10月27日（水）

Zoom ブレイクアウトルーム

口頭発表３・４

座長：夏目 やよい（医薬基盤・健康・栄養研究所）、川又 生吹（東北大学）、
池田 和由（理化学研究所/慶應義塾大学）

O3-01 Hideaki Mamada (Japan Tobacco Inc. / Meiji Pharmaceutical University)
"Novel QSAR approach for clearance prediction, combination DeepSnap-Deep Learning, and conventional machine learning"

O3-02 Satoko Namba (Department of Bioscience and Bioinformatics, Faculty of Computer Science and Systems Engineering, Kyushu Institute of Technology)
"From drug repositioning to target repositioning: omics-based prediction of therapeutic targets for a variety of diseases"

O3-03 Zhaonan Zou (Dept Drug Discov Med, Grad Sch Med, Kyoto Univ)
"Transcription factor binding profiling using chemically induced genes by ChIPEA"

O3-04 Tomokazu Shibata (Department of Bioscience and Bioinformatics, Kyushu Institute of Technology)
"Food digital transformation: large-scale prediction of food functions and elucidation of the mode-of-action"

O3-05 Xiaoran Hu (School of Computing, Tokyo Institute of Technology Molecular Robot Research Institute, Co., Ltd.)
"Construction of super-resolution DNA AFM images with VR DNA molecular models"

O3-06 Taisei Mori (Tohoku University)
"Simulating Self-replication of Linear Structures"

O3-07 Chen Ma (School of Computing, Tokyo Institute of Technology)
"Tracking microtubule groups with deep learning and optical flow"

O3-1

Novel QSAR approach for clearance prediction, combination DeepSnap-Deep Learning, and conventional machine learning

Hideaki Mamada[1, 2]
hideaki.mamada@jt.com

Yukihiro Nomura[1]
yukihiro.nomura@jt.com

Yoshihiro Uesawa[2]
uesawa@my-pharm.ac.jp

[1] Drug Metabolism and Pharmacokinetics Research Laboratories, Central Pharmaceutical Research Institute, Japan Tobacco Inc., 1-1, Murasaki-cho, Takatsuki, Osaka 569-1125, Japan

[2] Department of Medical Molecular Informatics, Meiji Pharmaceutical University, 2-522-1, Noshio, Kiyose-shi, Tokyo 204-858, Japan

Keywords: Clearance (CL), Quantitative structure-activity relationships (QSAR), DeepSnap-Deep Learning (DeepSnap-DL), ensemble model, consensus model

In drug discovery, there are some prediction targets for which the prediction accuracy by machine learning is not sufficient. Therefore, the development of new prediction models is required. In this study, rat clearance (CL) was selected as a challenging target because of poor prediction [1], and a new prediction model was developed. A classification model was constructed using 1545 in-house compounds for which rat CL data are available. The molecular descriptors calculated by Molecular Operating Environment (MOE), alvaDesc, and ADMET Predictor software were used to construct the prediction model. Molecular descriptors and random forest selected by DataRobot were used for conventional machine learning. The area under the curve (AUC) and accuracy (ACC) were 0.883 and 0.825, respectively. Conversely, compound images and Deep Learning were used for DeepSnap and Deep Learning (DeepSnap-DL) [2]. AUC and ACC were 0.905 and 0.832, respectively. The two models (conventional machine learning and DeepSnap-DL) were combined to develop a novel prediction model. The ensemble model using mean of the predicted probabilities from each model improved the evaluation scores (AUC=0.943 and ACC=0.874). Furthermore, using the results of the agreement between each classification as a consensus model resulted in higher ACC (=0.959). These combination models with a high level of predictive performance can be applied to rat CL as well as other pharmacokinetic parameters. These models will help the design of more rational compounds in drug discovery.

[1] McIntyre, T. A.; Han, C.; Davis, C. B. Prediction of animal clearance using naïve Bayesian classification and extended connectivity fingerprints, *Xenobiotica*, **2009,** *39*, 487–494. https://doi.org/10.1080/00498250902926906.

[2] Uesawa, Y. Quantitative structure–activity relationship analysis using deep learning based on a novel molecular image input technique, *Bioorganic & Medicinal Chemistry Letters,* **2018,** *28*, 3400–3403. https://doi.org/10.1016/j.bmcl.2018.08.032.

O3-2

From drug repositioning to target repositioning: omics-based prediction of therapeutic targets for a variety of diseases

Satoko Namba[1]
namba.satoko775@mail.kyutech.jp

Michio Iwata[1]
iwata121@bio.kyutech.ac.jp

Yoshihiro Yamanishi[1]
yamani@bio.kyutech.ac.jp

[1] Department of Bioscience and Bioinformatics, Faculty of Computer Science and Systems Engineering, Kyushu Institute of Technology, 680-4 Kawazu, Iizuka, Fukuoka 820-8502, Japan

Keywords: Therapeutic target, Drug discovery, Transcriptome, Gene knock-down, Gene over-expression

The identification of therapeutic targets, biomolecules that lead to therapeutic effects, for treating diseases is vital in drug development [1]. However, most therapeutic targets easily identified using pathological data have been thoroughly investigated. The conventional methods for investigating individual diseases are limited in their ability to discover novel therapeutic targets. Recently, there has been an accumulation of omics data on various diseases. Thus, there is a need to identify novel therapeutic targets by effectively using omics data resources about various diseases.

In this study, we proposed the novel concept of target repositioning, an extension of the concept of drug repositioning, to predict new therapeutic target for a variety of diseases. We developed a novel computational method using genetically perturbed and disease-specific gene expression signatures. We predicted inhibitory and activatory therapeutic targets separately, assuming that gene expression following gene knock-down of inhibitory targets reflects the functions of drugs that inhibit the targets, and gene expression following gene over-expression reflects the functions of drugs that activate the targets. Based on the inverse correlations between the disease-specific and genetically perturbed signatures, we predicted novel therapeutic targets, and performed an integrative analysis taking into account the similarities among the diseases. Our results revealed that the proposed method accurately predicted known inhibitory and activatory targets for diseases. We also made a comprehensive prediction of therapeutic targets for a variety of diseases, suggesting many potential therapeutic targets.

[1] Santos, R. et al. A comprehensive map of molecular drug targets. Nat Rev Drug Discov. 16, 19-34 (2017).

O3-3

Transcription factor binding profiling using chemically induced genes by ChIPEA

Zhaonan Zou[1]
zou.zhaonan.56e@st.kyoto-u.ac.jp

Michio Iwata[2]
iwata121@bio.kyutech.ac.jp

Yoshihiro Yamanishi[2]
yamani@bio.kyutech.ac.jp

Shinya Oki[1]
oki.shinya.3w@kyoto-u.ac.jp

[1] Department of Drug Discovery Medicine, Kyoto University Graduate School of Medicine, 53 Shogoin Kawahara-cho, Sakyo-ku, Kyoto 606-8507, Japan

[2] Department of Bioscience and Bioinformatics, Faculty of Computer Science and Systems Engineering, Kyushu Institute of Technology, 680-4 Kawazu, Iizuka, Fukuoka 820-8502, Japan

Keywords: Drug modes of action, Transcriptome, ChIP-seq, Transcription factor, Epigenetic landscape

Modification of disease-elicited gene expression is one of the core aspects in numerous drugs' modes of action. To predict drug–disease associations, transcriptomics-based approaches with pathway analysis, graph theory and supervised machine learning-based calculation were developed. However, the pharmacological mechanism employed by drugs remain largely unknown.

In this study, we focused on transcription factors (TFs) that integratively regulate differentially expressed genes (DEGs) in response to drug treatment. In particular, TF enrichment analysis by analyzing large-scale ChIP-seq data obtained from ChIP-Atlas database (ChIPEA) was performed for each chemical to identify TFs with enriched binding for chemically perturbed DEGs. Performance evaluation with area under the ROC curve (AUC) suggests the reliability of ChIPEA in drug target discovery (global AUC = 0.66). Furthermore, we successfully identified the pivotal factors that link drugs to diseases or side effects by utilizing protein–disease database (global AUC = 0.68). This approach is with high confidence because it is fully based on actual experiments of given transcriptome data and public ChIP-seq data. In the pharmaceutical field, ChIPEA is useful to shed light on compounds failed to be approved by identifying TFs primarily involved in the modes of action, together with the factors associated with potential side effects. Approved drugs including agents composed of unidentified ingredients such as traditional herbal medicines can also be re-examined for novel targets and actions, thus beneficial to drug repositioning research.

03-4

Food digital transformation: large-scale prediction of food functions and elucidation of the mode-of-action

Tomokazu Shibata[1] **Yusuke Tanaka**[2]
shiba535@bio.kyutech.ac.jp yusuke-tanaka@housefoods.co.jp

Hiromu Taguchi[2] **Ryusuke Sawada**[1] **Morihiro Aoyagi**[2]
h-taguchi@housefoods.co.jp sawad330@bio.kyutech.ac.jp m-aoyagi@housefoods.co.jp

Takashi Hirao[2] **Yoshihiro Yamanishi**[1]
t-hirao@housefoods.co.jp yamani@bio.kyutech.ac.jp

[1] Department of Bioscience and Bioinformatics, Kyushu Institute of Technology,
680-4 Kawazu, Iizuka, Fukuoka 820-8502, Japan
[2] Research and Development Headquarters, House Foods Group Inc.,
1-4 Takanodai, Yotsukaido, Chiba 284-0033, Japan

Keywords: food function, constituent compound, target protein, big data, machine learning

The aging of the population in developed countries, including Japan, has led to the problem of increasing medical costs. It is important to extend healthy life expectancy so that people can lead healthy and cultured lives. Daily food intake is closely related to health, and it would be ideal if we could maintain health through our daily diet. Since foods contain a wide variety of constituent compounds, there is a high possibility that they have unknown food functions and health effects. Even if food functions are known, it is very difficult to understand the mode-of-action.

In this study, we developed a machine learning method to comprehensively predict food functions and the mode-of-action based on a vast amount of food-related data. In contrast to our previous study on the prediction of health effects of food peptides [1], the scope of this study is not limited to peptides but covers all possible constituent compounds of foods. First, we collected information on the chemical structures of 69,594 constituent compounds for 757 foods from literature and databases. Next, using 1,830,624 compound-protein interaction pairs (1,288,343 compounds and 4,643 proteins) as training data, we constructed machine learning models to predict compound-protein interactions, and comprehensively predicted the proteins with which food constituent compounds interact. To estimate food functions, we linked the food constituent compounds to applicable diseases based on known therapeutic targets of 649 diseases. The correspondence between the therapeutic target proteins and the applicable diseases was manually collected from literature data. Finally, we predicted a large-scale network consisting of four types of nodes (foods, constituent compounds, target proteins, and applicable diseases), and elucidated the mode-of-action of the predicted food functions. For each food, the functional associations among target proteins were also examined at the pathway level. The proposed method is expected to be useful not only for prediction of food functions but also for elucidation of the mode-of-action.

[1] Fukunaga, I.; Sawada, R.; Shibata, T.; Kaitoh, K.; Sakai, Y.; Yamanishi, Y. Prediction of the Health Effects of Food Peptides and Elucidation of the Mode-of-action Using Multi-task Graph Convolutional Neural Network. *Molecular informatics*, **2020**, *39*, 1900134.

O3-5

Construction of super-resolution DNA AFM images with VR DNA molecular models

Xiaoran Hu[1,3]
hu.x.ab@m.titech.ac.jp

Masayuki Yamamura[1]
my@cs.titech.ac.jp

Hirotada Kondo[2]
kondo@vraide.jp

Akinori Kuzuya[2,3]
kuzuya@kansai-u.ac.jp

Gutmann Gregory Spence[1,3]
ggutmann13@jcu.edu

Akihiko Konagaya[3]
konagaya@molecular-robot.com

[1] School of Computing, Tokyo Institute of Technology
[2] Dept. Chemistry and Materials Engineering, Kansai University
[3] Molecular Robot Research Institute, Co., Ltd.

Keywords: Molecular Robotics, Super Resolution, Atomic Force Microscope Images, Deep Learning

An Atomic Force Microscope (AFM) is a high-resolution instrument that can detect various materials and samples in the atmosphere and liquid environment. It has become a basic tool in molecular robot research especially for DNA nano-structural design such as DNA origami technology. However, due to the limitation of AFM imaging resolution, it is difficult to observe the details of double-helix DNA structure, such as major groove and minor groove. In order to solve the difficulty to obtain high-resolution DNA pictures directly, this research first tries to establish a VR molecular model approach to obtain super resolution AFM images in atomic levels. Then, we use a virtual AFM probes in various scales to scan the DNA molecular models and to obtain DNA AFM images of different resolutions by simulating the AFM imaging process. Finally, the deep learning method is applied to build a super-resolution network to obtain high-resolution AFM images with different resolution DNA images as training sets.

[1] Gregory Gutmann, Ryuzo Azuma, Akihiko Konagaya: A Virtual Reality Computational Platform Dedicated for the Emergence of Global Dynamics in a Massive Swarm of Objects, *J. of the Imaging Society of Japan*, **2018**, 57(6), 647-653.

03-6

Simulating Self-replication of Linear Structures

Taisei Mori[1]
taisei.mori.t6@dc.tohoku.ac.jp

Ibuki Kawamata[1,2]
ibuki.kawamata@tohoku.ac.jp

Satoshi Murata[1]
satoshi.murata.a4@tohoku.ac.jp

[1] Department of Robotics, School of Engineering, Tohoku University, 6-6-01, AramakiAzaAoba, Aoba, Sendai, Miyagi, 980-0845, Japan
[2] Natural Science Division, Faculty of Core Research, Ochanomizu University, 2-1-1 Ohtsuka, Bunkyo-ku, Tokyo 112-8610, Japan

Keywords: Self-replication, Virtual Spring Model, Transition rule, Autocatalysis

Self-replication is the process by which a system creates ones identical to itself without external operations from outside. A typical example of self-replication is that of living organisms, but its process is extremely complicated involving so many chemical reactions that it is difficult to see what is essential in the process. Simulation models of self-replication help us to find out what are the essential conditions for a system to replicate itself [1-3]. In order to describe the self-replication process, we have proposed the Virtual Spring Model [4], in which the bonds between the elements are regarded as spring-mass-damper systems. In this model, the state transition of each element is used to represent the chemical reactions among them.

By using this model, we were able to represent the self-replication system consisting of up to three interconnected elements. As for the self-replication procedure, the use of catalytic elements increased the success rate of self-replication, but the problem was that the number of transition rules was too large despite the simplicity of the system. In order to achieve scalable self-replication independent of the size of the system, the rules need to be redesigned.

Here, we propose a new self-replication procedure inspired by the concept of "complementarity of DNA", that only complementary base pairs (A and T, C and G) selectively bind to each other to form a double helix [5]. In DNA replication, an enzyme (polymerase) takes advantage of this property to make a complementary copy of the strand. By representing the DNA replication process with the use of autocatalytic procedure in the Virtual Spring Model, a scalable self-replication system for linear structures can be realized. Under this framework, we present a set of transition rules and simulation results of self-replication of a strand (linear structure) formed by interconnected elements with various sequences. We expect that the framework will be also useful to establish self-replication systems of more complex structures and dynamic mechanisms.

[1] S. A. Kauffman, “Cellular Homeostasis, Epigenesis, and Replication in Randomly Aggregated Macromolecular Systems.” *Journal of Cybernetics* 1, pp.71-96, 1971.

[2] K. Tomita, H. Kurokawa, S. Murata, “Graph automata: natural expression of self-reproduction.” *Physica D* 171, pp.197-210, 2002.

[3] Z. Zeravcic, M. P. Brenner, “Self-replicating colloidal clusters.” *PNAS* 111(5), pp.1748-1753, 2014.

[4] K. Fujibayashi, S. Murata, K. Sugawara, M. Yamamura, “Self-Organizing Formation Algorithm for Active Elements.” *IEEE*, pp.416-421, 2002.

[5] J. D. WATSON, F. H. C. CRICK, “Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid.” *Nature* 171, pp.737-738, 1953.

O3-7

Tracking microtubule groups with deep learning and optical flow

Chen Ma[13]
ma.c.ab@m.titech.ac.jp

Masayuki Yamamura[1]
my@cs.titech.ac.jp

Mousumi Akter[2]
mousumi@sci.hokudai.ac.jp

Akira Kakugo[23]
kakugo@sci.hokudai.ac.jp

Gutmann Gregory Spence[13]
ggutmann13@jcu.edu

Akihiko Konagaya[3]
konagaya@molecular-robot.com

[1] School of Computing, Tokyo Institute of Technology.
[2] Faculty of Science, Hokkaido University.
[3] Molecular Robot Research Institute, Co., Ltd.

Keywords: Molecular Robotics, Gliding Assay, Deep Learning, Optical Flow

Microtubules often from groups at high density, which are able to glide in the same directions on Kinesin coated glass surface. At higher density of microtubules, they tend to move together (snuggling) to avoid collision and overriding of microtubules. As a result, microtubule groups emerge motion patterns showing straight, curved or wave like trajectories [1].

This research aims to analyze the condition of phase transition of the motion patterns of microtubules by deep learning which will bring new advancements in the field of molecular robotics. As a first step, we have developed an order parameter analysis workflow which consists of the U-Net like Fully Convolutional Neural Network (FCN) for noise filtering, Sparse Optical Flow (SOF) for tracking and SOF cluster for cluster matching among frames.

In this workflow, at first we trained the parameters using videos generated by the latest version of microtubule gliding assay simulation system [2], and then applied them on real experimental video data for evaluation. The workflow could drastically accelerate its performance by GPU with CUDA parallel programming.

[1] Daisuke Inoue, Greg Gutmann, Takahiro Nitta, Takahiro Nitta, Arif Md. Rashedul Kabir, Akihiko Konagaya, Kiyotaka Tokuraku, Kazuki Sada, Henry Hess, Akira Kakugo: Adaptation of Patterns of Motile Filaments under Dynamic Boundary Conditions, *ACS Nano,* **2019,** *13,* 11, 12452–12460.

[2] Greg Gutmann, Daisuke Inoue, Akira Kakugo, Akihiko Konagaya: Parallel Interaction Detection Algorithms for a Particle-based Live Controlled Real-time Microtubule Gliding Simulation System Accelerated by GPGPU, *Inter. J. of New Generation Computing (NGC)*, **2017**, *35*, 157–180.

O4-1

Transformer-based Generative Adversarial Networks for Generating Molecules with Desired Properties

Chen Li[1]
li260@bio.kyutech.ac.jp

Kazuma Kaitoh [1]
kaito168@bio.kyutech.ac.jp

Yoshihiro Yamanishi [1]
yamani@bio.kyutech.ac.jp

[1] Kyushu Institute of Technology, 680-4 Kawazu, Iizuka, Fukuoka, 820-8502, Japan

Keywords: Transformer, Generative adversarial network, Policy gradient, Molecular generation

Molecules can be represented by string-based sequences derived from molecular graphs, called the simplified molecular-input line-entry system (SMILES). Generative adversarial networks (GAN) with SMILES strings [1] have attracted widespread attention in generating molecules in drug discovery. Most models apply recurrent neural networks (RNNs) as the generator for the molecular generation with SMILES strings. However, RNNs are difficult to generate molecules with complex rings. In general, highly cyclic molecules have long sequence representations and more strict syntax than acyclic molecules. Slight changes in the syntax may result in the generation of molecules with totally different chemical property, or invalid molecules. Furthermore, RNNs cannot work on GPU versions because the current iteration must compute after the previous time step, which is not conducive to handle big data to explore infinite chemical space.

To overcome the above drawbacks, we propose a transformer-based objective-reinforced GAN model in this study. The model consists of two main parts: *generator* and *discriminator*. The generator is a generative model that tries to generate realistic fake data, and it is a transformer architecture with several stacked encoders and decoders. The discriminator is treated as a binary classifier, attempting to distinguish the generated data to avoid being fooled by the generator. The discriminator is based on a convolutional neural network (CNN), which is composed of a convolutional layer, a max-pooling layer, and a highway layer. Note that the generator and discriminator train in alternation. In addition, a reinforcement learning approach called the Monte Carlo policy gradient (MCPG) [2] is applied. While ensuring that the discriminator effectively guides the training of the generator, it also takes the desired chemical properties into account to generate desired molecules. Concretely, the discriminator first outputs the probability that the current input sequence is from the original SMILES dataset. Then, it calculates the chemical properties of the current input sequence, such as drug-likeness and solubility. Finally, the sum of the probability and the properties are used as a reward for MCPG. In experiments, we test our proposed method on molecular generation from the ZINC chemical dataset, and demonstrate the usefulness of our method in terms of uniqueness, novelty, and diversity in generating molecules.

[1] Guimaraes, G. L.; Sanchez-Lengeling, B.; Outeiral, C.; Farias, P. L. C.; Aspuru-Guzik, A. Objective-reinforced generative adversarial networks (ORGAN) for sequence generation models, *arXiv preprint arXiv:*1705.10843, 2017.

[2] Sutton, R.S.; McAllester, D.A.; Singh, S.P.; Mansour, Y. Policy gradient methods for reinforcement learning with function approximation, *Proc. In Advances in Neural Information Processing Systems 12*, 1057-1063, 1999.

04-2

RetroSynthWAVE: An Open-Source Software Platform for Efficient Chemical Synthesis Research

Haris Hasic [1, 2]
hasic@cb.cs.titech.ac.jp

Takashi Ishida [1]
ishida@c.titech.ac.jp

[1] Department of Computer Science, School of Computing, Tokyo Institute of Technology, W8-85, 2-12-1 Ookayama, Meguro-ku, Tokyo 152-8550, Japan
[2] Elix Inc., Daini Togo Park Building 3F, 8-34 Yonbancho, Chiyoda-ku, Tokyo 102-0081, Japan

Keywords: Machine Learning – AI Method Development, Data Curation, Data Visualization

The RetroSynthWAVE project aims to establish a systematic, open-source software platform focused on the field of chemical synthesis. It enables quick and efficient research for beginners as well as advanced users by providing software packages that cover the following synthesis-related functionalities: **W**idgets and helpers, **A**ggregated chemical compound and chemical reaction data, **V**ariety of popular existing model implementations, and **E**valuation metrics. Each of the software packages can be used independently, as well as within the project software stack.

RetroSynthWAVE: HANA is the first and fundamental software package of the project. The name is derived as an acronym for **H**elpers **AN**d **A**ccessories, and it represents a utility wrapper module that encapsulates all of the essential libraries (e.g., RDKit [1], RDChiral [2]) and offers additional fundamental functionalities while being easy to use.

RetroSynthWAVE: COCORO is the second software package of the project which is developed using the functionalities from the previous one. The name is derived as an acronym for **CO**llection of Chemical **CO**mpound and **R**eacti**O**n Data, and it represents an easy-to-use data platform that automates the retrieval, cleaning and featurization of available chemical information datasets. (e.g., ChEMBL [3], USPTO [4])

RetroSynthWAVE: CO-OP is the third software package of the project which is developed using the functionalities from the previous two. The name is derived as an acronym for **CO**llection **O**f **P**opular Models, and it represents an easy-to-use model platform for the reimplementation of popular existing (retro)synthesis models using the PyTorch library. It enables other users to submit the implementation of new models in a standardized fashion.

RetroSynthWAVE: REFEREE is the final software package of the project. The name is derived as an acronym for **R**etroactive **EF**ficiency M**E**t**R**ics **E**valuation Fram**E**work, and it represents a utility module that enables the definition of user-defined, pre-existing, and novel evaluation metrics for (retro)synthesis-focused models. Furthermore, by using the functionalities from all of the previous software packages, it enables the retroactive application of new metrics on pre-existing models thus enabling more advanced benchmarking of all relevant models.

[1] RDKit: Open-Source Cheminformatics Software. https://www.rdkit.org/. (2021.08.13)
[2] Coley, C.W.; Green, W.H.; Jensen, K.F.; RDChiral: An RDKit Wrapper for Handling Stereochemistry in Retrosynthetic Template Extraction and Application, *Journal of Chemical Information and Modeling*, 2019, 59, 6, 2529-2537.
[3] ChEMBL Database - EMBL-EBI: https://www.ebi.ac.uk/chembl/. (2021.08.13)
[4] Lowe, D. M. Extraction of Chemical Structures and Reactions from the Literature. *Ph.D. Thesis*, University of Cambridge, Department of Chemistry, Pembroke College, 2012.

04-3

Graph Convolutional Networks for Ligand-based Virtual Screening against the Androgen Receptor

Romeo Cozac[1] romeo.cozac@elix-inc.com
Nazim Medzhidov[1] nazim.medzhidov@elix-inc.com
Casey Galvin[1] casey.galvin@elix-inc.com
Shinya Yuki[1] shinya.yuki@elix-in.com

[1] Elix Inc., Daini Togo Park Building 3F, 8-34 Yonbancho, Chiyoda-ku, Tokyo 102-0081 Japan

Keywords: Androgen Receptor, Machine learning, Graph convolutional Networks, drug repurposing

Androgen receptor (AR) is a ligand-dependent transcription factor that belongs to the family of steroid hormone nuclear receptors. Androgens bind to the ligand binding domain of the AR with strong affinity and are capable of regulating transcription of AR-regulated genes. AR signaling has implications in pancreatic cancer as well as tumors in the lungs, kidney, liver, and bladder. The standard treatment approach for patients with prostate cancer is to lower testosterone levels in the body, however, this does not always prove effective since some patients do not respond to this form of treatment. Therefore alternative treatment options are necessary. Small molecule antagonists that interfere with androgens binding to AR have been under active investigation. In this study we utilize a graph based Machine Learning model to identify small molecule AR antagonists. In our approach, we design a flexible architecture that supports different graph convolutional layers. We used Bayesian optimization to find the best-performing graph kernel and hyperparameters, and we applied MC Dropout to measure the variance and confidence of the predicted values. The trained model was used to screen three datasets of commercially available compounds: the ZINC dataset, the eMolecules dataset, and a list of FDA-approved drugs. Using a high confidence threshold for the predicted activity and confidence, we reduced the original set of 364K molecules to 55 hits which were not present in our training set. Following a patent search on the 55 hits, we noticed that 36% had at least one relevant patent describing high activity against the Androgen Receptor, proving that graph-based predictive models can be efficient tools for virtual screening.

[1]Chothia C, Lesk AM, Tramontano A, Levitt M, Smith-Gill SJ, Air G, Sheriff S, Padlan EA, Davies D, Tulip WR, et al. Conformations of immunoglobulin hypervariable regions. Nature. 1989 Dec 21-28;342(6252):877-83. doi: 10.1038/342877a0. PMID: 2687698.

[2]Jeffrey A Ruffolo, Carlos Guerra, Sai Pooja Mahajan, Jeremias Sulam, Jeffrey J Gray, Geometric potentials from deep learning improve prediction of CDR H3 loop structures, Bioinformatics, Volume 36, Issue Supplement_1, July 2020, Pages i268–i275

[3] Alexander Rives, Siddharth Goyal, Joshua Meier, Demi Guo, Myle Ott, C. LawrenceZitnick, Jerry Ma, Rob Fergus, Biological Structure and Function Emerge from Scaling Unsupervised Learning to 250 Million Protein Sequences, *Proceedings of the National Academy of Sciences* doi: 10.1073/pnas.2016239118

O4-4

Leveraging Self-Supervised Contextual Language Models for Deep Neural Network Antibody CDR-H3 Loop Predictions

David Jimenez[1] **Nazim Medzhidov**[1]

david.jimenez@elix-inc.com nazim.medzhidov@elix-inc.com

[1] Elix Inc., Daini Togo Park Building 3F, 8-34 Yonbancho, Chiyoda-ku, Tokyo 102-0081 Japan

Keywords: Protein modeling, Unsupervised Learning, Antibody Structure Prediction, Transformers, Self-Supervision

Immunoglobulins take a structural conformation that is conserved in most parts, except for the antigen-binding fragment (Fab) that includes six complementarity-determining regions (CDRs). These CDRs are peptide loops on both antibody heavy and light chains that impact antigen recognition. Therefore accurate prediction of CDR structures from amino acid sequences is of great importance in antibody design. While existing methods are capable of predicting folding of five of these CDRs [1], determining the structure of the CDR H3 loop remains a challenge. This is due to the complex diversity in observed folding conformations in CDR H3 loop compared to the canonical folds in other five CDR loops. Deep neural networks strive at capturing complex patterns, which makes them a promising tool for protein structure prediction. However, the domain of antibody structure prediction has relatively scarce annotated data compared to general proteins, which usually limits the depth and complexity of the models that can be trained. To bypass this limitation, we propose to use a contextual language model trained unsupervised on a large general protein dataset using a proxy task, which is then joined with a H3-Loop predicting model. Here, the language model spans a representation space reflecting protein biochemical knowledge, which is exploited by the H3-Loop model for the antibody structure prediction. This results in a deeper and more expressive model that outperforms the prediction capabilities of the H3-Loop model alone.

[1]Chothia C, Lesk AM, Tramontano A, Levitt M, Smith-Gill SJ, Air G, Sheriff S, Padlan EA, Davies D, Tulip WR, et al. Conformations of immunoglobulin hypervariable regions. Nature. 1989 Dec 21-28;342(6252):877-83. doi: 10.1038/342877a0. PMID: 2687698.

[2]Jeffrey A Ruffolo, Carlos Guerra, Sai Pooja Mahajan, Jeremias Sulam, Jeffrey J Gray, Geometric potentials from deep learning improve prediction of CDR H3 loop structures, Bioinformatics, Volume 36, Issue Supplement_1, July 2020, Pages i268–i275

[3] Alexander Rives, Siddharth Goyal, Joshua Meier, Demi Guo, Myle Ott, C. LawrenceZitnick, Jerry Ma, Rob Fergus, Biological Structure and Function Emerge from Scaling Unsupervised Learning to 250 Million Protein Sequences, *Proceedings of the National Academy of Sciences* doi: 10.1073/pnas.2016239118

04-5

Using Attribution-based Explainability to Guide Deep Molecular Optimization

Pierre Wüthrich[1] **Jun Jin Choong**[1]
pierre.wuthrich@elix-inc.com junjin.choong@elix-inc.com

[1] Elix Inc., Daini Togo Park Building 3F, 8-34 Yonbancho, Chiyoda-ku, Tokyo 102-0081 Japan

Keywords: Molecular Optimization, Genetic Algorithms, Interpretability, Graph Neural Networks

De novo molecular design is an optimization task where the objective is to find candidate molecules with desired properties. This task is however challenging given the size of the drug-like chemical space. The recently proposed Genetic expert guided learning (GEGL) [1] framework has demonstrated impressive performances on several de novo molecular design tasks. Despite the displayed state-of-the art results, the proposed system relies on an expert-designed Genetic expert. Although hand-crafted experts allow to navigate the chemical space efficiently, designing such experts requires a significant amount of effort and might contain inherent biases which can potentially slow down convergence or even lead to suboptimal solutions. In this research, we propose a novel genetic expert which is free of design rules and can generate new molecules by combining extracted molecular fragments. Fragments are obtained by using an additional graph convolutional neural network [2] which computes attributions [3] for each atom for a given molecule. Molecular substructures which contribute positively to the task score are kept and combined to propose novel molecules. We experimentally demonstrate that our attribution-based genetic expert is competitive on most tasks and even outperforms the previous state-of-the-art expert-designed genetic expert [4] when evaluating proposed candidate molecules is limited. Furthermore, we empirically show that combining several experts that share a fixed sampling budget at each optimization round either improves or maintains the overall performance of the framework.

[1] Ahn, S.-S.; Kim, J.; Lee, H.; Shin, J. Guiding Deep Molecular Optimization with Genetic Exploration. *Advances in Neural Information Processing Systems* **2020**, *33*, 12008–12021.
[2] Kipf, T. N.; Welling, M. Semi-Supervised Classification with Graph Convolutional Networks. Proceedings on the 5th International Conference on Learning Representations. 2017.
[3] Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; Torralba, A. Learning Deep Features for Discriminative Localization. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016; pp 2921–2929.
[4] Jensen, J. H. A Graph-Based Genetic Algorithm and Generative Model/Monte Carlo Tree Search for the Exploration of Chemical Space. *Chemical Science* **2019**, *10*, 3567–3572.

04-6

Improving Molecular Property Prediction using Self-supervised Learning

Laurent Dillard[1]
laurent.dillard@elix-inc.com

[1] Elix Inc., Daini Togo Park Building 3F, 8-34 Yonbancho, Chiyoda-ku, Tokyo 102-0081 Japan

Keywords: Self-supervised learning, transfer learning, molecular property prediction, graph neural networks,

Fast computation of molecular properties holds great potential to boost the efficiency of drug discovery pipelines. In recent years, there has been a surge of interest in developing deep learning models for such applications. A popular choice of architecture to process molecular data are Graph Neural Networks (GNNs). However, as with most deep learning models, training GNNs faces the challenge of gathering large amounts of labeled data. Since labels mostly come from experimental results, the data collection process is both time-consuming and costly. On the other hand, it is easy to access large databases of molecules making it attractive for approaches that do not rely explicitly on labels. Self-supervised learning techniques leverage large amounts of unlabeled data to train models on pretext tasks for which labels can be generated from the raw data. These pretext tasks help the model learn to extract useful feature representations from the data and the trained model can then be fine tuned on downstream tasks. Recently, self-supervised learning techniques have been applied to a growing number of fields, including chemistry. In this work, we introduce a self-supervised framework for GNNs tailored specifically for molecular property prediction. Our framework uses three different pretext tasks, each focusing on a different scale of molecules (atoms, fragments and complete molecules). For the atom and molecule level tasks, a predefined list of fragments is used to encode atom contexts and molecule labels to train the model in a classification setting. For the fragment level task, molecules are decomposed into several fragments and the model is trained to recognize which fragments belong to the same molecule through a binary classification task. Using a subset of ZINC15 molecule database[1] as the pretraining dataset, we evaluate the efficiency of our framework on the MoleculeNet[2] benchmark datasets as well as ADME datasets collected from the literature[3-6]. Our results show that self-supervised learning can successfully improve performance compared to training from scratch, especially in low data regimes. The improvement varies depending on the dataset and model architecture reaching up to +2.6% in area under the curve (AUC) for classification tasks and up to +7% in coefficient of determination (R2) for regression tasks.

[1] Shoichet Brian K Irwin John J. Zinc–a free database of commercially available compounds for virtual screening. Journal of chemical information and modeling, 2005.
[2] Zhenqin Wu, Bharath Ramsundar, Evan N. Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S. Pappu, Karl Leswing, and Vijay S. Pande. Moleculenet: A benchmark for molecular machine learning. CoRR,abs/1703.00564, 2017.
[3] Dominique Douguet. Data sets representative of the structures and experimental properties of fda-approved drugs.ACS Medicinal ChemistryLetters, 9(3):204–209, 2018.
[4] Shuangquan Wang, Huiyong Sun, Hui Liu, Dan Li, Youyong Li, and Tingjun Hou. Admet evaluation in drug discovery. 16. predicting herg blockers by combining multiple pharmacophores and machine learning approaches.Molecular Pharmaceutics, 13(8):2855–2866, 2016. PMID:27379394.
[5] Mark Wenlock and Nicholas Tomkinson. Experimental in vitro dmpk and physicochemical data on a set of publicly disclosed compounds.
[6] Youjun Xu, Ziwei Dai, Fangjin Chen, Shuaishi Gao, Jianfeng Pei, and Luhua Lai. Deep learning for drug-induced liver injury. Journal of Chemical Information and Modeling, 55(10):2085–2093, 2015. PMID:26437739.

Estimation of disease preventive drugs and therapeutic targets using clinical big data

Sae Okamoto[1]
okamoto.sae404@mail.kyutech.jp

Ryusuke Sawada[1]
sawad330@bio.kyutech.ac.jp

Yoshihiro Yamanishi[1]
yamani@bio.kyutech.ac.jp

[1] Department of Bioscience and Bioinformatics, Faculty of Computer Science and Systems
Engineering, Kyushu Institute of Technology, 680-4 Kawazu, Iizuka,
Fukuoka 820-8502, Japan

Keywords: preventive drugs, therapeutic targets, clinical big data, adverse event

Drug development is the most important issue for medical care. However, it is extremely difficult and it requires a huge amount of time and money. Especially, the depletion of therapeutic targets has become a serious problem in recent drug discovery, and the conventional methods for investigating individual diseases are limited in their ability to discover novel therapeutic targets [1]. Recently, there has been an accumulation of clinical and molecular data on various diseases. Thus, there is a strong need to identify novel therapeutic targets by effectively using various big data resources about various diseases.

In this study, we propose a new computational method to predict therapeutic targets via large-scale analyses of clinical big data on patients with various diseases. First, we estimate the potential preventive drugs that are effective in preventing the onset of the target disease by calculating the reporting odds ratio based on the reports of clinical medication history (more than 40 million reports on drug responses and adverse events). Second, we predict proteins, with which the preventive drugs interact, as candidates for therapeutic targets of diseases of interest based on chemical structures and chemical-protein interactome [2]. We applied the proposed method to various diseases, and evaluated its performance in terms of reproducibility for known therapeutic targets. It was observed that the proteins with high prediction scores tended to correspond to the known therapeutic targets of many diseases at the statistically significant level. For example, in the application to Alzheimer's disease (AD), we confirmed that some of the predicted drugs were reported to be effective against AD in recent literature. We also confirmed that some of the predicted target proteins corresponded to known therapeutic targets of AD. For example, butyrylcholinesterase (BCHE) and acetylcholinesterase (ACHE) were detected with high prediction scores. These results show the validity of the proposed method. Other predicted target proteins are expected to be the potential candidates for therapeutic targets.

[1] Santos, R. et al. A comprehensive map of molecular drug targets. Nat Rev Drug Discov. 16, 19-34 (2017).

[2] Sawada, Ryusuke et al. “Target-Based Drug Repositioning Using Large-Scale Chemical-Protein Interactome Data.” *Journal of chemical information and modeling* vol. 55,12 (2015): 2717-30.