Introduction to AI in Life Sciences

Mikael Huss

(1) Department of Learning, Informatics, Management and Ethics (LIME), Karolinska Institutet, Stockholm, Sweden.
(2) Codon Consulting AB, Stockholm, Sweden.

AI has had a new resurgence in the past few years, mostly due to the recent burst of state-of-the-art results from various deep learning models (although some voices of caution are warning against relying too much on this particular type of AI). In the life sciences, these techniques are being taken up from the larger AI community with a relatively short lag time. Some early promising results have been obtained in medical imaging, drug discovery, and fundamental research such as protein folding and genomics. At the same time, deep learning techniques are not always trivially applicable to problems in life science. In this talk, I will try to put into context the recent advances in AI in life science by relating them to the AI field as a whole. I will present some example applications and discuss where there are problems that result from data properties or differences in culture between the larger AI community and the life science AI community. Further, I will briefly discuss resource sharing and possible areas in biomedicine where deep learning AI might hold promise.
Epistasis analysis from genome-scale population-wide sequence data

Erik Aurell

KTH-Royal Institute of Technology, AlbaNova University Centre, SE-106 91 Stockholm, Sweden

Genetic co-variation of host and pathogen is known to play a role in infectious diseases (cf. Lees et al, “Joint sequencing of human and pathogen genomes reveals the genetics of pneumococcal meningitis”, Nature Comm vol 10:2176 (2019)). Co-variation of alleles at separate loci in bacterial pathogens have been interpreted as signs of epistasis, or synergistic fitness effects (Skwark et al, “Interacting networks of resistance, virulence and core machinery genes identified by genome-wide epistasis analysis”, PLoS Genet 13(2): e1006508 (2017)). If well characterized, such effects may offer possibilities for the discovery, design and optimization of new combination treatments.

This presentation is based on Skwark et al 2017 (op cit) and a later methodological development (Chen-Yi Gao et al “DCA for genome-wide epistasis analysis: the statistical genetics perspective” Phys Biol 16 026002 (2019)) which both relied on a set of around 3,000 whole genome sequences of the human pathogen Streptococcus pneumoniae, obtained from patient samples in the Maela refugee camp in Thailand (Chewapreecha et al “Dense genomic sampling identifies highways of pneumococcal recombination” Nature Genetics 46: 305 (2014)). With increasing population-wide genomic data this type of analysis will become feasible for many more human pathogens (bacterial or other), and eventually for human and human-microbiome data.

I will discuss the computational challenges in such analysis and how speed-up can be obtained by pre-filtering. I will also present conjectures for the limits of applicability of such analysis.
In recent years, with the development of measurement technology and the spread of artificial intelligence technology, data science has been rapidly introduced to medical/life science research. In medical data analysis, it is often difficult to directly apply a basic biological model due to differences in time scale and hierarchy. In such case, data-driven approach is essential to create individualized models based on comprehensive observation of the target disease without specific hypotheses. Among the data-driven approaches, machine learning technique is often used to extract potential patterns of diseases considering different types of variables and dependencies among variables, and to perform accurate predictions. Supervised learning can predict the characteristics of diseases with high accuracy compared to the conventional statistical method. In addition, unsupervised learning enabled us to discover patterns/clusters that have not been noticed even by clinicians. In this talk, we will introduce how to use machine learning as a means to support human knowledge discovery and hypothesis formation, and discuss new framework of medical research for the next generation of medicine.
Trans-omics: integration of multiple omic data on the basis of reaction kinetics

Katsuyuki Yugi

YCI Laboratory for Trans-omics, RIKEN Center for Integrative Medical Sciences (IMS)

Metabolic homeostasis is realized by global networks of molecular interactions across multiple ‘omic’ layers such as genome, transcriptome, proteome, and metabolome. We have proposed ‘trans-omics’, a methodology to reconstruct large-scale networks that span across the multiple omic layers on the basis of time-series omic data and reaction kinetics [1,2]. Here we show an application of trans-omics to reconstruction of a metabolic regulatory network that underlies acute insulin action in rat hepatoma FAO cells [3,4]. The network integrates phosphoprotome and metabolome data, and involves 13 protein kinases, 26 phosphorylated metabolic enzymes, and 35 allosteric effectors, resulting in quantitative changes in 44 metabolites. Kinetic modeling analysis predicted selective control of a subnetwork around phosphofructokinase by specific phosphorylation and allosteric regulation. Thus, we provide an unbiased method that reconstructs the trans-omic network from phosphoproteome and metabolome data, which will be applicable to other cellular responses. Furthermore, possible future extension of trans-omics particularly incorporation of machine learning technologies will be discussed.

References

Despite the fundamental nature of transcription acting as an interface between the DNA and the protein synthesis machinery, it is still unclear to what extent the variation of transcript copy numbers is encoded in the DNA and what are the key determinants of this variation. Through the analysis of over 20,000 RNA-Seq experiments using deep neural networks, we show the gene expression levels can be predicted with less than 20% error based merely on information encoded in the genes and their surrounding sequence from all kingdoms of life. We identify interactions between specific regulatory sequence elements crucial for transcription regulation in *Saccharomyces cerevisiae*, including those in promoters, untranslated regions and termination sequences, that are fundamental determinants of transcript levels. This is reflected in the coevolution of gene coding sequences and their flanking regulatory regions, demonstrated by measuring mutation rates between genes regions in 90% of orthologous genes from 13 yeast species. Our advances to the understanding of the transcriptional regulatory code suggest that protein coding regions with adjacent flanking elements are an evolved transcriptional regulatory unit yielding a mechanism by which whole gene sequences with prespecified expression patterns can be designed. We anticipate that this AI-driven approach may have dramatic consequences to the fields of biotechnology and biomedicine greatly advancing our understanding of cellular processes but also opening doors for new opportunities in synthetic biology and eventually leading lead to better therapy design and other applications to medicine.
Obesity is a complex disease involving an excessive amount of body fat. In 2016, more than 1.9 billion adults are obese or overweight, leading to a rise in related metabolic complications, including insulin resistance, type 2 diabetes, cardiovascular and liver disease. The epidemic of obesity urges a better understanding of the mechanisms of white adipose tissue expansion and its linkage with metabolic disorders, which is accompanied by chronic low-grade inflammation. Chronic inflammation often presents in an acute adaptive inflammatory response to an initial trigger, followed by a long-term maladaptive phase that leads to complications.

While the detailed mechanism is still not clear, it is known that both adipocytes and adipose tissue immune cells can regulate inflammatory status by secreting cytokines (adipokines in case of adipocytes). Besides, numerous signaling pathways have been described at the interface between inflammation and metabolism. However, so far therapies using inflammation as the target have modest effects. One reason for this is the lack of biomarker evidence of the anti-inflammatory effects of the drugs. Thus, it is important to obtain information of changes in protein expression of different cell populations in adipose tissue.

In our study, we applied next-generation proteomics technologies to precisely quantify expression of thousands of proteins in various cell populations, e.g. adipocytes, macrophages, TCR+ cells in white adipose tissue from mice under high-fat and control diet. In addition, samples from short and long-term high-fat diet were compared. The generated data sets allow separation of cells types in an unsupervised data-driven approach. With these comprehensive data sets, we aim to identify key molecules that determine inflammatory status of adipose tissue and metabolic homeostasis of the body. We expect that artificial intelligence will be a powerful tool to reach this goal and we look forward to discussing this at the symposium.
Cell heterogeneity in a population
-what can we investigate by deep single cell RNA-sequencing?

Hideyuki Yoshida

YCI Laboratory for Immunological Transcriptomics,
RIKEN Center for Integrative Medical Sciences (IMS)

T cells (or T lymphocytes) are one of the main components of the immune system, of which hallmark is the presence of a T-cell receptor (TCR) on the cell surface which recognize pathogens. By recombination, random insertion, deletion and substitution, the set of genes encoding the TCR have the potential to shape as many as $1 \times 10^{15}$ different TCRs. While this tremendous diversity is crucial for T cells to recognize wide-range of offending pathogens, it can produce self-reactive T cells by chance. As these self-reactive T cells have the potential to damage the healthy tissue, they have to be eliminated and/or regulated.

Negative selection is the process where auto-reactive T cells are removed during the T cell differentiation in a thymus. A bunch of peripheral tissue specific antigens (PTA), such as insulin and mucin which are specific for pancreas and stomach respectively, are expressed in thymic medullary epithelial cells (TECs) and the developing T cells are eliminated if they respond to these PTA.

The transcriptional factor AIRE (autoimmune regulator) has been shown to function inducing the expressions of thousands of PTA genes in TECs. The prevalent view is that the expressions of PTA genes are simply promoted as AIRE is induced along the differentiation of TECs. However, our single cell RNA sequencing data revealed another layer of gene regulation which enhances the heterogeneity of expressed PTAs in individual TECs in coordination with AIRE.

I would like to discuss about the possible analysis to examine the cell heterogeneity and its regulation utilizing single cell RNA-sequencing data.
Acute myeloid leukemia (AML) has a poor prognosis in both adults and children, with a long-term survival of only 25% and 60% respectively. AML is associated with perturbed epigenetic regulation, with early mutations in and chromosomal translocations of different epigenetic regulators. This indicates that epigenetic mechanisms may play an essential role in the development and AML and are potentially very potent drug targets. A network of epigenetic factors regulates DNA methylation, posttranslational histone modifications and chromatin structure, and relays information to the transcriptional program that dictates hematopoietic cell fate and differentiation. We have previous demonstrated the importance of epigenetic mechanisms in hematopoietic differentiation and how specific mutations in AML leads to a specific DNA methylation pattern. Especially we have showed that epigenetic regulation of enhancer activity is crucial for normal myelopoiesis and AML (Rönnerblad et al. Blood 2014, Qu et al. Blood 2017). We demonstrated that the generation of leukemic-specific gene expression involves an interplay of combinatorial epigenetic mechanisms at specific enhancer elements with their cognate promoters (Qu et al. Blood 2017). In addition, we have showed that DNA methylation pattern can be used to predict prognosis (Deneberg et al Blood 2010). Based on those results we hypothesize that different chromatin marks and higher order chromatin structures provide a pathological pin code that enables a much earlier and more accurate staging of the disease.
Behavior analysis is a sensitive readout of brain function. We previously reported that Pdcd1−/− mice, which lack the inhibitory receptor PD-1, showed altered behavior manifested with enhanced anxiety and exacerbated fear responses. We found that augmented T cells responses in Pdcd1−/− mice led to the depletion of amino acids from the serum due to their trapping in the activated lymphocytes. Tryptophan and tyrosine are essential amino acids for the synthesis of two important brain monoamine neurotransmitters, serotonin and dopamine, respectively. We found that the serum depletion of tryptophan and tyrosine led to reduced levels of serotonin and dopamine in the brain, resulting in changes of emotional behaviors of Pdcd1−/− mice. We are currently trying to normalize the behavior of Pdcd1−/− mice, and found that tryptophan supplementation diet has only partial effect. Our data also suggest that environmental factors, such as gut microbiota, might be involved in abnormal behavior of Pdcd1−/− mice. Indeed, tryptophan supplementation diet significantly changed the composition of gut microbiota, but the impacts of such environmental shifts on mice behavior is yet difficult to investigate. Thus, a wide array of behavioral analyses and host-microbiota co-metabolism are required to determine the impact of both genetic and environmental factors on behavior. However, the complexity, variability and the amount of data sets related to behavioral analyses, metabolome and metagenome precludes us to perform large-scale screenings and unbiased analyses. Here we propose to apply artificial intelligence for mouse behavioral analysis. The automated pattern recognition of mouse behavior will allow us to efficiently evaluate the effect of therapeutic interventions and other environmental factors strongly impacting on the brain functions.
Applications of natural language processing methods using routine healthcare data

Sumithra Velupillai

Institute of Psychiatry, Psychology & Neuroscience, King’s College London/NIHR Maudsley Biomedical Research Centre

Routine healthcare data such as electronic health records (EHRs) are an incredibly rich resource for secondary data analysis, and appropriate utilization of these could have a dramatic impact on healthcare research and delivery. A large proportion of the documentation in EHRs is in written text form. In mental health care, this proportion is larger than in other clinical domains, as the most important features of mental health care do not lend themselves to structured fields (e.g. mental health symptoms, determining treatment initiation, outcome evaluation). To enable large-scale analysis of this information, Natural Language Processing (NLP) applications are needed. In this talk, I will describe past and ongoing projects where NLP methods have been applied on mental health records from a large secondary mental health care provider in South London, UK. I will give some examples of use-cases and approaches, and discuss where NLP and combinations of ML approaches have been successfully applied for particular clinical use-cases.
Learn genomics from AI for interpreting GWAS findings

Masaru Koido¹, Chung-Chau Hon¹, Hideya Kawaji¹, Yasuhiro Murakawa¹, Yoichiro Kamatani¹², Piero Carninci¹, Chikashi Terao¹³

¹RIKEN Center for Integrative Medical Sciences (IMS), Japan
²Graduate School of Frontier Sciences, the University of Tokyo, Japan
³Clinical Research Center, Shizuoka General Hospital, Japan

Genome-wide association studies (GWAS) have found many non-coding single nucleotide variants (SNVs) associated with human complex traits. So far, catalogs of expression quantitative trait loci (eQTL) have provided plausible and novel interpretations of GWAS findings in a cell-type dependent manner; however, eQTL for lowly expressed non-coding RNAs including enhancer RNAs is poorly understood due to the lack of large-scale studies measuring them. Here we show that, by using only genome sequence surrounding transcriptional start site (TSS) as input data, machine learning (ML) can predict cell-type specific expressions of promoters and enhancers (as a binary “on-off” switch) in Cap Analysis of Gene Expression (CAGE). We designed ML models by combining deep convolutional neural networks (from sequence to epigenetic features, using publicly available pre-trained models) and binary classifier using gradient boosting trees (from epigenetic features to transcription probabilities). We leveraged the transcriptomes of 347 major human cell-types and tissues from the Functional Annotation Of the Mammalian genome (FANTOM5). We trained the ML models with the use of the autosomal promoter and enhancer expressions profiled by CAGE except for chromosome 8 and tested the accuracy using those on chromosome 8. As a result, ML models using ±100-kb sequences from TSS showed high accuracy to predict transcription (mean (± SD) area under the ROC curve (AUROC) = 0.83 ± 0.22 (n = 347)). Notably, 295, 125, and 152 among the 347 models achieved AUROC >0.7 for divergent-lncRNA, intergenic-lncRNA, and enhancer peaks, respectively. Finally, we calculated effects of SNVs on proximal promoter/enhancer transcription and found high concordance between the predicted effects and known eQTL effects. Collectively, these findings indicate the potential use of our ML models to efficiently investigate genetic architecture for non-coding transcripts without additional population studies.
How are the functions of Zinc Finger domains in Transcription Factors regulated?

Ichiro Taniuchi
Laboratory for Transcriptional Regulation, RIKEN Center for Integrative Medical Sciences (IMS)

A Zinc finger (ZF) domain in proteins has been characterized to function as a platform to interact with nucleic acids as well as proteins. In mammals, there are more than 500 proteins that possess single or multiple ZF domain(s), which include transcription factors. ZF transcription factors can be categorized into sub-families according to their primary structure. For instance, Ikaros-zinc finger (IKZF) family proteins have a common structure, four N-terminal ZFs for DNA binding and two C-terminal ZFs for dimerization, while Bcl11 family has two and three C2H2-type ZFs at the middle and C-terminal, respectively. Recently, we isolated causal variant for human B-cell deficiency in \( \text{IKZF3} \) gene, which results in Glycine to Arginine replacement at position 159 (\( \text{IKZF3}^{G159R} \)) within the second ZF domain in its encoding AIOLOS protein. Other variants in \( \text{BCL11b} \) gene that causes Asparagine to Lysine at position 441 (\( \text{Bcl11b}^{N441K} \)) within the middle ZF or at position 807 (\( \text{Bcl11b}^{N807K} \)) within the C-terminal ZF were reported to be causal for T-cell deficiency. We have generated mouse models for these three human variants and revealed that a single amino acid replacement mutation in such ZF domains interfere with functions of heterodimer partner: IKAROS and Bcl11a function is hijacked by Aiolos\( ^{G158R} \) and Bcl11b\( ^{N441K} \), respectively. Thus, we propose that heteromeric interference is a novel mechanism of autosomal dominance that causes human disease by impairing protein functions via mutation of its heteromeric partner. Currently, understanding of such molecular pathogenesis requires experimental procedures. However, accumulating knowledges that links amino acid sequences with 3D protein structures would make it possible to predict whether and how a single amino acid change alters protein functions. I would like to discuss whether and how AI can contribute to predict proteins structures and function from amino acid sequences.
Less than 2% of the human genome encodes proteins. Before reference human genomes were mapped, and in the first part of the post-genomic period, biologists focused on building a comprehensive understanding of how this small fraction of the genome is expressed. Perhaps unsurprisingly, the schema of genome function build on this understanding is incomplete, explaining only a portion of the expected genetic contribution to health and disease. Moving forward in the post-genomic era, understanding the remaining 98% of the human genome will take a higher priority. 45% of the human genome is recognizably derived from sequences which first arrived as a consequence of horizontal transfer (e.g. sequences derived from transposable elements and viruses). I will present evidence that our ability to recognize such sequences in the human genome is limited, and thus 45% is almost certainly an underestimate. Many examples of such sequences contributing to genome function have recently been reported. I will summarize these briefly and explain my group’s studies on the function of sequences horizontally-acquired from viruses. I will report some initial efforts to use artificial intelligence to identify such sequences, and speculate on the benefits of improved functional annotation of horizontally-acquired sequences in human genomes.
Artificial Intelligence aided Deep Phenotyping

Kazuhiro Sakurada
RIKEN Medical Sciences Innovation Hub Program (MIH)

Organisms are nonlinear dynamic systems (NLDSs). Understanding NLDSs is challenging in various scientific fields. In modern Biology, the purpose of research is to identify causal mechanism to describe common feature of organism including diseases. However, a linear approximation and a spatial characterization have limits to describe tangible nonlinear dynamic behavior of nature. Employing high quality big data and AI aided interpretation would be the way to overcome this problem. To understand an individual living thing as becoming not as being, physical space should be described as spatial-temporal features. Although genome information is necessary, reductionist approaches such as genetics or genotyping are not sufficient.

Life course data is extremely high dimension and discrete. In addition, nonlinear interaction, self-organization and emergent features of organism should be considered at an individual level. To overcome epistemological limitation of mechanism, I have introduced the concept of deep clinical phenotyping to characterize diseases by comprehensive phenotypes using “information geometry”. By considering points in multidimensional space as informational representation, we can manipulate probability distribution exactly as a new flavor with geometric interpretation.

Homeostasis has been described by feedback model in gene expression network. This model cannot describe collective behaviors of organism. A large ensemble of dynamic systems can behave collectively by synchronization. Rhythm generation and synchronization are the universal principle of non-linear systems. Organisms have the property of non-linear oscillator at all scale. Complex spatiotemporal patterns of organisms are generated by synchronization and desynchronization of discrete entities. By introducing this “first principle”, we can redefine health and diseases.

Current progress of AI will emerge new synthesis of Biology based on NLDSs.
Deep analysis of images from ultra-microstructural microscopy using machine learning

Yasuhiro Murakawa, Norio Kobayashi, Satoshi Kume

1. RIKEN-IFOM Joint Laboratory for Cancer Genomics, RIKEN Center for Integrative Medical Sciences (IMS)
2. Computational Engineering Applications Unit, Head Office for Information Systems and Cybersecurity
3. Laboratory for Cellular Function Imaging, RIKEN Center for Biosystems Dynamics Research (BDR)

Biological tissues have fine and complex morphological characteristics, and electron microscopy (EM) is indispensable for observing these ultra-microstructures. We have developed a large-scale EM system that comprehensively and quantitatively analyze 3D whole images of the tissues using serial and automatic imaging of tissue specimens by the latest scanning electron microscope. Such comprehensive imaging can be regarded as “Morphomics”. Furthermore, we have shared images of the tissue microstructures and their metadata in the RIKEN Microstructural Imaging Metadatabase (http://clst.multimodal.riken.jp/RIKENImageDB/) and determined a common procedure for sharing, integration, and analysis of data.

Here, we introduced the results of image data analysis using machine learning (ML). We used ML to segment the images, allowing us to extract and quantify the morphologies of the cell nuclei and mitochondria with high accuracy. Combining large-scale imaging and ML enables the comprehensive morphological analyses of the tissue microstructures under various pathophysiological conditions. We also discuss the possibility of applying this technology to the histopathological analysis of human specimens.
SciLifeLab is a national Swedish infrastructure for molecular life sciences. In 2018, more than 1,200 researchers were supported by 40 different SciLifeLab service platforms. Data management has a key role, as data is one of the main deliverables from the infrastructure. The need for improved access to high quality IT services and support for research data management has increased rapidly because of the data intensive nature of state-of-the-art life science research, and the application of new computational analysis methods for life science, such as new artificial intelligence based methods. Moreover, requirements on open access to both raw data and results from funders and journals, together with the scientific benefits that come from improved data sharing, has led to an increased need for support to researchers.

To address these new requirements for support to data intensive research, SciLifeLab has established a Data Centre, which primarily supports the service platforms that in turn support the users – the researchers. By setting its focus on the data producing platforms, the Data Centre aims at ensuring that all produced data from the national infrastructure has the highest quality, meeting the FAIR criteria as far as possible, and that the platforms have access to state-of-the-art compute and storage services, data management services, and data resources. By ensuring good practices for data management already at data production stage, we facilitate the best possible data handling further downstream in the research workflow, with the goal to maximize the scientific impact of the produced data.