The 5th RIKEN-Karolinska Institutet/SciLifeLab Joint Symposium

Artificial Intelligence Meets Life Sciences

Abstract Book

This symposium is part of the RIKEN Symposium series
This event is also supported by Japan Society for the Promotion of Science (JSPS), Stockholm Office

Table of Contents

Revealing hidden patterns via integration of biomedical Big Data by unsupervised Deep Learning

Nikolay Oskolkov, National Bioinformatics Infrastructure (NBIS), SciLifeLab

Next Generation Sequencing (NGS) technologies gave rise to manifolds of Biological and Biomedical Big Data which is particularly manifested in the area of single cell transcriptomics where the number of samples approaches hundreds of thousands of cells sequenced (X. Han et al., Cell 2018). The rapidly growing amount and diversity of data provides new unique opportunities as well as poses a number of challenges for the analysis.

Biomedical Big Data from different sources (OMICs data) should have synergistic effects which allows to model the behavior of biological cells. In this way, OMICs integration can identify novel biological pathways that are not necessarily distinguishable in the separate OMICs layers. Further, new mathematical methodologies are needed to deal with Big Data, among them Artificial Intelligence (AI) and Deep Learning (DL) are ideally suited for processing and integration of large amounts of data as well as generating predictive models that can be used in e.g. Clinical Diagnostics within the concept of Precision Medicine. Finally, taken into account the highly non-linear nature of single cell sequencing data, the key advantage of DL is its superior ability to learn non-linear relationships in the data.

In this work, we utilize single cell transciptomics (scRNAseq), DNA methylation (scBSseq) and chromatin accessibility (scATACseq) sequencing data from S. J. Clark et al., Nature Communications 2018, to demonstrate the power of non-linear data integration via Deep Autoencoder which is an unsupervised Deep Artificial Neural Network (DANN). We show that combining data through Deep Autoencoder provides the most flexible approach to discovering biological similarities between cells which might be hidden in each individual OMIC due to technological noise. The Deep Autoencoder searches for consistencies across multiple OMICs by learning most fundamental features in each data set and thus reduces the noise level and makes the biological signal more pronounced. We show that the use of regularizations as well as non-linear dimensionality reduction and clustering implemented in the Deep Autoencoder results in discovering novel biological patterns and provides wide potential for applications in Biology, Medicine and Clinical Diagnostics.

Machine learning for complex disease genetics: risk prediction and interpreting GWAS findings

Masaru Koida and Yoichiro Kamatani, RIKEN Center for Integrative Medical Sciences

Genome-wide association study (GWAS) is an unbiased way to test associations between complex traits and millions of genotypes. To date, GWAS has identified more than tens of loci robustly associated with each disease (P<5´10^{-8}) and provided plausible and novel insights of disease etiology. In contrast, the risk effect of each variant was quite small, which makes hard to utilize this information in clinic or in the society. Meanwhile, statistical approaches have elucidated that much more variants across the genome might be associated, in agreement with the earlier concept suggested a hundred years ago called infinitesimal model. Therefore, hundreds or thousands of variants may work cooperatively in a linear and non-linear manner. However, due to the combinatorial explosion, the frequentism-based approach has failed to explicitly model these linear and non-linear effects among the variants, also known as additive, dominance and epistatic genetic effects in the context of classical genetics. To overcome these difficulties and increase predictive ability by genetic variants, we have been applying machine learning (ML) methodologies to GWAS framework. We applied ML methods such as Lasso, Random Forest and Support Vector Machine (SVM) into a dataset of BioBank Japan (BBJ), which had enrolled ~200K patients with 47 target diseases and collected whole-genome SNP data with clinical information. We identified that ML using nonsignificant variants in frequentist statistics achieved moderate predictive accuracies of disease risks (c-index>0.7, 10-fold cross-validations) other than those by SVM (c-index~0.5), indicating that careful ML design may lead to clinical applications of genetic variants. In this presentation, we will also discuss our ongoing research for ML-driven interpretation of heritability by integrating epigenomic and transcriptomic data.

Novel Health Quotients through AI-based big health data analyses

Yasuyoshi Watanabe and Kei Mizuno, RIKEN Center for Biosystems Dynamics Research

We usually just say "I'm healthy" unless we realize evidence of disease onset. However, of importance is the extent of health, namely, extremely healthy, moderately healthy, less healthy, or pre-disease state?? There are a lot of indices of pre-disease state or prelude of disease onset or ahead sick condition. For example, index of metabolic syndrome and/or hemoglobin A1C are the biomarker toward diabetes mellitus. However, we are not always toward just one specific disease, but the risks toward a variety of diseases are creeping up to us at every moment. So, we need novel health quotients with thorough evaluation of biomarkers of pre-disease state and disease (we proposed the word "Precision Health"). Especially, the system error or dysfunction of sensing (checking) mechanisms toward pre-disease state and also recovery engine system from the pre-disease state. From long years' researches on "fatigue" and "chronic fatigue," we focused on autonomic nerve function, biological oxidation by reactive oxygen species, less repair energy, and local inflammation/immune response for the candidates of health quotients. We have been developing many novel molecular and functional imaging technologies for such health disturbance. And also we introduced the packages of simpler functional test such as cognitive, locomotive, and skin functions. Including various questionnaire, blood test, blood biochemistry test, breathing gas test, and skin gas test, we made 232 factors/items measurement for 692 healthy (by self-judgement) subjects. After multi-dimensional big data analyses with machine learning, we succeeded to find the clusters for different types of pre-disease states. By using these quotients, we could invent the goods and services to prevent from different angles of vulnerability of health. Now, we started to confirm these health quotients and the way or mathematical function to calculate the health quotients with 10,000 subjects.

The Description of biological phenomena as an open system using Machine learning and Markov constraint

Kazuhiro Sakurada, RIKEN Medical Sciences Innovation Hub Program

Personalized medicine is a new paradigm that represents a shift from a statistical abstraction of the patients toward the view that each patient is unique. This is a new scientific challenge as well as a new social challenge. Although linear causations and correlations have been used in the explanation of biological phenomena, biological systems form complex network whose collective behavior cannot be reduced to simple correlations. In addition, explanations usually eliminate information on differences between each individual patient. To overcome this problem, we are developing a new biomedical science based on pure description of diseases by using multi-omics data.

Human beings are not genetically programmed systems—they are historical systems that organize themselves through cooperation. For this reason, we humans must define ourselves not by *spatial* properties like structure and function, but by *temporal* properties. In order to define what exactly these "time properties" are, I am working to collect and organize, in machine-readable form, "life course data" related to the physical condition of a person and then leverage the power of artificial intelligence to analyze that data.

Using a hidden hierarchical Markov model to calculate state transition probability, I have discovered a method to represent individual characteristics by using the concept of degree of freedom, as well as the limits of that freedom. The state allocation is done by the reduction of dimensionality and data granularity using machine learning and energy land scale analysis. Now these new concepts are applied to the data gathered from patients with immune disorders, cancer and developmental disorders.

Omic analysis with AI drives precision medicine

Tatsuhiko Tsunoda, RIKEN Center for Integrative Medical Sciences

Precision medicine requires the prediction of an individual's body and disease state to estimate the incidence risk for disease prevention or the efficacy and side-effects of each treatment when deciding the optimum therapy. Recent advancement of omic profiling technologies will greatly contribute to this goal, but they need novel analysis techniques to cope with the high dimensionality and spatio-temporal dependency of the data. Fundamental techniques in artificial intelligence (AI), including machine learning (ML) and deep learning (DL), are likely to be leveraged for their ability to extract features and predict outcomes. With this in mind, I introduce our recent results: (1) type 2 diabetes risk prediction with GWAS data (*PLoS One*, **9**:e92549 (2014)), (2) prognosis prediction for breast cancer with genome-wide gene expression data (*Cancer Medicine*, **6**:1627-1638 (2017)), (3) survival prediction for hepatocellular carcinoma with whole-genome sequencing and omic data (*Nature Genetics*, **48**:500-509 (2016)), and (4) drug efficacy prediction for lung cancer with omic data applied with a semi-supervised technique. In addition, I discuss what can be expected from the implementation of DL, feature extraction, dimensional reduction, and integration of omic and image data, particularly in the context of cancer immunology.

Developing prediction methods for disease treatment and prevention taking into account individual variability

Mayumi Kamada, RIKEN Cluster for Science, Technology and Innovation Hub

Precision medicine including genomic medicine is starting to be realized through the utilization of big data in life science and healthcare. Genomic medicine aims to make individualized diagnostic and therapeutic decision based on patient's genomic information. For clinical realization in Japan, we have been developing a disease-related genomic information database, Medical Genomics Variant Database Japan (MGeND). However, many of the detected genomic variants are unclear in relation to mechanism of disease and often do not lead to clinical determination. Enormous effort to aggregate evidences and specialized knowledge are required to give a clinical interpretation to those variants, expectations are rising for the efficiency of interpretation utilizing AI. In this presentation, I would like to talk about AI development for genomic medicine and introduce the activities of Life Intelligence Consortium (LINC), which is a consortium for development of AI and big data technology in the life science fields.

Computational Neuroimaging for Understanding Primate Connectomics

Takuya Hayashi, RIKEN Center for Biosystems Dynamics Research

Primate brain has gained the ability for processing higher-order multi-modal information for achieving goal-directed behaviors and adapting environments. The human brain is 200-220 times larger in its volume, 60-90 times larger in cortical area and number of cortical neurons, as compared with those in marmoset, an experimental primate animal model. We have recently established the imaging technologies for comparative neuroscience, in which high-quality structural and functional MRI data can be obtained in the same MRI scanner across three species of primates, macaque, marmoset and humans. We would establish the way to precisely delineate cortical convolution, functionally parcellate cerebral cortex, and investigate organization of large-scale connectivity, which we call 'connectome'. The analysis includes supervising machine learning for brain and cortical parcellations, but we also need to establish unsupervised approach for fast and reliable analysis of big data. The goal of our approach is to establish brain and cortical functional registration between species and to find comparable architecture and its evolution to help understanding dynamics and pathologies of human brain.

Integration of big image and omics data for modeling of animal development

Shuichi Onami, RIKEN Center for Biosystems Dynamics Research

Applications of computational image processing to four-dimensional microscope images of cells enable high-throughput quantification of spatiotemporal dynamics of cells; The dynamics cells includes that of cell position, cell shape and cellular gene expression. The resultant large data can be applied to data-driven analysis. In this talk, I will present our data-driven analysis of *C. elegans* embryo, as an example of data-driven analysis of animal development enabled by bioimage informatics. We developed a computational method for inferring causal network among phenotypic characters' expressions during embryogenesis by finding correlations between phenotypic characters' expressions in our large image data collection of wild-type embryos. We also developed another computational method for inferring genes involved in the causal interactions among phenotypic characters' expressions by using our large image data collection of gene knockout embryos against essential embryonic genes. We created a model of *C. elegans* early embryogenesis by integrating a causal network of phenotypic characters' expressions consisting of 3,372 causal interactions among 421 phenotypic characters, and gene regulatory network deduced by using multi-omics data. I will also present our SSBD database for storing and sharing quantitative data of biological dynamics and bioimages, and RIKEN life science data platform project that aims to integrate all life science basic research data in RIKEN. Then I would like to discuss on current and future contribution of deep learning and AI in our projects and data-driven developmental biology.

Deep learning as a tool in microscopy data analysis and digital pathology

Carolina Wählby, Department of Information Technology, Uppsala University

Digital image processing and analysis makes it possible to use microscopes not only as a means of producing pretty pictures, but also as a tool to extract quantitative measurements from experiments involving live or fixed biological samples imaged in multiple dimensions and over time. Automatically extracted quantitative information becomes even more important when conducting large-scale experiments, such as high-throughput image-based screens of potential drugs or genetic perturbations. A number of commercial software as well as free and open source alternatives are available for setting up analysis approaches designed for identifying objects of interest, extracting descriptive measurements, making classifications, and drawing conclusions. This 'classical approach' to image processing and analysis requires knowledge of the underlying biology, to hypothesize a range of morphological responses, as well as knowledge of image analysis, to select a set of parameters suitable for extracting measurements describing these morphological responses. These 'classical approaches' are now being outperformed by systems based on deep learning, overcoming many of the limitations in our ability to translate visual ques to descriptive measurements. We have shown that learning-based approaches may even identify morphological changes that were previously not observed by visual inspection of the image data. We also show how combinations of imaging modalities, rather than tedious manual annotations, may be used to train deep learning networks. In an era of great enthusiasm, it is however also important to be aware of the pitfalls, and the fact that what is learnt may not be what was initially intended. We also see how it is possible to overcome limitations in microscope optics, fluorophore chemistry, and sample exposure by deep-learning based image restoration.

A deep learning approach to breast cancer risk assessment

Kevin Smith, Department of Computational Science and Technology, Royal Institute of Technology

Breast cancer is one of the most common forms of cancer worldwide, and the incidence is rising. It is crucial to identify women with the disease at an early stage. One approach is to select certain women for screening by magnetic resonance imaging (MRI) in addition to the established mammographic examination. MRI is more sensitive, but also costly and time-consuming, and cannot be provided to the entire population. To accurately select the women most likely to benefit from additional diagnostics, we trained a state of the art deep learning network to estimate the risk of developing cancer by distinguishing between historic mammograms from women who developed breast cancer and women who remained healthy. We compare the performance of the deep learning risk predictions against the current standard practice: estimated mammographic density.

Artificial intelligence for images based diagnostics

Nina Linder, FIMM-Institute for Molecular Medicine Finland

Visual assessment of tissue morphology remains the gold standard method for diagnosis of solid tumours. However, subjectiveness of manual tissue examination reduces the accuracy and reproducibility of diagnostic decision-making. Advances in artificial intelligence allow to build powerful classifiers, eliminating the need of feature engineering, i.e. complex labels can be learnt automatically and directly from raw inputs such as images of cancer tissue samples. With the rise of deep learning techniques and quantity of digitized samples it has become feasible to address more demanding tasks such as prediction of disease outcome, bypassing intermediate tissue characterisation such as grading of cancer. Outcome prediction is crucial for patient stratification and to aid clinical decision-making, e.g. choice of adjuvant therapy, to achieve a more personalized and cost-effective treatment.

We take advantage of artificial intelligence to learn directly from large amounts of raw image data eliminating the biases introduced by manual annotations and feature engineering thus reducing human bias. We compare the results of machine learning-based analysis with that of visual assessment of colorectal and breast cancer performed by an expert as well as with clinical parameters. Also, we visualize hot-spots on tissue images that are considered by the machine learning model as the most predictive of the desired target output. This allows to better understand what drives the neural network decisions. Our studies reveal novel imaging biomarkers for colorectal and breast cancer and leads to improved diagnostics and stratification of cancer patients. Moreover, computer-aided analysis improves standardization and reproducibility within image-based cancer diagnostics.

Image based cancer survival modeling with machine learning

Dimitrii Bychkov, FIMM-Institute for Molecular Medicine Finland

Visual assessment of tissue morphology remains the gold standard method for diagnosis of solid tumours. However, subjectiveness of manual tissue examination reduces the accuracy and reproducibility of diagnostic decision-making. Advances in artificial intelligence allow to build powerful classifiers, eliminating the need of feature engineering, i.e. complex labels can be learnt automatically and directly from raw inputs such as images of cancer tissue samples. With the rise of deep learning techniques and quantity of digitized samples it has become feasible to address more demanding tasks such as prediction of disease outcome, bypassing intermediate tissue characterisation such as grading of cancer. Outcome prediction is crucial for patient stratification and to aid clinical decision-making, e.g. choice of adjuvant therapy, to achieve a more personalized and cost-effective treatment.

We take advantage of artificial intelligence to learn directly from large amounts of raw image data eliminating the biases introduced by manual annotations and feature engineering thus reducing human bias. We compare the results of machine learning-based analysis with that of visual assessment of colorectal and breast cancer performed by an expert as well as with clinical parameters. Also, we visualize hot-spots on tissue images that are considered by the machine learning model as the most predictive of the desired target output. This allows to better understand what drives the neural network decisions. Our studies reveal novel imaging biomarkers for colorectal and breast cancer and leads to improved diagnostics and stratification of cancer patients. Moreover, computer-aided analysis improves standardization and reproducibility within image-based cancer diagnostics.

Machine learning and subcellular localization

Marco Salvatore, Stockholm University, Science for Life Laboratory Stockholm

Knowledge of the correct protein subcellular localization is necessary for understanding the function of a protein. Recently large-scale experimental studies have provided new data that can be used to improve the accuracy of subcellular predictions.
Computational prediction of protein subcellular localization using sequence-based information was introduced 30 years ago by the study of signal peptides. The first method able to predict multiple localizations, PSORT, was developed 25 years ago and later many other methods have been developed. Today, prediction methods can be specialized for a specific localization, either for a few or for a wide range of localizations.
The most successful subcellular predictors use a combination of features that can roughly be classified to be either sequence- or annotation-based. These features are used as inputs to some machine learning method.
Machine Learning and more recently Deep Learning have been applied to various biological domains, leading to a number of impressive advances.
This presentation will highlight some of the important past and current developments on the interface between machine learning and biology—with a particular focus on Random Forest and Deep Neural Network that can be used to predict sorting signals as well as the final destination of a protein.

Reference:
SubCons: a new ensemble method for improved human subcellular localization predictions. Salvatore, M. Et al. Bioinformatics, 2017. (33) 16, 2464-2470.
The SubCons web-server: A user friendly web interface for state-of-the-art subcellular localization prediction. Salvatore, M., Shu, N., Elofsson, A. Protein Science. 2017 Sep 13. doi: 10.1002/pro.3297

Fast and Scalable Estimation of Uncertainty using Bayesian Deep Learning

Emtiyaz Khan, RIKEN Center for Advanced Intelligence Project

Uncertainty estimation is essential to design robust and reliable systems, but this usually requires more effort to implement and execute compared to maximum-likelihood methods. In this talk, I will summarize some of our recent work that enables fast and scalable estimation of uncertainty using deep models, such as Bayesian neural network. The main feature of our method is that they are extremely easy to implement within existing deep-learning softwares. I will also summarize some of the current challenges faced by the Bayesian deep-learning community and how real-world applications can be useful for our research.

Applying Deep Learning for Life Science

Masatoshi Hamanaka, RIKEN Center for Advanced Intelligence Project

We present three kinds of applications for deep learning. The first application is for drug discovery. Computational prediction of compound-protein interactions (CPIs) is of great importance for drug design as the first step in in-silico screening. The results of cross-validation show that the accuracy of our system based on deep learning reaches up to 98.2% (ƒÐ< 0.01) with 4 million CPIs. The second application is for music analysis. For over 15 years, we have been implementing music analyzers on the basis of the generative theory of tonal music (GTTM), which was proposed by Lerdahl and Jackendoff. The experimental results demonstrated that our system based on deep learning outperformed the previous analyzers for a GTTM in F-measure for generating the grouping and metrical structures. The last application is for estimating the position of drone. We propose a method for estimating the flying area of drone by using a 3D map created by deep learning. Our method could estimate the flight area with 92.2% accuracy in a simulation and 98.4% accuracy in a field experiment.

Mapping of expression profiles in tissue

Simone Codeluppi, Department of Medical Biochemistry and Biophysics, Karolinska Institute

Current advances in spatial transcriptomics enable the discovery, characterization and mapping of cell states in both healthy and pathological conditions. However, the spatial definition of an expression profile requires the identification of the tissue region occupied by a cell or part of it. In order to identify cell regions we developed a multimodal approach based on clustering of specific markers signal and segmentation of image metadata such as nuclei and total RNA staining using a mask region based convolutional neural network.

Our current effort is focused onto adding more advanced and efficient algorithms for cell segmentation in human and mouse brain tissue sections together with decreasing the computing time required for RNA molecules detection. In addition, we are actively working on integrating some of the newly developed and optimized functions into pysmFISH, the open source analysis framework developed for smFISH data analysis.

Genomics applications of recent advances in deep learning

Mikael Huss, Peltarion, Department of Learning, Informatics, Management and Ethics, Karolinska Institutet

Neural networks have been used for classification and prediction in various application areas, including biological research, for decades. However, in the past few years, the field (now rebranded as deep learning) has evolved rapidly and new techniques and concepts have emerged, enabling deep learning practitioners to reach unprecedented results in tasks like image classification, object recognition, and machine translation. These techniques tend to work well on data types with intrinsic local correlations, like images, and/or when there is a massive amount of training data available (like with language). But some of the newer techniques are starting to be applied in genomics as well. I will talk about some of those, such as deep convolutional models, attention mechanisms and deep generative models.

Learning machines for the life sciences

Magnus Boman, EECS, Royal Institute of Technology

A learning machine can be defined as an autonomous self-regulating open reasoning machine system that actively learns in a mostly unsupervised and decentralized manner, over multiple domains. Its purpose is usually not to replace, but to augment, humans in its vicinity. Unlike machine learning, the term learning machine thus has a relatively crisp definition, which stresses the autonomy of the system. How such autonomous software can prove useful to the life sciences will be illustrated by research questions and engineering solutions from an ongoing project on Internet psychiatry.