this problem is to utilise the power of crowdsourcing. In this report, we describe how we investigated the potential of crowdsourced modelling for a life science task by conducting a machine learning competition, the DNA Data Bank of Japan (DDBJ) Data Analysis Challenge. In the challenge, participants predicted chromatin feature annotations from DNA sequences with competing models. The challenge engaged 38 participants, with a cumulative total of 360 model submissions. The performance of the top model resulted in an area under the curve (AUC) score of 0.95. Over the course of the competition, the overall performance of the submitted models improved by an AUC score of 0.30 from the first submitted model. Furthermore, the 1st- and 2nd-ranking models utilised external data such as genomic location and gene annotation information with specific domain knowledge. The effect of incorporating this domain knowledge led to improvements of approximately 5%–9%, as measured by the AUC scores. This report suggests that machine learning competitions will lead to the development of highly accurate machine learning models for use by experimental scientists unfamiliar with the complexities of data science.

Key words: chromatin features prediction, deep learning, machine learning competition, sequence read archive

INTRODUCTION

The next-generation sequencers that appeared in

2005 have rapidly been developed to reduce the costs of analyses and realise the goal of the 1000 dollars human genome (DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program [<http://www.genome.gov/sequencingcostsdata>]) (van Nimwegen et al., 2016). As these large-scale sequences and sequence annotations are accumulated, experimental researchers increasingly

Edited by Toshihiko Shiroishi

* Corresponding author. E-mail: ekaminum@nig.ac.jp

DOI: <http://doi.org/10.1266/ggs.19-00034>

seek to construct machine learning models. However, many experimental researchers are not familiar with the analytical techniques of data science, and they often lack the connections to find data scientists with whom to collaborate. One solution to this problem is the crowdsourcing approach, i.e., outsourcing these tasks as challenges to the crowd (Meyer et al., 2011; Saez-Rodriguez et al., 2016; Wazny, 2017). By holding contest-style crowdsourcing events in the form of ‘machine learning competitions’, experimental researchers can leverage the power of crowd workers (Bender, 2016). In the domain of biomedical research, many crowdsourcing competition events have been held, such as the CASP project for protein structure prediction (Moult, 2005), the Bio-CreAtIvE project for text mining methods in molecular biology (Hirschman et al., 2005), the CAPRI project for protein-protein docking (Janin, 2005) and the DREAM project for general tasks in systems biology and medicine (Prill et al., 2011). There are also crowdsourcing platform services specifically for data science. These include Kaggle (<https://www.kaggle.com>), which conducts machine learning competitions such as the biomedical task ‘American Epilepsy Society Seizure Prediction Challenge’, which attracted 504 teams with 654 competitors (Brinkmann et al., 2016). However, one issue with these crowdsourcing competition projects is that they typically only report their results on websites and are only occasionally documented in academic journals with an information science background. While guidelines for running computational competitions (Friedberg et al., 2015) already exist, they provide only an outline of competition rules. Thus, it is difficult for experimental scientists to obtain concrete procedures as competition protocols and to develop further knowledge regarding the organisation of competitions and participants. Meanwhile, there is an education-based data science platform, the UniversityOfBigData (Baba et al., 2018), which is open to the public and features leaderboard panels with participant nicknames, model performance scores and submission status. The platform was constructed for educational purposes and participants are able to visually monitor submission activity and improvements in model performance.

In this report, we conducted a machine learning competition, the DDBJ Data Analysis Challenge, to investigate the quantitative evaluation of participant contributions and the improvement of submitted models for the purpose of supporting experimental scientists unfamiliar with data science. The DDBJ (Kodama et al., 2018) maintains large-scale DNA sequence databases such as the Sequence Read Archive (SRA) (Kodama et al., 2012) for members of the International Nucleotide Sequence Database Collaboration (Cochrane et al., 2016). The machine learning competition was conducted using the UniversityOfBigData platform and took place during the

summer of 2016. The task of the challenge was to predict chromatin feature annotations from DNA sequences while referring to the data protocol of a secondary annotation database of DDBJ sequences. We can obtain chromatin feature annotations as peak-call regions based on ChIP-seq/DNase-seq. Thus, the task in the challenge was designed by utilising chromatin feature annotations of *Arabidopsis thaliana*. This report focuses on the competition protocol, with our desired outcome being that experimental researchers will be able to hold competitions themselves in the future, from coordinating tasks to managing participant submissions, by following the procedures for competition implementation that we describe here.

OVERVIEW OF THE DDBJ CHALLENGE COMPETITION

Overview of the DDBJ Data Analysis Challenge competition The DDBJ Data Analysis Challenge (DDBJ Challenge) was a machine learning competition using the International Nucleotide Sequence Database, a large biomolecular database that exists as an international collaboration between DDBJ in Japan, EMBL/EBI in Europe and NCBI in the U.S.A. The competition was held during the 2016 Japanese school summer period, from July 6 to August 31, for the purpose of quantitatively evaluating the effects of crowdsourcing a data science project. We opened a website dedicated to the DDBJ Challenge (<https://www.ddbj.nig.ac.jp/activities/ddbj-challenge-e.html>) within the framework of the larger DDBJ website, and supported the competition with the high-performance computational resources of the NIG supercomputer.

In the competition, participants submitted prediction outputs using their machine learning models for a prediction task, automatically annotating DNA sequences into chromatin features. Two common types of machine learning competition tasks are prediction problems and insight problems (Kohavi et al., 2000). Solutions to prediction problems were previously known, which meant they could be utilised for model evaluation. However, solutions to insight problems were not known before the competition and thus the evaluation cost for these problems tended to be high. Therefore, we selected prediction problems for this competition.

A total of 38 participants took part in the ‘Predicting chromatin features from DNA sequences’ challenge task, with a cumulative total of 360 model submissions. Three top prize winners and a student prize winner were selected, with the student prize winner (top rank of all student participants) corresponding to fifth place overall. Out of the top five-ranking participants, only one (in fourth place) had an information science background. The other participants had bioinformat-

ics backgrounds, such as company workers and research scientists. In the following sections, we will present the detailed results from the DDBJ Challenge.

Data analysis challenge task with biological features The task involved in the challenge was to predict chromatin annotations from DNA fragment sequences of *A. thaliana* based on the protocol of the ChIP-Atlas database (Oki et al., 2018). The original ChIP-Atlas database contains peak-call annotations from the next-generation sequencing of five model organisms (human, mouse, fruit fly, nematode and budding yeast). The analytical conditions were ‘ChIP-seq’ and ‘DNase-Hypersensitivity’ from LIBRARY_STRATEGY and ‘illumina’ from PLATFORM, where both parameters were metadata attributes of the SRA database. As for human datasets, the SRA database included datasets from big analytical projects such as the ENCODE project (The ENCODE Project Consortium, 2012) and the Roadmap Epigenomics project (Bernstein et al., 2010).

We generated 163 new *A. thaliana* datasets for the competition task by analysing the SRA database under the ChIP-Atlas protocol. The versions of the analytical tools are summarised in Supplementary Table S1. From these, datasets with only a few peaks were discarded, and then eight datasets from two classes of Antigen and Cell Type were selected based on an informative number of annotated peaks (Table 1). Antigen classes can be specified by peak types of DNase-seq, ChIP-seq to characterise binding sites of transcription factors, ChIP-seq for identifying histone modification sites and others. The eight class combination labels were masked to prevent cheating by participants, and we documented these to prohibit any reannotation of the SRA *Arabidopsis* database.

The following training datasets for machine learning models were generated from the TAIR10 database of the *Arabidopsis* genome and annotations (<https://abrc.osu.edu/>).

- Input training data: 60,000 DNA fragment sequences
- Input test data: 10,000 DNA fragment sequences
- Output training data: 60,000 rows $\times$ 8 conditions (True/False boolean labels)

A true prediction against test data of 10,000 rows with 8 conditions was requested for submission to the UniversityOfBigData website.

Ethical approval and consent to participate The DDBJ Challenge crowdsourcing research was approved by the Institutional Review Board (IRB) (Graber and Graber, 2013) of the National Institute of Genetics, Japan in March 2016. Informed consent with IRB approval was displayed to the crowd participants in the DDBJ Challenge and user agreement was obtained prior to users being registered on the UniversityOfBigData data science infrastructure of Kyoto University.

Concealing life science domain-specific descriptions in the competition task Predicted labels for training data comprise boolean code with one binary digit. True (1) or False (0) indicates that the DNA sequence contains or does not contain the characteristic chromatin region, respectively. This boolean description is popular in the domain of data science, while the description of DNA sequences was unfamiliar to the participants with no life science background. Thus, we converted nucleotide sequences with nucleotide letters requiring specific domain knowledge to zero-one-hot encoding vectors using binary digits to make them easier to handle for participants without domain-specific knowledge. Therefore, sequence fragments comprising 200 nucleotide letter codes were converted to 800 binary codes.

Participant recruitment and submission management of prediction outputs Participants were recruited through DDBJ newsletters and various email mailing lists from bioinformatics communities in

Table 1. List of eight selected conditions from the *Arabidopsis* SRA dataset. The SRA accession SRX# denotes the *EXPERIMENT* condition

	SRA		ChIP-Atlas		
	SRA experiment accession	LIBRARY_STRATEGY	Antigen class	Antigen	Cell type
1	SRX391997	DNase-seq	DNase-seq	–	flower
2	SRX391993	DNase-seq	DNase-seq	–	seed
3	SRX022326	ChIP-Seq	Histone	H3K9ac	rosette leaf
4	SRX346160	ChIP-Seq	Histone	H3K4me3	leaf
5	SRX145429	ChIP-Seq	Histone	H3K9me2	leaf
6	SRX156079	ChIP-Seq	RNA polymerase	RNA polymerase V	seedling
7	SRX128177	ChIP-Seq	TF	PRR5	whole plant
8	SRX159029	ChIP-Seq	TF	MYC	seedling

TF: Transcription factor.

Japan. The prediction results of chromatin feature annotations from the test DNA sequences used as input data were submitted using Kyoto University’s educational data science platform, UniversityOfBigData. The competition dataset was distributed from two data sites, (1) the web server of the UniversityOfBigData, and (2) disk space on the NIG Supercomputer (Ogasawara et al., 2013). After submission, half of the predicted output data were utilised for computing tentative participant ranks based on the evaluation scores of the receiver operating curve–area under the curve (ROC–AUC). The remaining half of the dataset was retained to calculate the final scores (Supplementary Table S2).

Supporting GPU computational resources and tool installation requests To construct machine learning models, participants need to prepare their computational resources. The DDBJ Challenge committee supported 16 GPU nodes, part of the computing resources of the NIG supercomputer, where the special usage of GPU nodes expired during the final day of the competition. Moreover, user requests for open source software installation were accepted. Thus, data science tools such as BVLC Caffe (Jia et al., 2014) and PFN Chainer (Tokui et al., 2015) were installed on the NIG supercomputer. From Mathworks Japan, MATLAB licences were offered for the student competition during the DDBJ Challenge period.

DDBJ CHALLENGE AWARDS AND PARTICIPATORY FEATURES

Outline of the first-place model The first-place model in the DDBJ Challenge was constructed by Dr. Masahiro Mochizuki. He designed a compositional architecture model based on two classifiers of extremely randomized trees (ERT) (Geurts et al., 2006) and convolutional neural networks (CNNs) (LeCun et al., 1998). A stacking generalisation algorithm (Wolpert, 1992) was then applied for ensemble learning to integrate the two classifier models. He also utilised two external parameters, genomic coordination and gene annotation information, in addition to the default features of the challenge query sequences. The first ERT model was constructed as an $N \times M$ matrix, with N as the index of input query sequences and M as the chromosome number. The genomic coordinate position was computed by alignment analysis of query sequences against the TAIR10 *Arabidopsis* reference genome sequence. We will refer to this genomic coordinate-based ERT model as the genomic coordinates based model (GCBM). The second component of the model, the CNN, incorporated query DNA sequences and gene structural annotation information, where gene annotation data were downloaded as a general feature format (Eilbeck et al., 2005) file from the TAIR10 database. As shown in Fig. 1A, gene structural

annotation was modelled by feature matrix representation with two parameters as follows:

$(1-r)^d$ if the base is included in a gene of the strand,

Otherwise 0.

The variable r is the attenuation rate and the variable d is the distance of a genomic coordinate from the gene start coordinate. For instance, if the variable r is zero, the feature is equal to 1, which indicates that the target coordinate position is included in an annotated gene coordinate. Likewise, when the variable r is greater than zero, the gradient value from the gene start coordinate is computed. This CNN model based on gene annotation is called the gene annotated sequences based model (GASBM), indicating the effect of domain knowledge. Figure 1B depicts a schematic explanation of the CNN from feature matrix input to chromatin feature outputs. Figure 1C displays the benchmark results of the algorithms utilised in the first-place model. The performance of the individual GCBM and GASBM models was evaluated using their AUC evaluation scores. The query sequence-based model without gene annotation data was compared as a control.

Model performances and algorithms of the top three ranking models After the close of the DDBJ Challenge submission period, the top three overall winners and the student prize winner were asked to submit reports explaining the detailed structures of their proposed machine learning models. Table 2 summarises the performance and model algorithm designs based on these reports, including information on individual analytical tools. All of the top three winners used CNNs, and all deployed ensemble learning as a strategy to incorporate multiple algorithms. External parameters were incorporated by the first- and second-place winners. The first-place winner used two external parameters, genomic position and gene annotation with strand information, as stated above, while the second-place winner used only genomic position. The performance of the top model resulted in an AUC score of 0.95, where the performance of the first submitted model was 0.65. The value of the first submitted model was equivalent to the performance score of our tutorial with the linear discriminant analysis algorithm. Over the course of the competition, the overall performance of the submitted models improved by an AUC score of 0.30 from the first submitted model.

Incorporating domain knowledge into prediction models and the participants’ background knowledge of life science In general, domain knowledge is known to significantly raise performance scores in domain-specific machine learning tasks. For example, a study aimed at predicting the risk of breast cancer

Fig. 1. Overview of the first-place model in the DDBJ Data Analysis Challenge. (A) A schematic view of feature matrices with both strands representing a gene annotated sequence of the GASBM model. (B) A schematic view of the neural network structures of the GASBM model. (C) A benchmark result of the GASBM, GCBM and ensemble algorithms in the first-place model.

Table 2. Model performances, designs and tools of the top prize winners

Rank	Model performance (AUC)	Model design	Programming tool
1st	0.946	- 2 classifiers (CNN, ERT) - Ensemble learning (Stacking) - External data (Genomic position, Gene structural annotation)	Python 3.5 Scikit-learn 0.17.1 Chainer 1.10.0
2nd	0.899	- 2 classifiers (CNN, Product of genomic distance decay parameter and nearest training data output) - Ensemble learning (Averaged) - External data (Genomic position)	Julia 0.4.6 Python 2.7.10 TensorFlow 0.8.0
3rd	0.854	- 7 classifiers (CNN, Logistic regression, Gradient boosting, Naive Bayes for multivariate Bernoulli models, Random forest, Extremely randomized trees, Extreme gradient boosting) - Ensemble learning (Stacking)	Python 2.7.11 NumPy 1.10.4 Scikit-learn 0.17 Chainer 1.11.0 XGBoost 0.4a30

found that the effect of domain knowledge of SNPs was a 6%–8% increase in accuracy (Bochare et al., 2014). Both the first- and second-place models incorporated external parameters as domain knowledge. From Table 2, the differences in the AUC scores between the models with

domain knowledge (first and second place) and without domain knowledge (third place) were 0.092 and 0.045, respectively. The rough effect of domain knowledge, 5%–9%, is close to that found by Bochare et al. in the above-mentioned study. Furthermore, assessing the

effect of domain knowledge using only the first-place model under the average of 8 conditions, we can confirm three differences in AUC scores as 0.034, -0.034 and 0.091 between the sequence-based model (Fig. 1C(a)) and the three respective models incorporating patterns, GASBM (Fig. 1C(c)), GCBM (Fig. 1C(d)) and the stacking model (Fig. 1C(e)), respectively. Regarding the participants' background knowledge, the top three winners had background knowledge of life science, while the fourth-ranking participant had a background in information science. In the UniversityOfBigData platform, the DDBJ Challenge was the first life science domain task, which meant that the previous seven tasks, such as 'Business Card OCR', did not require life science knowledge. This domain knowledge effect might imply some barriers for participants from non-life science backgrounds when it comes to tackling life science machine learning competitions.

Programming tools of top-ranking models As shown in Table 2, all the winning models used the Python programming language, with its plentiful machine learning libraries, including scikit-learn (Pedregosa et al., 2011), and the deep learning libraries of Google's TensorFlow (Abadi et al., 2015) and PFN's Chainer. The student prize winner also utilised Python 2.7 and the Lasagne package (Dieleman et al., 2015) as a wrapper library for the University of Montreal's Theano (The Theano Development Team, 2016).

Analysis of participatory metrics 1: Participation inequality Participatory metrics analysis may be valuable for readers who want to plan machine learning competitions. We counted submission statistics using functions of the UniversityOfBigData platform. Figure 2A displays the pattern of submissions by participants, with the horizontal axis showing the days of the competition period and the vertical axis showing the intermediate prediction score of the competitions. In the field of crowdsourcing research, there is typically inequality in the number of submissions between the participants (Ortega et al., 2008), in that a small number of participants tend to account for a large number of submissions, as a fraction of the whole submission volume (Sauermann and Franzoni, 2015). Such participation inequality can be quantified by the Gini coefficient (Yang et al., 2016), which is often used in economics. The Gini coefficient, G , is defined as a ratio of the areas on the Lorenz curve diagram. If the area between the line of perfect equality and the Lorenz curve is denoted as A , and the area under the Lorenz curve is denoted as B , then the Gini coefficient can be calculated as $G = A/(A + B)$. The Gini coefficient is a global index, where 0 represents complete equality and 100 complete inequality. The Gini coefficient of submissions in our competition was 0.48, while those of other competition tasks in the UniversityOfBigData platform ranged from 0.43 to 0.54.

Fig. 2. References for the participatory metrics analyses. (A) The pattern of submissions by participants, with the horizontal axis showing the days of the competition period and the vertical axis showing the intermediate prediction score of the competition. (B) Pearson's correlation coefficient between the final performance scores (vertical axis) and the number of model submissions (horizontal axis) on the participants' leaderboard activity.

Analysis of participatory metrics 2: Correlation between leaderboard activity and performance

Pearson's correlation coefficient between the final performance scores and the number of submissions, as the participants' leaderboard activity, was 0.35 ($P = 0.03$), which was a weak positive correlation, as shown in Fig. 2B. In crowdsourcing research, it has been reported that engagement score and submission number are not correlated based on annotation tasks (Good et al., 2015). As for crowdsourced machine learning modelling tasks, an online article discussing Kaggle competitions states that there is an association between making a greater total number of submissions and an increase in final ranking (<https://rpubs.com/pedmiston/kaggle>). However, another study (Küffner et al., 2015) reported unclear results, with both positive and negative correlations between the number of submissions and final performance score depending on the datasets used. Concerning the correlation analysis of performance score with the number of submissions, it may be necessary to collect data on further tasks.

CONCLUDING REMARKS

We described the DDBJ Data Analysis Challenge, a machine learning competition in the life science domain held during the 2016 summer period. This report will provide reference knowledge for ensuing challengers and task planners, and encourage future machine learning competitions in life science domains.

We thank all of the participants in the competition: orion, Ryota, tsukasa, emn, extraterrestrial Fuun species, ηzw, mkoido, ksh, AoYu@Tohoku, hiro, emihat, MorikawaH, hmt-yamamoto, tonets, suzudora, take2, bicycle1885, morizo, forester, doiyasan, yudai, tag, nwatarai, soki, himkt, saoki, tsunechan, Ken, A.K, singular0316, IK, yk_tani, yota0000. We are also grateful to Ayako Oka, Yasuhiro Tanizawa, Takako Mochizuki, Fumi Hayashi, Naoko Sakamoto and Tarzo Ohta for their support in preparing the datasets for the task, and to Fumitaka Ootobe, Takuya Ohtani, Hikari Amano, Takafumi Ohbiraki, Yuji Ashizawa, Tomohiko Yasuda, Naofumi Ishikawa, Tomohiro Hirai, Tomoka Watanabe, Chiharu Kawagoe, Emi Yokoyama, Kimiko Suzuki and Junko Kohira for their computational infrastructure and management support. Data analysis was partially performed using the Research Organization of Information and Systems (ROIS) NIG Supercomputer System. This research was partially supported by management expenses grants from the DNA Data Bank of Japan, the ROIS, and JST CREST Grant Number JPMJCR1501, Japan.

REFERENCES

- Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., et al. (2015) TensorFlow: Large-scale machine learning on heterogeneous distributed systems. Software available from tensorflow.org. arXiv:1603.04467.
- Baba, Y., Takase, T., Atarashi, K., Oyama, S., and Kashima, H. (2018) Data analysis competition platform for educational purposes: lessons learned and future challenges. Proceedings of the 8th Symposium on Educational Advances in Artificial Intelligence 2018 (EAAI-18), 7887–7892.
- Bender, E. (2016) Challenges: crowdsourced solutions. *Nature* **533**, S62–S64.
- Bernstein, B. E., Stamatoyannopoulos, J. A., Costello, J. F., Ren, B., Milosavljevic, A., Meissner, A., Kellis, M., Marra, M. A., Beaudet, A. L., Ecker, J. R., et al. (2010) The NIH roadmap epigenomics mapping consortium. *Nat. Biotechnol.* **28**, 1045–1048.
- Bochare, A., Gangopadhyay, A., Yesha, Y., Joshi, A., Yesha, Y., Brady, M., Grasso, M. A., and Rishe, N. (2014) Integrating domain knowledge in supervised machine learning to assess the risk of breast cancer. *Int. J. Med. Eng. Inform.* **6**, 87–99.
- Brinkmann, B. H., Wagenaar, J., Abbot, D., Adkins, P., Bosshard, S. C., Chen, M., Tieng, Q. M., He, J., Muñoz-Almaraz, F. J., Botella-Rocamora, P., et al. (2016) Crowdsourcing reproducible seizure forecasting in human and canine epilepsy. *Brain* **139**, 1713–1722.
- Cochrane, G., Karsch-Mizrachi, I., Takagi, T., and International Nucleotide Sequence Database Collaboration (2016) The international nucleotide sequence database collaboration. *Nucleic Acids Res.* **44**, D48–D50.
- Dieleman, S., Schlüter, J., Raffel, C., Olson, E., Sønderby, S. K., Nouri, D., Maturana, D., Thoma, M., Battenberg, E., Kelly, J., et al. (2015) Lasagne: First release. doi:10.5281/zenodo.27878.
- Eilbeck, K., Lewis, S. E., Mungall, C. J., Yandell, M., Stein, L., Durbin, R., and Ashburner, M. (2005) The Sequence Ontology: a tool for the unification of genome annotations. *Genome Biol.* **6**, R44.
- Friedberg, I., Wass, M. N., Mooney, S. D., and Radivojac, P. (2015) Ten simple rules for a community computational challenge. *PLoS Comput. Biol.* **11**, e1004150.
- Geurts, P., Ernst, D., and Wehenkel, L. (2006) Extremely randomized trees. *Mach. Learn.* **63**, 3–42.
- Good, B. M., Nanis, M., Wu, C., and Su, A. I. (2015) Microtask crowdsourcing for disease mention annotation in PubMed abstracts. *Pac. Symp. Biocomput.* 282–293.
- Graber, M. A., and Graber, A. (2013) Internet-based crowdsourcing and research ethics: the case for IRB review. *J. Med. Ethics* **39**, 115–118.
- Hirschman, L., Yeh, A., Blaschke, C., and Valencia, A. (2005) Overview of BioCreAtIvE: critical assessment of information extraction for biology. *BMC Bioinformatics* **6**, S1.
- Janin, J. (2005) Assessing predictions of protein-protein interaction: the CAPRI experiment. *Protein Sci.* **14**, 278–283.
- Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., and Darrell, T. (2014) Caffe: Convolutional architecture for fast feature embedding. arXiv:1408.5093.
- Kodama, Y., Mashima, J., Kosuge, T., Kaminuma, E., Ogasawara, O., Okubo, K., Nakamura, Y., and Takagi, T. (2018) DNA Data Bank of Japan: 30th anniversary. *Nucleic Acids Res.* **46**, D30–D35.
- Kodama, Y., Shumway, M., and Leinonen, R. (2012) The sequence read archive: explosive growth of sequencing data. *Nucleic Acids Res.* **40**, D54–D56.
- Kohavi, R., Brodley, C. E., Frasca, B., Mason, L., and Zheng, Z. (2000) Kdd-cup 2000 organizers' report: peeling the onion. *ACM SIGKDD Explor. Newsl.* **2**, 86–92.
- Küffner, R., Zach, N., Norel, R., Hawe, J., Schoenfeld, D., Wang, L., Li, G., Fang, L., Mackey, L., Hardiman, O., et al. (2015)

- Crowdsourced analysis of clinical trial data to predict amyotrophic lateral sclerosis progression. *Nat. Biotechnol.* **33**, 51–57.
- LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998) Gradient-based learning applied to document recognition. *Proceedings of the IEEE* **86**, 2278–2324.
- Meyer, P., Alexopoulos, L. G., Bonk, T., Califano, A., Cho, C. R., de la Fuente, A., de Graaf, D., Hartemink, A. J., Hoeng, J., Ivanov, N. V., et al. (2011) Verification of systems biology research in the age of collaborative competition. *Nat. Biotechnol.* **29**, 811–815.
- Moult, J. (2005) A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction. *Curr. Opin. Struct. Biol.* **15**, 285–289.
- Ogasawara, O., Mashima, J., Kodama, Y., Kaminuma, E., Nakamura, Y., Okubo, K., and Takagi, T. (2013) DDBJ new system and service refactoring. *Nucleic Acids Res.* **41**, D25–D29.
- Oki, S., Ohta, T., Shioi, G., Hatanaka, H., Ogasawara, O., Okuda, Y., Kawaji, H., Nakaki, R., Sese, J., and Meno, C. (2018) ChIP-Atlas: a data-mining suite powered by full integration of public ChIP-seq data. *EMBO Rep.* **19**, e46255.
- Ortega, F., Gonzalez-Barahona, J. M., and Robles, G. (2008) On the inequality of contributions to Wikipedia. *Proceedings of the 41st Annual Hawaii International Conference on System Sciences (HICSS 2008)*, 304.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011) Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830.
- Prill, R. J., Saez-Rodriguez, J., Alexopoulos, L. G., Sorger, P. K., and Stolovitzky, G. (2011) Crowdsourcing network inference: the DREAM predictive signaling network challenge. *Sci. Signal.* **4**, mr7.
- Saez-Rodriguez, J., Costello, J. C., Friend, S. H., Kellen, M. R., Mangravite, L., Meyer, P., Norman, T., and Stolovitzky, G. (2016) Crowdsourcing biomedical research: leveraging communities as innovation engines. *Nat. Rev. Genet.* **17**, 470–486.
- Sauermann, H., and Franzoni, C. (2015) Crowd science user contribution patterns and their implications. *Proc. Natl. Acad. Sci. USA* **112**, 679–684.
- The ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74.
- The Theano Development Team (2016) Theano: A Python framework for fast computation of mathematical expressions. *arXiv:1605.02688*.
- Tokui, S., Oono, K., Hido, S., and Clayton, J. (2015) Chainer: a next-generation open source framework for deep learning. *Proceedings of Workshop on Machine Learning Systems in the 29th Annual Conference on Neural Information Processing Systems (NIPS)* **5**, 1–6.
- van Nimwegen, K. J., van Soest, R. A., Veltman, J. A., Nelen M. R., van der Wilt, G. J., Vissers L. E. L. M., and Grutters, J. P. C. (2016) Is the \$1000 genome as near as we think? A cost analysis of next-generation sequencing. *Clin. Chem.* **62**, 1458–1464.
- Wazny, K. (2017) “Crowdsourcing” ten years in: a review. *J. Glob. Health* **7**, 020602.
- Wolpert, D. H. (1992) Stacked generalization. *Neural Netw.* **5**, 241–259.
- Yang, A., Fan, H., Jing, N., Sun, Y., and Zipf, A. (2016) Temporal analysis on contribution inequality in openstreetmap: a comparative study for four countries. *Int. J. Geo-inf.* **5**, 5.