TOHOKU UNIVERSITY

Graduate Program in Data Science

2nd Data Science Cafe

"Data Science Workshop" held by the Graduate School of Information Sciences Tohoku University





2023 March 24 FRI

13:00 - 13:25 Talk 1: **Dr. Lin Zhang** (Tohoku University)

13:30 - 14:30 Talk 2 (Invited): **Dr. Shawn McGlynn** (Tokyo Institute of Technology)

14:35 - 15:15 Talk 3 (Invited): **Dr. Anthony Poole** (University of Auckland)

15:20 - 15:45 Talk 4 (Invited): **Ms. Nicole Tamer** (University of Zurich)

15:50 - 16:25 Talk 5: **Dr. Takeshi Obayashi** (Tohoku University)

＊Refreshments and snacks will be served during the discussion time.

MAP

Venue 会 場

Large Meeting Room (3F), Innovation Center for Creation of a Resilient Society
東北大学青葉山キャンパスレジリエント社会構築イノベーションセンター 3 階大会議室

Contact

JP

東北大学大学院情報科学研究科
山田和範
Phone: 022-752-2206
Email: yamada@tohoku.ac.jp

EN

Siwalee Choilek
Graduate School of Information Sciences
Phone: +81-22-795-4691
Email: choilek.siwalee.a5@tohoku.ac.jp

English

Japanese

Integrated analyses of single-cell RNA-seq public data reveal the gene regulatory network landscape of respiratory epithelial and peripheral immune cells in COVID-19 patients.

Lin Zhang

Graduate School of Information Sciences, Tohoku University

Introduction: Infection with SARS-CoV-2 leads to coronavirus disease 2019 (COVID-19), which can result in acute respiratory distress syndrome and multiple organ failure. However, its comprehensive influence on pathological immune responses in the respiratory epithelium and peripheral immune cells is not yet fully understood.

Methods: In this study, we integrated multiple public scRNA-seq datasets of nasopharyngeal swab and peripheral blood results to investigate the gene regulatory networks (GRNs) of healthy individuals and COVID-19 patients with mild/moderate and severe disease, respectively. Similar and dissimilar regulons were identified within or between epithelial and immune cells during COVID-19 severity progression. The relative transcription factors (TFs) and their targets were used to construct GRNs among different infection sites and conditions.

Results: Between respiratory epithelial and peripheral immune cells, different TFs tended to be used to regulate the activity of a cell between healthy individuals and COVID-19 patients, although they had some TFs in common. For example, XBP1, FOS, STAT1, and STAT2 were activated in both the epithelial and immune cells of virus-infected individuals. In contrast, severe COVID-19 cases exhibited activation of CEBPD in peripheral immune cells, while CEBPB was exclusively activated in respiratory epithelial cells. Moreover, in patients with severe COVID-19, CEBPD upregulated S100A8 and S100A9 in CD14 and CD16 monocytes, while S100A9 genes were co-upregulated by different regulators (SPEDEF and ELF3) in goblet and squamous cells. The cell-cell communication analysis suggested that epidermal growth factor receptor signaling among epithelial cells contributes to mild/moderate disease, and chemokine signaling among immune cells contributes to severe disease.

Conclusions: This study identified cell type- and condition-specific regulons in a wide range of cell types from the initial infection site to the peripheral blood, and clarified the diverse mechanisms of maladaptive responses to SARS-CoV-2 infection.

New evolutionary and algorithmic views on early life.

Shawn McGlynn

(1) Earth-Life Science Institute, Tokyo Institute of Technology
(2) Blue Marble Space Institute of Science, and
(3) Center for Sustainable Resource Science, RIKEN

There are two records of early evolution on Earth: One is material and seen in the form of stable isotope ratios and physical biomarkers such as stromatolites. The second is that of molecular evolution as we know from phylogenetics. In this talk I'll discuss some of our recent work in phylogenetics, as well as the limitations of this work. Recognizing the limitations, I'll introduce recent progress using network expansion algorithms to find a pathway for metabolisms emergence.

Keywords: Phylogeny, network expansion, LUCA

Building evolutionary trees from protein structure.

Anthony M. Poole

School of Biological Sciences, University of Auckland, New Zealand
a.poole@auckland.ac.nz

Evolutionary trees are traditionally made from multiple sequence alignments of DNA or protein sequences. This has been a spectacularly successful approach to charting the evolutionary relationships between all life on Earth. The approach makes a number of assumptions, and focuses on the information found in sequences. However, it is widely accepted that protein structure evolves slower than sequence, so very distant relationships may be evidenced through structural comparisons, even when it is difficult to detect similarities between sequences. We have therefore been working to develop methods for building evolutionary trees that make use of protein structure.

Our approach is to compare protein structures to generate scores of how dissimilar proteins are. We can then use these data to build an evolutionary tree. To test how reasonable the trees are, we sample variations in structure from molecular dynamics simulations and then we use these variant structures to test whether there is statistical support for our phylogeny derived from structure.

This approach allows us to build trees for datasets where the structures are conserved, but sequence data are poorly conserved. It therefore allows us to build trees for protein superfamilies that are too divergent for traditional sequence-based phylogenies.

I will present our work to date, using examples from the Ferritin Superfamily, the Histone Fold, and the Jelly Roll Fold that makes up viral capsids. I will also outline our plans for producing ‘total-evidence’ trees, that make use of both sequence and structure, and I will consider how we might be able to make use of computational predictions of protein structure such as those made by AlphaFold.

An interdisciplinary investigation of categorical call distinction in meerkat vocalisations

Nicole Tamer

(1) Department of Comparative Language Science, University of Zurich
(2) The Swiss National Centre of Competence in Research

In human language, the first word segment can help distinguish between grammatical categories. Non-human animal species have also demonstrated the ability to differentiate between various call categories and react in response suitable to the context. Particularly meerkats have a complex vocalisation system with distinct call categories. However, meerkat calls were looked at as a whole in previous studies and little is known about which part of the call conveys information about call categories and how they are processed. Since meerkat alarm calls contain vital information and thus need to be processed instantly, they appear to be a promising call type to investigate whether the first call segment can help discern them. We hypothesise that the first segment of meerkat calls differentiates call categories and that alarm calls are more likely to be distinguished by the first call segment because of their urgency. By combining methods used on human languages and conducting a behavioural study with playback experiments, we hope to shed more light on how meerkats process and distinguish call categories.

Gene coexpression resource for predicting all metabolic pathways in a species

Takeshi Obayashi

Graduate School of Information Sciences, Tohoku University

Thanks to sequencing technology, we can determine the genome sequence of any species. However, functional prediction of the genes in a genome is a rather difficult problem that also limits studies with non-model species. Gene coexpression information, *i.e.*, a similarity of expression profiles, is useful for identifying functionally related genes and constructing gene networks. We have built such databases based on publicly available transcriptome data: ATTED-II [atted.jp] for plants, COXPRESdb [coxpresdb.jp] for animals, and ALCOdb [alcodb.jp] for microalgae. In this talk, I will discuss the construction and visualization of a gene coexpression resource and demonstrate its utility for metabolic pathway prediction using several examples.