Big data science at AMED-BINDS

Haruki Nakamura¹

Received: 14 January 2020 / Accepted: 23 January 2020 / Published online: 6 February 2020

© International Union for Pure and Applied Biophysics (IUPAB) and Springer-Verlag GmbH Germany, part of Springer Nature 2020

Science is considered to rest upon four basic pillars. Traditionally, experiment and theory are considered as the first and second of these. Advances in computing power have driven simulation to become the third pillar. The development of large repositories of data—so called Big Data—along with the ability to analyze these data libraries using inferential and interpolative statistical methods, has led big data analysis, in other words, data science, to now constitute the fourth scientific pillar.

The reason why big data analysis has become possible in such fields as genome science and astronomy is due to rapid developments in information technology (IT) that have allowed the observation and storage of very large data sets. Analysis of such huge and precise data sets, including image data, has made it possible to obtain new knowledge, which has never before been available. The famous “DIKW pyramid” by Rowley (2007) indicates the basic concept for such big data-driven science: The lowest level is raw big data (**D**ata), which are then validated and assembled to information. Then, analysis of **I**nformation can provide **K**nowledge, which finally produces **W**isdom. “Wisdom” in this context should be defined as the ability to solve scientific problems by making reliable predictions. Making science-based predictions can be achieved using either data-driven (empirical) or theory-driven chains of argument. Wisdom is frequently sought in the field of life science, where lots of phenomena are observed without necessarily producing any theoretical principles. With its ability to establish predictive confidence based on interpolative assessment of big data sets, such big data-driven approaches are particularly useful in life science.

Let us consider what is necessary for data science, once we define “wisdom” as above. The answer should be very

simple: “Any problem lacking a firm theoretical underpinning cannot be solved without data”, or more correctly, “Any such problem cannot be solved without *correct* data”. Such a statement is easy to formulate but it is very difficult in practice to confirm whether the data used by data science are actually correct or not. There are two approaches towards this issue based on (i) post-validation of data sets and (ii) pre-validation of data sets using standardized data production methods.

(i) Post-validation of data sets: The first approach is to make the data well validated. Namely, any observed or simulated data should be objectively scrutinized and described by users with enough metadata, so that the observation or simulation can be reproducible. Several examples from the international databank for protein structures, wwPDB (worldwide Protein Data Bank) (wwPDB consortium 2019), have addressed this issue of data validation (Gore et al. 2017; Young et al. 2018). In addition, the data should be updated with the versioning system, so that the user can utilize the newest data with a contemporaneous understanding of the history of data production and revision.

(ii) Pre-validation of data sets using standardized data production methods: The second approach is to create new data by experiments or simulations such that the produced data is automatically archived with enough metadata to allow it to be validated and reproduced, keeping data quality high. If data production is made using a standardized high-throughput (HTP) procedure, new and correct big data are created. Such approaches are becoming popular in many fields in science, and governmental scientific funding agencies, along with other funding bodies, are increasingly requesting that fundees make their data open to society during or following the publication process.

Creating innovative drugs requires state-of-the-art protein science technology. Such technological approaches utilize the big data associated with the genome, proteome, medical and clinical data, and protein structural data deposited in the PDB. In addition to utilizing big data, drug discovery research itself produces big data. The BINDS

✉ Haruki Nakamura
harukin@protein.osaka-u.ac.jp

¹ Institute for Protein Research, Osaka University, 3-2 Yamadaoka, Suita, Osaka 565-0871, Japan

(Basis for supporting INnovative Drug discovery and life Science research) program started in April of 2017 as one of the AMED (Japan Agency for Medical Research and Development) programs to promote drug discovery in Academia at the pre-clinical and life science research stages. AMED is a rather new funding agency in Japan since 2015 for medical and life science research, integrating the funding from the Ministry of Health, Labor and Welfare; Ministry of Education, Culture, Sports, Science and Technology; and Ministry of Economy, Trade and Industry. A characteristic feature of the BINDS program is that it is composed of 59 research groups in diverse fields which include pharmaceutical science, medicine, chemistry, genomics, structural biology, informatics, and computer science. In addition, BINDS members support researchers outside of the program through the sharing of key technologies such as, synchrotron beams, free-electron lasers, cryo-EM (electron microscopy) instrumentation, NMR devices, supercomputer resources, chemical libraries equipped with HTP assay systems, and next generation DNA sequencers.

In this letter, I will not directly address drug discovery results generated within the BINDS program which will instead be described in other papers in the near future. Instead, here I focus on the scientific activities of the BINDS program from the view point of data science. Directed to this point, several BINDS researchers have engaged in structural bioinformatics studies in order to predict and analyze protein complex structures. Kentaro Tomii at AIST has continuously developed original algorithms to predict protein tertiary structures and their complex structures (Shiota et al. 2015) and has achieved considerable success at the blind worldwide contest, CASP (Nakamura et al. 2017). Most recently, his group has developed a new algorithm with deep neural networks for protein contact prediction, which makes it possible to build tertiary structural models (Fukuda & Tomii 2020). Their core technology is based on multiple alignments of protein sequence big data. Hidetoshi Kono at QST has built many complex structural models using molecular simulation with constraints provided by experimental data of SAXS (small angle X-ray scattering) and cryo-EM, integrating local structures provided by the PDB structural database. In particular, their simulations could sample lots of possible conformations having low free energies with the advantage of their generalized sampling approach (Kono et al. 2018). These days, cryo-EM shows its strong power to reveal the near atomic structures of large complexes of proteins and nucleic acids. However, there are still technical difficulties, of which attempts are being made to overcome them using data science approaches. In order to prepare good cryo-EM grids, which is a key technology for facilitating data collection, Keiichi

Namba group at Osaka University has developed software programs, Gwatch and Rwatch, to select appropriate particle images by automatically averaging millions of two-dimensional (2D) images. Toru Terada at University Tokyo and Kazutoshi Tani at Mie University have developed a deep-learning-based method to identify good regions of a cryo-EM grid without the aid of specialist knowledge and only using several hundreds of 2D images. Takeshi Kawabata at Osaka University has developed his own software program, gmfit, for fitting subunits of proteins into density map of protein complexes using a Gaussian mixture model (GMM) (Kawabata 2008; Kawabata 2018a). His approach should be useful for model building not only with low-resolution cryo-EM data (Kawabata 2018b) but also with other data of atomic force microscopy (AFM) (Dasgupta et al. 2020).

Almost all the members in the BINDS program produce lots of different kind of data through their own scientific activities. The members involved in the “Platform function optimization unit” have made efforts to construct archives for those data. One recent archive is called “Antibody Square, <http://antibodysq.info/>,” which has been developed by Yukinari Kato at Tohoku University and Hirofumi Suzuki at Waseda University, as the repository for antibody developers, users, and suppliers (Fig. 1). Genji Kurisu at PDBj, Osaka University, manages the mirror site of the raw image database of cryo-EM (<https://empiar.pdbj.org/>), EMPIAR at EMBL-EBI, for supporting data-in and data-out applications involving huge raw image data sets, which have been produced by the BINDS members. In the near future, the EMPIAR-PDBj site will directly accept the depositions of the raw image data. Gert-Jan Bekker (2020) at Osaka University has developed an archive of the static and dynamic structural models, BSM-Arc (biological structure model archive, <https://bsma.pdbj.org/>), which were produced during the BINDS activities for modeling and molecular simulations (Fig. 2). Several other data archives, such as chemical compound library for screening and gene expression and epigenetics given by next generation sequencer (NGS), are also planned and will be released in the near future.

In conclusion, BINDS members have advanced their activities utilizing big data by the data science approach. In addition, the BINDS program has been producing many excellent results both from the individual research contributions of BINDS members and from the support and assistance provided to other researchers, who submit their proposals through the BINDS platform. Many BINDS results are themselves now archived as big data in addition to their publication within original scientific articles. Such a dual track system of publication and data deposition in big data repositories helps to spread scientific knowledge

Antibody Square - Information repository for antibody developers, users, and suppliers -

+ Search

- Data

Showing 1 - 10 of 95 items

Show/hide all

- anti-PATag, "NZ-1", rat

Antigen	Type	Peptide tag
	Tag name	PATag
	Sequence	GVAMPGAEDDVV (12 res.)
	Trademark	Kato-lab (5671624)
Antibody	Antibody name	NZ-1
	Antibody animal	rat
	Isotype	IgG2a, lambda
Description		
Application	<ul style="list-style-type: none"> protein purification, immunoprecipitation, immunohistochemistry, western blotting, flow cytometry, etc. For details of protein purification, see Fujii et al., PEP2014; Fujii et al., JCS2016 Has been used for WB in several articles 	
	article 1 , literature . <ul style="list-style-type: none"> Biochem. Biophys. Res. Commun., 2006, 349(4), 1301-1307 Inhibition of tumor cell-induced platelet aggregation using a novel anti-podoplanin antibody reacting with its platelet-aggregation-stimulating domain. Yukinari Kato, Mika Kato Kaneko, Atsushi Kuno, Noboru Uchiyama ... FEBS Lett., 2007, 581(2), 331-336 Functional glycosylation of human podoplanin: glycan structure of platelet aggregation-inducing factor. Mika Kato Kaneko, Yukinari Kato, Akihiko Kameyama, Hiromi Ito ... 	
Reference	Show more 27 items	
Contact	Kato laboratory, Tohoku univ.	
Supplier	FUJIFILM Wako Pure Chemical	

Menu

- This page
 - Show/hide all
 - Exclusive
 - Search
 - Data
 - About
- External links
 - Development note
 - Antibodybank
- Options
 - [English] 日本語
 - Font size

Fig. 1 The web page of “Antibody Square,” the repository for antibody developers, users, and suppliers. <http://antibodysq.info/>. Accessed 15 January 2020

BSM-Arc English 日本語

Login using ORCID Help Browse

Help menu

- Help top page
- How to submit a new entry
- How to add text panels
- File upload system
- How to annotate files/folders
- Adding authors
- Adding PDB IDs
- How to set a graphical abstract
- How to request publication of an entry

Welcome to BSMA

The Biological Structure Model Archive (BSMA or BSM-Arc for short) top page.

We welcome depositions of raw data and supplementary information corresponding to published, peer-reviewed papers.

If you have any questions, please [contact us](#).

Please note that this archive is still under development and thus we appreciate any feedback you might have.

2018-10-25 (last update: 3 months ago)

Latest entries

Mutual population-shift driven antibody-peptide binding elucidated by molecular dynamics simulations
Gert-Jan Bekker, Ikuro Fukuda, Junichi Higo, Narutoshi Kamiya
Deposition date: 2019-10-02
Modification date: 2020-01-29
Release date: 2020-01-29

A novel rare variant R292H in RTN4R affects growth cone formation and possibly contributes to schizophrenia susceptibility
Hiroki Kimura, Yuki Fujita, Takeshi Kawabata, Kanako Ishizuka, Chenyao Wang, Yoshimi Iwayama, Yoko Okahisa, Itaru Kushima, Mako Morikawa, Yota Uno, Takashi Okada, Masashi Ikeda, Toshiya Inada, Aleksic Branko, Daisuke Mori, Takeo Yoshikawa, Nakao Iwata, Haruki Nakamura, Toshihide Yamashita, Norio Ozaki
Deposition date: 2019-12-20
Modification date: 2019-12-20
Release date: 2019-12-20

Rare genetic variants in CX3CR1 and their contribution to the increased risk of schizophrenia and autism spectrum disorders
Kanako Ishizuka, Yuki Fujita, Takeshi Kawabata, Hiroki Kimura, Yoshimi Iwayama, Toshiya Inada, Yoko Okahisa, Jun Egawa, Masahide Usami, Itaru Kushima, Yota Uno, Takashi Okada, Masashi Ikeda, Aleksic Branko, Daisuke Mori, Takeo Yoshikawa, Nakao Iwata, Haruki Nakamura, Toshihide Yamashita, Norio Ozaki
Deposition date: 2019-12-20
Modification date: 2019-12-20
Release date: 2019-12-20

Structural and thermodynamic characterization of endo-1,3-β-glucanase: Insights into the substrate recognition mechanism
Masayuki Oda, Satomi Inaba, Narutoshi Kamiya, Gert-Jan Bekker, Bunzo Mikami
Deposition date: 2019-12-12
Modification date: 2019-12-12
Release date: 2019-12-12

Dynamic Docking of a Medium-Sized Molecule to Its Receptor by Multicanonical MD Simulations
Gert-Jan Bekker, Mitsugu Araki, Kanji Oshima, Yasushi Okuno, Narutoshi Kamiya
Deposition date: 2019-07-09
Modification date: 2019-07-09
Release date: 2019-07-09

Fig. 2 The web page of “BSM-Arc (biological structure model archive),” the repository for the static and dynamic structural models. <https://bsma.pdbj.org/> (Bekker et al. 2020). Accessed 15 January 2020

throughout the world, further contributing to drug discovery and life science.

References

- Bekker G, Kawabata T, Kurisu G (2020) The biological structure model archive: an archive for in-silico models and simulations. *Biophys Rev*. (in the current issue)
- Dasgupta B, Miyashita O, Tama F (2020) Reconstruction of low-resolution molecular structures from simulated atomic force microscopy images. *BBA* 1864:129420. <https://doi.org/10.1016/j.bbagen.2019.129420>
- Fukuda H, Tomii K (2020) DeepECA: an end-to-end learning framework for protein contact prediction from a multiple sequence alignment. *BMC Bioinformatics* 21:10. <https://doi.org/10.1186/s12859-019-3190-x>
- Gore S, García ES, Hendrickx PMS, Gutmanas Westbrook JD, Yang H, Feng Z, Baskaran K, Berrisford JM, Conroy M, Hudson BP, Ikegawa Y, Kobayashi N, Lawson CL, Mading S, Mak L, Mukhopadhyay A, Oldfield TJ, Patwardhan A, Peisach E, Sahni G, Sekharan MR, Sen S, Shao C, Smart OS, Ulrich EL, Yamashita R, Quesada M, Young JY, Nakamura H, Markley JL, Berman HM, Burley SK, Velankar S, Kleywegt GJ (2017) Validation of the structures in the Protein Data Bank. *Structure* 25:1916–1927. <https://doi.org/10.1016/j.str.2017.10.009>
- Kawabata T (2008) Multiple subunit fitting into a low-resolution density map of a macromolecular complex using a Gaussian mixture model. *Biophys J* 95:4643–4658. <https://doi.org/10.1529/biophysj.108.137125>
- Kawabata T (2018a) Gaussian-input Gaussian mixture model for representing density maps and atomic models. *J Struct Biol* 203:1–16. <https://doi.org/10.1016/j.jsb.2018.03.002>
- Kawabata T (2018b) Rigid-body fitting of atomic models on 3D density maps of electron microscopy. *Adv Exp Med Biol* 1105:219–235. https://doi.org/10.1007/978-981-13-2200-6_14
- Kono H, Sakuraba S, Ishida H (2018) Free energy profiles for unwrapping the outer superhelical turn of nucleosomal DNA. *PLoS Comput Biol* 14:e1006024. <https://doi.org/10.1371/journal.pcbi.1006024>
- Nakamura T, Oda T, Fukasawa Y, Tomii K (2017) Template-based quaternary structure prediction of proteins using enhanced profile–profile alignments. *Proteins* 86:274–282. <https://doi.org/10.1002/prot.25432>
- Rowley J (2007) The wisdom hierarchy: representations of the DIKW hierarchy. *J Inform Sci* 33:163–180. <https://doi.org/10.1177/0165551506070706>
- Shiota T, Imai K, Qiu J, Hewitt VL, Tan K, Shen H-H, Sakiyama N, Fukasawa Y, Hayat S, Kamiya M, Elofsson A, Tomii K, Horton P, Wiedemann N, Pfanner N, Lithgow T, Endo T (2015) Molecular architecture of the active mitochondrial protein gate. *Science* 349:1544–1548. <https://doi.org/10.1126/science.aac6428>
- wwPDB consortium (2019) Protein Data Bank: the single global archive for 3D macromolecular structure data. *Nucl Acids Res* 47:D520–D528. <https://doi.org/10.1093/nar/gky949>
- Young JY, Westbrook JD, Feng Z, Peisach E, Persikova I, Sala R, Sen S, Berrisford JM, Swaminathan GJ, Oldfield TJ, Gutmanas A, Igarashi R, Armstrong DR, Baskaran K, Chen L, Chen M, Clark AR, Costanzo LD, Dimitropoulos D, Gao G, Ghosh S, Gore S, Guranovic V, Hendrickx PMS, Hudson BP, Ikegawa Y, Kengaku Y, Lawson CL, Liang Y, Mak L, Mukhopadhyay A, Narayanan B, Nishiyama K, Patwardhan A, Sahni G, Sanz-Garcia E, Sato J, Sekharan MR, Shao C, Smart OS, Tan L, van Ginkel G, Yang H, Zhuravleva MA, Markley JL, Nakamura H, Kurisu G, Kleywegt GJ, Velankar S, Berman HM, Burley SK (2018) Worldwide Protein Data Bank biocuration supporting open access to high-quality 3D structural biology data. *Database* 2018:1–17. <https://doi.org/10.1093/database/bay002>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.