Overview of the big data bioinformatics symposium (2SCA) at BSJ2019

Tsuyoshi Shirai¹ • Tohru Terada²

Received: 10 January 2020 / Accepted: 6 February 2020 / Published online: 14 February 2020

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

The third wave of AI (artificial intelligence) has been boosted by the developments of high-performance computing technologies and big data, both of which were absent in the first and second AI waves. The same would be also true for the studies in bioinformatics or biological data science. DNA sequence data from next-generation sequencing represent such a recent explosive increase of biological data, which has enlarged the quantitative gap between sequence and structural data. However, the advances of structure/interaction analysis techniques, as represented by cryo-EM (electron microscopy), are steadily narrowing the gap by accumulating the structural data of biological supramolecules. Consequently, this trend demands concurrent improvements in bioinformatics methods, and it prompted us to plan a symposium to review recent progress in structural bioinformatics in the BSJ2019 (57th Annual Meeting of the Biophysical Society of Japan).

The symposium session titled “Challenges of bioinformatics for the era of molecular structure big-data” was started on the morning of September 25, 2019 at the Seagaia Convention Center. The session was opened by the keynote speech “Big Data Science at AMED-BINDS” from Dr. Haruki Nakamura (Japan Agency for Medical Research and Development). The objective of the BINDS (Basis for Supporting Innovative Drug Discovery and Life Science Research) program is to establish an innovative platform to accelerate the therapeutic applications of early-stage drug discovery and medical technology advances by providing and sharing key technological infrastructures and technical/scientific supports from expert researchers in respective fields. In his talk, the mission of

BINDS was introduced with a special emphasis on the notion that every basic life science research would be eventually related to medicine.

The next talk, “Prediction of protein residue contacts and protein-ligand interactions with deep neural networks”, was from Dr. Kentaro Tomii (National Institute of Advanced Industrial Science and Technology (AIST)). He introduced a novel machine learning method to predict protein–ligand interactions. In this method, a GNN (graph neural network) for ligand molecules and a CNN (convolutional neural network) for protein sequences were combined, and the results showed the presented method out-performed existing methods (Tsubaki et al. 2019). He also introduced a recent application of the method for residue–residue contact predictions based on MSA (multiple sequence alignment) (Fukuda and Tomii 2020). Such efforts, including a neural network used in analyzing protein dynamics (Tsuchiya et al. 2019), are reported in this issue (Tsuchiya and Tomii 2020).

Dr. Hidetoshi Kono (National Institutes for Quantum and Radiological Science and Technology) introduced the recent study on structural modeling of overlapping dinucleosome in his talk, “Integrated approach of experimental data and computer modeling and simulation for understanding chromatin structure and dynamics.” The crystal structure of overlapping dinucleosome revealed considerable deviations from SANS/SAXS (small angle neutron/X-ray scattering) data especially at high-resolution areas. He and his colleagues have generated a structural library of the overlapping dinucleosome through molecular modeling and simulation including histone tail regions, which were invisible by crystal structure analysis. They successfully find the structures which are well fitting to the SANS/SAXS data from the library and suggest possible conformational change and dynamics in solution (Matsumoto et al. 2020). A recent modeling study of HP1-bound nucleosome was also described in his talk (Kumar and Kono 2020).

Dr. Takeshi Kawabata (Osaka University) first introduced the databases EMPIAR-PDBj and BSM-Arc in his talk, “EM informatics: archiving raw 2D images and fitting atomic

✉ Tsuyoshi Shirai
t_shirai@nagahama-i-bio.ac.jp

¹ Department of Computer Bioscience, Nagahama Institute of Bio-Science and Technology, 1266 Tamura, Nagahama, Shiga 526-0829, Japan

² Interfaculty Initiative in Information Studies, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan

models into a map.” The EMPIAR-PDBj is a mirror site of an EM data repository, EMPIAR, in Japan, and BSM-Arc is the database for the structural data from modeling/simulation studies, for which appropriate public repository has not been so far offered (Bekker et al. 2020). He then introduced an improved application of gmfit, which fit structural models into a density map with a GMM (Gaussian mixture model) method. It was pointed out that an increment of the number of Gaussian functions from ~ 10 for low-resolution map to more than 500 for high/atomic resolution map was required for efficient fitting.

In his talk, “Development of a deep-learning-based method to identify ‘good’ regions of a cryo-EM grid,” Dr. Tohru Terada (The University of Tokyo) presented a new machine learning method to evaluate the quality of ice layers in the holes on cryo-EM grids. The test data revealed a considerable variability among different protein samples, and the parameters trained on one sample generally did not show high performance on another sample. According to this result, he proposed to use a few hundred low magnitude images from the target grid for a training set, and demonstrated high-accuracy discrimination (Yokoyama et al. 2020).

Finally, we heard from Dr. Keiichi Namba (Riken SPring-8 and Osaka University) about “Improvements in grid preparation method and software for facilitating cryoEM data collection.” Recently, he successfully analyzed the structure of apoferritin to a 1.53-Å resolution (EMDB ID: 6840) by using a recently installed electron microscope (Kato et al. 2019). This is the highest resolution analysis of proteins via EM at this point of time. He pointed out, despite the excellent improvements in EM, evaluation of grid quality was still a major bottleneck for high-throughput analyses. In order to overcome this problem, he developed a pipeline application named Gwatch, which evaluated the grid qualities based on the class averages obtained from each grid.

Wrapping up the session, the presented studies in this symposium emphasized that the collaboration between

experimental and theoretical structural biology researchers is necessary and quite effective. Bioinformatics, especially machine learning methods tailored for molecular structural data, will serve as a key component of the current integrated structural analysis in its application to medicine/drug discovery researches.

References

- Bekker G, Kawabata T, Kurisu G (2020) Biological structure model archive : an archive of *in silico* models and simulations of biological molecules. *Biophys Rev*. <https://doi.org/10.1007/s12551-020-00632-5>
- 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
- Kato T, Makino F, Nakane T, Terahara N, Kaneko T, Shimizu S, Motoki S, Ishikawa I, Yonekura K, Namba K (2019) CryoTEM with a cold field emission gun that moves structural biology into a new stage. *Microsc Microanal* 25(Suppl. S2):998–999
- Kumar A, Kono H (2020) Interactions of heterochromatin protein 1 (HP1) structural elements. *Biophys Rev* 12(2) Current Issue
- Matsumoto A, Sugiyama M, Zhenhai L, Martel A, Porcar L, Inoue R, Kato D, Osakabe A, Kurumizaka H, Kono H (2020) Structural studies of overlapping dinucleosomes in solution. *Biophys J*. <https://doi.org/10.1016/j.bpj.2019.12.010>
- Tsubaki M, Tomii K, Sese J (2019) Compound-protein interaction prediction with end-to-end learning of neural networks for graphs and sequences. *Bioinformatics* 35(2):309–318
- Tsuchiya Y, Tomii K (2020) Neural networks for protein structure prediction and dynamics analysis. *Biophys Rev* 12(2) Current Issue
- Tsuchiya Y, Taneishi K, Yonezawa Y (2019) Autoencoder-based detection of dynamic allostery triggered by ligand binding based on molecular dynamics. *J Chem Inf Model* 59(9):4043–4051
- Yokoyama Y, Terada T, Shimizu K, Nishikawa K, Kozai D, Shimada A, Mizoguchi A, Fujiyoshi Y, Tani K (2020) Development of a deep-learning-based method to identify “good” regions of a cryo-electron microscopy grid. *Biophys Rev* 12(2) Current Issue

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