Development of FMODB and Auto-FMO protocol through 2022

Chiduru Watanabe[1] chiduru.watanabe@riken.jp
Daisuke Takaya[2] takaya-d@phs.osaka-u.ac.jp
Kikuko Kamisaka[1] kikuko.kamisaka@riken.jp
Teruki Honma[1] honma.teruki@riken.jp

[1] RIKEN BDR, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa 230-0045, Japan
[2] Osaka Univ., 2-8 Yamadaoka, Suita, Osaka 565-0871, Japan

Keywords: FMO, database, structure preparation, interaction energy analysis

Elucidating biomolecular interactions such as protein–ligand, protein–protein, and nucleic acid interactions is essential for structure-based drug discovery. Our group has been focusing on quantum mechanics (QM), which incorporates the effects of donating and withdrawing electrons and can appropriately deal with the CH/π, π-π, and cation/π interactions. Fragment molecular orbital (FMO) method [1] enables us to perform *ab initio* QM calculations for large biomolecules efficiently. The benefit of this fragmentation scheme is the availability of inter-fragment interaction energy (IFIE) and pair interaction energy decomposition analysis (PIEDA).

Since 2014, we have performed the FMO calculations for various drug targets (e.g., kinase, nuclear receptor, protease, protein-protein interaction, and COVID-19) as the activities of the FMO drug design (FMODD) consortium. To accumulate the FMO data, we developed an FMO database (FMODB) and an automated FMO calculation (Auto-FMO) protocol [2,3]. First, FMODB has already completed its basic design and is currently working on extending its analysis capabilities to make it more user-friendly. While only individual entry data could be analyzed in the past, we are working to enable IFIE/PIEDA analysis for a series of related data, such as tens to hundreds of FMO calculations based on molecular dynamics snapshots [4], in a batch. In addition, a function to generate an interaction energy diagram between fragments is under development. Next, the Auto-FMO protocol has also been developed by BIOVIA Pipeline Pilot and MOE software as a primary type. Currently, we are working on developing an Auto-FMO protocol in a configuration using free software such as python and Amber Tools. The protocol was validated using X-ray crystal structure data already registered with FMODB and AlphaFold2 model structures, which are in high demand these days. The goal is to construct a system that allows computational experts and not experts, such as experimental researchers, to perform FMO calculations and analyze IFIE/PIEDA efficiently.

Acknowledgement

The authors thanks Dr. Kazumi Tsuda, Mr. Daisuke Murayama, Dr. Shu Koyama, and Dr. Tomoharu Isobe of Science & Technology Systems, Inc., Japan for technical support. This research was done in activities of FMO Drug Design Consortium, https://fmodd.jp/top-en/. The results of FMO calculations were obtained using the Fugaku (project IDs: hp220143 and ra000017) and HOKUSAI (ID: Q22306).

Reference

1. K. Kitaura *et al.*, *Chem. Phys. Lett.*, **1999**, *313*, 701–706.
2. C. Watanab *et al.*, *CBI J.*, **2019**, *19*, 5–18.
3. D. Takaya *et al., J. Chem. Inf. Model.*, **2021**, *61*, 777–794.
4. K. Takaba *et al.*, *J. Comput. Chem.*, **2022**, *43*, 1362–1371.

Hit Identification for SARS-CoV-2 Main Protease Using Convolutional Neural Network

Nobuaki Yasuo[1]
yasuo.n.aa@m.titech.ac.jp

Hiroshi Yoda[2]

Masakazu Sekijima[2]
sekijima@c.titech.ac.jp

[1] Academy for Convergence of Materials and Informatics (TAC-MI), Tokyo Institute of Technology, S6-404, Ookayama 2-12-1, Meguro-ku, Tokyo, 152-8550, Japan
[2] Department of Computer Science, Tokyo Institute of Technology, 4259-J3-23 Nagatsutacho, Midori-ku, Yokohama, Kanagawa, 226-8501, Japan

Keywords: Virtual screening, Ligand docking, Machine learning, SARS-CoV-2

Docking-based virtual screening is widely applied in drug discovery campaigns. Visual inspection of binding modes is frequently used to filter out false positive-like docking results[1, 2]. In this study, we propose a convolutional neural network-based virtual screening method named VisINet (Visual Inspection Network). VisINet predicts whether a ligand binds to the target protein by images that is taken from docking pose using PyMOL. We also performed virtual screening to find hit compounds of SARS-CoV-2 Main Protease.

[1] Fischer, A.; Smieško, M.; Sellner, M.; Lill, M. A. Decision Making in Structure-Based Drug Discovery: Visual Inspection of Docking Results. *Journal of Medicinal Chemistry* **2021**, *64* (5), 2489–2500.

[2] Singh, K.; Coopoosamy, R. M.; Gumede, N. J.; Sabiu, S. Computational Insights and in Vitro Validation of Antibacterial Potential of Shikimate Pathway-Derived Phenolic Acids as Nora Efflux Pump Inhibitors. *Molecules* **2022**, *27* (8), 2601.

Evaluating Computer-assisted Single-step Retrosynthesis: How Significant are Recent Improvements?

Haris Hasic[1,2]
haris.hasic@elix-inc.com

Takahiro Inoue[1]
takahiro.inoue@elix-inc.com

Tatsuya Okubo[1]
tatsuya.okubo@elix-inc.com

Takashi Ishida[2]
ishida@c.titech.ac.jp

[1] Elix Inc., Daini Togo Park Building 3F, 8-34 Yonbancho, Chiyoda-ku, Tokyo 102-0081, Japan
[2] Department of Computer Science, School of Computing, Tokyo Institute of Technology, W8-85, 2-12-1 Ookayama, Meguro-ku, Tokyo 152-8552, Japan

Keywords: Machine Learning – AI Method Development, Synthetic Pathway Prediction

Computer-assisted synthesis grew to become one of the most active fields in chemoinformatics, with scientists all over the globe frequently introducing new and improved approaches. In recent years, the research focus shifted towards the retrosynthesis technique concepts of transforming the target compound into less complex precursor compounds. The improvement of the single-step retrosynthesis mechanism (*i.e.*, the disconnection suggestion procedure within the route planning strategy) is among the most popular objectives. However, none of the recent improvements proved to be a robust enough foundation to establish a practically applicable, standalone solution.

The current practices in modelling the problem of single-step retrosynthesis might conceal a hard ceiling on performance, but it is impossible to know because of insufficient chemical reaction data. Nevertheless, analyzing the current state-of-the-art more rigorously using available open-source datasets could help to shed more light on that question. Thus, the purpose of this research is to consolidate, analyze and discuss the significance of the recent frequent improvements in the field of computer-assisted single-step retrosynthesis.

First, the Top-3 template-based **[1-3]** and the Top-3 template-free **[4-6]** single-step retrosynthesis approaches as of July 2022 are implemented and adjusted to customized open-source chemical reaction datasets. Secondly, the performance of these approaches is analyzed using various metrics and statistical significance tests. Finally, using the acquired results, the impact of each of these approaches is illustrated, thereby consolidating the best practices for the analysis and evaluation of future research in this field.

[1] Dai, H.; Li, C.; Coley, C. W.; Dai, B. W.; Song, L. Retrosynthesis Prediction with Conditional Graph Logic Network, *Advances in Neural Information Processing Systems*, **2019**, *32*.
[2] Somnath, V. R.; Bunne, C.; Coley, C. W.; Krause, A. W.; Barzilay, R. Learning Graph Models for Retrosynthesis Prediction, *Advances in Neural Information Processing Systems*, **2021**, *34*, 9405-9415.
[3] Chen, S.; Jung, Y. Deep Retrosynthetic Reaction Prediction using Local Reactivity and Global Attention, *JACS Au*, **2021**, *1*, *10*, 1612-1620.
[4] Tetko, I. V.; Karpov, P.; Van Deursen, R.; Godin, G. State-of-the-Art Augmented NLP Transformer Models for Direct and Single-step Retrosynthesis, *Nature Communications*, **2020**, *11*, 5575-5586.
[5] Ucak, U. V.; Ashyrmamatov, I.; Ko, J.; Lee, J. Retrosynthetic Reaction Pathway Prediction through Neural Machine Translation of Atomic Environments, *Nature Communications*, **2022**, *13*, 1186-1196.
[6] Zhong, Z.; *et al.* Root-aligned SMILES: A Tight Representation for Chemical Reaction Prediction, *arXiv:2203.11444v4 [cs.LG]*, **2022**.

Developing a network-based combination therapy approach for complex diseases

Midori Iida[1] redgreen@bio.kyutech.ac.jp
Yurika Kuniki[2] c401703080@tokushima-u.ac.jp
Kenta Yagi[3] yagi.kenta@tokushima-u.ac.jp
Mitsuhiro Goda[2,4] mgoda@tokushima-u.ac.jp
Satoko Namba[1] namba.satoko775@mail.kyutech.jp
Jun-ichi Takeshita[5] jun-takeshita@aist.go.jp
Ryusuke Sawada[1] sawad330@bio.kyutech.ac.jp
Michio Iwata[1] iwata121@bio.kyutech.ac.jp
Yoshito Zamami[2,6] zamami-y@okayama-u.ac.jp
Keisuke Ishizawa[2,3,4] ishizawa@tokushima-u.ac.jp
Yoshihiro Yamanishi[1] yamani@bio.kyutech.ac.jp

[1] Kyushu Institute of Technology, Iizuka, Fukuoka 820-8502, Japan
[2] Department of Clinical Pharmacology and Therapeutics, Tokushima University Graduate School of Biomedical Sciences, Kuramoto-cho, Tokushima 770-8503, Japan
[3] Clinical Research Center for Developmental Therapeutics, Tokushima University Hospital, Kuramoto-cho, Tokushima 770-8503, Japan
[4] Department of Pharmacy, Tokushima University Hospital, Kuramoto-cho, Tokushima 770-8503, Japan
[5] Research Institute of Science for Safety and Sustainability, National Institute of Advanced Industrial Science and Technology (AIST), Tsukuba, Ibaraki 305-8569, Japan
[6] Department of Pharmacy, Okayama University Hospital, Kita-ku, Okayama 700-8558, Japan

Keywords: Combination therapy, Cancer, Network medicine, Interactome, Transcriptome

Combination therapy, a treatment modality that combines two or more therapeutic drugs, has become the standard treatment for complex diseases—however, the discovery of drug combinations relies on serendipity by clinical experience and intuition [1]. A network-based approach allows us to discover efficient drug combination by quantifying the network-based relationship between drug targets and disease proteins in the human protein–protein interactome [2]. However, conventional methods do not take into account changes in gene expression due to drug administration, resulting in low prediction accuracy. In this study we developed a network-based method to predict synergistic drug effects by integrating interactome and transcriptome data. The proposed method outperformed the previous methods in terms of higher prediction accuracy for several cancers. To validate the prediction results by our proposed method, we carried out in vitro cell surviving assays. The results showed that more than half of the top 20 predicted drug pairs had synergistic effects on the cell survival. Pathway enrichment analysis demonstrated several biological pathways are enriched with the genes used for prediction, suggesting that those pathways contribute to the synergistic effect. This approach allows us to identify drug combinations for complex diseases, offering a powerful methodology to find efficacious combination therapies.

[1] Prasad, S.; Gupta, S. C.; Aggarwal, B. B. Serendipity in Cancer Drug Discovery: Rational or Coincidence? *Trends in Pharmacological Sciences,* **2016**, *37* (6), 435–450.
[2] Cheng, F.; Kovács, I. A.; Barabási, A. L. Network-Based Prediction of Drug Combinations. *Nature Communications,* **2019**, 10 (1).

Prediction of Protein Binding Region on RNA with Transformer and Attention Augmentation

Takayuki Kimura[1]
kimura.t.bf@m.titech.ac.jp

Nobuaki Yasuo[2]
yasuo@cbi.titech.ac.jp

Masakazu Sekijima[1]
sekijima@c.titech.ac.jp

[1] Department of Computer Science, Tokyo Institute of Technology, Tokyo, 152-8550, Japan
[2] Academy for Convergence of Materials and Informatics, Tokyo Institute of Technology, Tokyo, 152-8550, Japan

Keywords: RNA-protein binding prediction, Transformer, Statistical potential

RNA-protein interaction is important in post-transcriptional regulations [1]. However, the size of public databases is limited [2], and the cost of experimental approaches [3] is not low. Therefore, computational prediction is necessary that predicts RNA-protein binding. The computational model's evolution is nevertheless insufficient [4]. For example, accurate prediction models for protein binding regions in RNA have not been reported yet. In this study, we developed a Transformer model [5] that predicts a protein binding region (101 nt). We adopted cross-attention module architecture [6] and used statistical potential [7] as prior knowledge to augment the attention weights [8]. The cross-attention module can generate feature vectors from the perspectives of both the RNA and the protein. The test AUROC was higher than a baseline model of bidirectional-LSTM and existing models. We also confirmed the effect of attention augmentation with statistical potential on the binding prediction.

[1] Hentze MW.; Castello A.; Schwarzl T.; Preiss T, *A brave new world of RNA-binding proteins, Nature reviews Molecular cell biology*, **2018**, *19*, 327-41.
[2] ENCODE Project Consortium, *An integrated encyclopedia of DNA elements in the human genome*, *Nature*, **2012**, *489*, 57.
[3] Van Nostrand, E.L.; Pratt, G.A.; Shishkin, A.A.; Gelboin-Burkhart, C.; Fang, M.Y.; Sundararaman, B.; Blue, S.M.; Nguyen, T.B.; Surka, C.; Elkins, K. and Stanton, R, *Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP), Nature methods*, **2016**, *13*, 508-514.
[4] Wei J.; Chen S.; Zong L.; Gao X.; Li Y, *Protein–RNA interaction prediction with deep learning: structure matters, Briefings in bioinformatics,* **2022,** *23*, bbab540.
[5] Vaswani A.; Shazeer N.; Parmar N.; Uszkoreit J.; Jones L.; Gomez AN.; Kaiser Ł.; Polosukhin I, *Attention is all you need, Advances in neural information processing systems*, **2017,** 30.
[6] Koyama K.; Kamiya K.; Shimada K, *Cross attention DTI: Drug-target interaction prediction with cross attention module in the blind evaluation setup, 19th International Workshop on Data Mining in Bioinformatics*, **2020**, BIOKDD, San Diego
[7] Kimura T.; Yasuo N.; Sekijima M.; Lustig B, *Statistical potentials for RNA-protein interactions optimized by CMA-ES, Journal of Molecular Graphics and Modelling*, **2022**, *110*, 108044.
[8] Maziarka Ł.; Danel T.; Mucha S.; Rataj K.; Tabor J.; Jastrzębski S, *Molecule attention transformer, arXiv preprint arXiv*, **2020**, 2002.08264.

Molecular Generation using Sequence-based Transformer Generative Adversarial Network

Chen Li[1]
li260@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: Transformer, Deep generative model, Reinforcement learning, Molecular generation

Generating molecules with the desired chemical properties from scratch is a challenging task in de novo drug design. Deep generative models such as variational autoencoder (VAE) and generative adversarial network (GAN) have been receiving much attention in recent years. In deep generative models, molecules are usually represented as molecular graphs and simplified molecular-input line-entry system (SMILES) strings. In previous work, a Transformer-based objective-reinforced GAN (TransORGAN) [1] was proposed to generate molecules with desired chemical properties using SMILES strings. The generator of TransORGAN is a transformer architecture, which uses a self-attention mechanism to capture features in sequences. Self-attention has a global receptive field that can capture long-range dependencies. Since TransORGAN uses the SMILES strings from the training set as a condition rather than generating molecules starting from noise or scratch, the molecular generation is restricted. Furthermore, the instability of TransORGAN leads to a high variance in the distribution of the chemical properties of the generated molecules, which is an obstacle in practice.

In this study, we propose a pure transformer encoder-based GAN for generating molecules with desired properties to solve the above problems. The generator and discriminator of the proposed model are variants of the transformer encoders, where atoms can access each other at each encoder layer to capture complex semantic and syntactic rules of SMILES strings. Furthermore, we implement an enhanced algorithm incorporating the mini-batch discrimination [2] and Wasserstein GAN [3] to mitigate the mode collapse problem. In addition, we present variant SMILES, a data augmentation to sufficiently train the generator in the training phase. Finally, we use Monte Carlo policy gradient reinforcement learning [4] to improve the desired chemical properties of generated molecules. To our knowledge, we are the first to produce molecules with chemical properties from SMILES strings using only transformer encoders. The experimental results demonstrate the usefulness of the proposed models for generating molecules with desired properties such as drug-likeness. Ablation studies validate the effectiveness of our proposed techniques. The proposed method is expected to be useful for de novo drug design.

[1] C. Li; C. Yamanaka; K. Kaitoh; Y. Yamanishi. Transformer-based objective-reinforced generative adversarial network to generate desired molecules, *Proc. In the 31th International Joint Conference on Artificial Intelligence* (*IJCAI)*, 3884-3890, 2022.
[2] T. Salimans; I. Goodfellow; W. Zaremba; V. Cheung; A.Radford; X.Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29:2234–2242, 2016.
[3] I. Gulrajani; F. Ahmed; M. Arjovsky; V. Dumoulin; A. Courville. Improved training of wasserstein gans. *arXiv preprint arXiv:*1704.00028, 2017.
[4] R.S. Sutton; D.A. McAllester; S.P. Singh; Y. Mansour. Policy gradient methods for reinforcement learning with function approximation, *Proc. In Advances in Neural Information Processing SystemsAdv Neural Inf Process Syst. 12*, 1057-1063, 1999.