DDLS Annual Conference 2023

November 15-16, 2023

#	POSTER
1	**Integrative Data Processing Pipeline for PROseq Uncovers Mechanisms of Transcription across Functional Genomic Regions** Serhat Aktay, KTH Royal Institute of Technology
2	**FunCoup 5: Functional Association Networks in All Domains of Life, Supporting Directed Links and Tissue-Specificity** Davide Buzzao, Stockholm University
3	This abstract has been withdrawn
4	**Understanding the microbial diversity in the healthy vagina** Gabriella Edfeldt, Karolinska Institutet
5	**REFSTRA (REference-Free Single-cell Transcriptome Autotyping) for deconvolution of paired Spatial Transcriptomics data** Alper Eroglu, SciLifeLab
6	**CoolBeans: user-friendly workflow for identifying multi-metabolite signatures of exposure.** Núria Estanyol-Torres, Chalmers University of Technology
7	**Processing spatial networks through UMI** David Fernandez Bonet, KTH Royal Institute of Technology
8	**Revealing the RBP regulome in hepatocellular carcinoma via consensus GRN inference** Mateusz Garbulowski, SciLifeLab/ Stockholm University
9	**Utilizing large-scale genomic data to explore variable disease penetrance** Sanna Gudmundsson, SciLifeLab/KTH
10	**U-FISH: a universal deep learning approach for accurate FISH spot detection across diverse datasets** Weize Xu, AI Cell Lab, KTH
11	**Model based DNA basecallers** Joakim Jaldén, KTH Royal Institute of Technology
12	**Detecting resin wood in Scots pine with laboratory-bespoke X-ray computed tomography imaging** Sheng Joevenller, Luleå University of Technology
13	**Inferring dissipation from cell-membrane fluctuations** Sreekanth K Manikandan, Department of Chemistry, Stanford University
14	**An Advanced Deep Learning Pipeline for Genome-Wide Imaging Screen Analysis Uncovering Cell Death Regulators** Salma Kazemi Rashed, Lund University
15	**Beam search decoder for enhancing sequence decoding speed in single-molecule peptide sequencing data** Javier Kipen, KTH Royal Institute of Technology
16	**BIIF - Support on BioImage Analysis in Sweden** Anna Klemm, SciLifeLab
17	**Single-cell RNA sequencing-based program-polygenic risk scores associated with pancreatic cancer risks in the UK Biobank cohort** Yelin Zhao, Karolinska Institutet
18	**Deep Learning with Big Data for Genetic Epidemiology** Max Kovalenko, Uppsala University

#	POSTER
19	**Spatial transcriptomics and GWAS data identify putative causal tissue structures for complex traits** Linda Kvastad, SciLifeLab/KTH
20	**Bacterial vaginosis: Understanding the effect of antibiotic-free treatment on the vaginal microbiome** Emilia Lahtinen, Karolinska Institutet
21	**SciCommander Track provenance of any shell command** Samuel Lampa, Karolinska University Hospital
22	**Kidney automatic segmentation and cystic renal occupying lesions classification system based on the Deep Learning.** Hui Li, Chalmers University of Technology
23	**Metagenomics of insect bulkDNA for population genetics** Samantha López Clinton, Swedish Museum of Natural History
24	**Multimodal signal recordings in neuroscience: modeling strategies** Melisa Maidana Capitan, Linköping University
25	**National Genomics Infrastructure (NGI) Next Generation Sequencing and Genotyping for Swedish Research** Tom Martin, National Genomics Infrastructure (NGI)
26	**Multi-metabolic signature of controlled modification of dietary carbohydrate quality** Cecilia Martinez Escobedo, Chalmers University of Technology
27	**From Genes to Causal Maps: A Benchmark for Gene Regulatory Network Inference** Mariia Minaeva, KTH Royal Institute of Technology
28	**Identification of metabolomic networks linked with incident heart failure.** Jakub Morze, SGMK Copernicus University / Chalmers University of Technology
29	**Predicting Protein-Protein Interactions using Machine Learning** Sarah Narrowe Danielsson, SciLifeLab/ Stockholm University
30	**Recalibrating differential gene expression analysis by variance in gene dosage** Philipp Rentzsch, SciLifeLab
31	**Mapping the genetic architecture of sarcoidosis across populations** Natalia Rivera, Karolinska Institutet
32	**The Swedish Childhood Tumor Biobank** Johanna Sandgren, Karolinska Institutet
33	**Targeted Proteomics of Blood Plasma from the hPOP Cohort** Thanadol Sutantiwanichkul, KTH Royal Institute of Technology
34	**Barcode-free prediction of cell lineages from scRNA-seq datasets** Marcel Tarbier, SciLifeLab/KI
35	**Deep Learning for Time Series Classification of Parkinsons Disease Eye Tracking Data** Gonzalo Uribarri, SciLifeLab/KTH
36	**Benchmarking orthologous clustering programs for proteins a case study in Pseudomonas aeruginosa** Virág Varga, Chalmers University of Technology

#	POSTER
37	**Predicting plastic-degrading potential of the Baltic Sea: a data-driven approach with taxonomic information** Máté Vass, Chalmers University of Technology
38	**Predicting Preterm Birth Using Machine Learning** Nicole Wagner, Karolinska Institutet
39	**Navigating the Toolbox: A Comparative Analysis of Metagenomic Tools for Taxonomic and Resistance Gene Identification** Marcus Wenne, Chalmers University of Technology
40	**SciLifeLab Data Platform** Liane Hughes, SciLifeLab/UU
41	**SciLifeLab Serve - enabling sharing of machine learning models and applications** Arnold Kochari, SciLifeLab
42	**Services and support from SciLifeLab Data Centre** Katarina Öjefors Stark, SciLifeLab Data Centre

1. Integrative Data Processing Pipeline for PROseq Uncovers Mechanisms of Transcription across Functional Genomic Regions

Serhat Aktay, KTH (serhat.aktay@scilifelab.se)

High-throughput sequencing has transformed genomics, and Precision Run-On Sequencing (PRO-seq) is a method for mapping RNA Polymerase positions at nucleotide precision. Here we present a versatile PRO-seq data pipeline offering data processing, mapping, detection of divergent transcription, and mapping functional genomic regions. It enables the identification of transcription start sites, promoters, divergent transcription sites, gene bodies, termination windows, and differentially expressed genes. The pipeline's efficacy is demonstrated with PROseq data from five species (fly, mouse, dog, human, and thale cress), unveiling insights into heat-induced transcriptional reprogramming across organisms.

Keywords:
PRO-seq, computational pipeline, functional genomic regions

2. FunCoup 5: Functional Association Networks in All Domains of Life, Supporting Directed Links and Tissue-Specificity

Davide Buzzao, Stockholm University (DBB) (davide.buzzao@scilifelab.se)
Emma Persson[1], Miguel Castresana-Aguirre[2], Davide Buzzao1, Dimitri Guala[1], Erik L.L. Sonnhammer[1]

[1]Department of Biochemistry and Biophysics, Stockholm University, Science for Life Laboratory, Box 1031, 171 21 Solna, Sweden
[2]K7 Department of Oncology-Pathology, Karolinska Institute, 171 77 Stockholm, Sweden

FunCoup (https://funcoup.sbc.su.se) is one of the most comprehensive functional association networks of genes/proteins available. Functional associations are inferred by integrating evidence using a redundancy-weighted naïve Bayesian approach. FunCoups high coverage comes from using eleven different types of evidence, and extensive transfer of information between species. Since the latest update of the database, the availability of source data has improved, and user expectations have grown. To meet these requirements, we have made a new release of FunCoup with updated source data and improved functionality, now including 22 species from all domains of life. In this release, directed regulatory links can be visualized for the human interactome, and subnetworks can be filtered for genes expressed in specific tissues. FunCoup 5 includes the SARS-CoV-2 proteome, allowing users to visualize and analyze interactions between SARS-CoV-2 and human proteins. This new release of FunCoup constitutes a major advance for the users, with updated sources, new species and improved functionality for analysis of the networks.

Keywords:
Functional Association Networks; Network Biology; Systems Biology; Benchmark; SARS-CoV-2

3. This abstract has been withdrawn

4. Understanding the microbial diversity in the healthy vagina

Gabriella Edfeldt, Karolinska Institutet (gabriella.edfeldt@ki.se)
Gabriella Edfeldt (presenter)[1], Johanna Norenhag[2], Emilia Lahtinen[1], Emma Fransson[1], Juan Du[1], Luisa W. Hugerth[1], Marica Hamsten[1], Alexandra Pennhag[1], Maike Seifert[1], Ina Schuppe Koistinen[1], Matts Olovsson[2], Lars Engstrand[1]
[1] Centre for Translational Microbiome Research, Department of Microbiology, Tumor and Cell Biology (MTC), Karolinska Institutet, Stockholm, Sweden
[2] Department of Womens and Childrens Health, Uppsala University, Uppsala, Sweden

The vaginal microbiome plays a major role in female and reproductive health. Healthy vaginal microbiome has been frequently shown to protect women from various infections, and to reduce the risks of adverse pregnancy outcomes. However, compared to the gut microbiome, the microbiome of the reproductive tract is still relatively understudied. The Vaginal Microbiome in Gynecological Health (VaMiGyn) is the largest metagenomic study on the vaginal microbiome of healthy Swedish women and aims to define the normal vaginal microbiota composition based on the samples collected through the national cervical cell screening program. 1460 vaginal samples with shotgun metagenomic data are associated with health and sociodemographic questionnaires to identify potential microbial patterns between women and to define the taxonomic and functional composition of the healthy vaginal microbiome. The poster will present the cohort and data on the factors that contribute to the composition of the vaginal microbiome.

Keywords:
bioinformatics, metagenomics, microbiome analysis

5. REFSTRA (REference-Free Single-cell Transcriptome Autotyping) for deconvolution of paired Spatial Transcriptomics data

Alper Eroglu, Scilifelab (alper.eroglu@scilifelab.se)

6. CoolBeans: user-friendly workflow for identifying multi-metabolite signatures of exposure.

Núria Estanyol-Torres, Chalmers University of Technology (nuria.estanyol@chalmers.se)
Núria Estanyol-Torres[1], Cecilia Martinez Escobedo[1], Daniela A. Garcia-Soriano[2], Luke W. Johnston[3], Clemens Wittenbecher[1]
[1]Division of Food and Nutrition Science, Department of Life Sciences. Chalmers University of Technology. SciLifeLab. SE-412 96 Gothenburg, Sweden
[2] E-Commons. Chalmers University of Technology. SE-412 96 Gothenburg, Sweden
[3] Department of Clinical Medicine, Aarhus University. Steno Diabetes Center Aarhus. Aarhus, Denmark.

High-throughput metabolomics approaches in human studies provide large datasets with complex correlation structures that reflect genetic, phenotypical, lifestyle and environmental influences. At the same time, metabolomics data are strongly predictive of multiple disease outcomes and the multi-metabolite patterns (aka signatures) have proved to best capture the exposure to complex lifestyles factors.
Here, we introduce CoolBeans, a tool that leverages metabolomics data for multi-metabolite biomarker assessment. It comprises of a workflow that integrates data processing, feature selection with multiple test correction, and state-of-the-art machine learning techniques to construct the multi-metabolite signatures while accounting for confounder adjustment. Additionally, the pipeline adheres to best practices in coding to promote reproducibility of results. We are building CoolBeans as an open and user-friendly tool, available to researchers as both an R package and a web-based interface, fostering collaboration and further research within the metabolomics and precision health field.

Keywords:
Metabolomics, Multi-metabolite signatures, R pipeline, machine learning, precision health

7. Processing spatial networks through UMI

David Fernandez Bonet, KTH Royal Institute of Technology (dfb@kth.se)
Fernandez David, Lang Shuai*, Dahlberg Simon*, Benson Erik*, Hoffecker Ian**
KTH Royal Institute of Technology, SciLifeLab, Gene Technology, Molecular Programming Group

Sequencing-based DNA Microscopy is an emerging field in spatial biology. Unlike traditional microscopy, it does not rely on optical lenses to create an image. This allows for the analysis of molecular structures in a way that is not limited by the diffraction limit of light, while also promising high throughput by means of Next-Generation Sequencing. Moreover, it does not rely on fluorescent labels or other markers that can introduce limitations on the number of molecule types that can be identified.
However, extracting and denoising UMIs has to be as accurate as possible in order to not introduce artifacts during the process of image reconstruction. PCR and sequencing errors complicate the task by introducing noise, and the larger the dataset the more difficult denoising can become.
We propose to solve the problem by crafting an efficient pipeline involving optimal UMI extraction, reliablity filters such as high-copy numbers and low edit-distances and the use of spatial information to detect fuse UMIs through graph representation learning algorithms.

Keywords:
Imaging, Biological Networks, UMI, Graph Representation Learning

8. Revealing the RBP regulome in hepatocellular carcinoma via consensus GRN inference

Mateusz Garbulowski, Stockholm University, SciLifeLab (mateusz.garbulowski@scilifelab.se)
Mateusz Garbulowski[1], Riccardo Mosca[2], Carlos J. Gallardo-Dodd[2], Claudia Kutter[2], Erik L. L. Sonnhammer[1]
[1]Department of Biochemistry and Biophysics, Stockholm University, Science for Life Laboratory, Solna, Sweden
[2]Department of Microbiology, Tumor, and Cell Biology, Karolinska Institute, Science for Life Laboratory, Solna, Sweden

RNA binding proteins (RBPs) is a group of RNA-targeting proteins that is associated to post-transcriptional mechanisms in cells. In cancer, RBPs contribute to drug resistance and oncogenesis. In this research, we aim at identifying the hepatocellular carcinoma (HCC) regulome using consensus gene regulatory network (GRN) inference for RBP knockdown. More specifically, the HCC GRN is constructed with perturbation-based design using 12 methods, including regression and machine learning. In addition, methodology was benchmarked with synthetic data that revealed decrease of false positive links while using consensus approach. Furthermore, to corroborate regulator-target interactions, we performed validation including such resources as GTEx and TCGA gene expression, GRAND scores, FunCoup5 scores, and RBP binding sites from eCLIP-seq and RAP-seq. For example, LIN28B positively regulates PTBP1, what was corroborated with GRAND, eCLIP-seq, TCGA and consensus links. Finally, we performed community enrichment for the GRN that revealed subnetworks related to diverse cancer and survival pathways.

Keywords:
gene expression, RNA binding proteins, gene regulatory network, machine learning

9. Utilizing large-scale genomic data to explore variable disease penetrance

Sanna Gudmundsson, SciLifeLab/KTH (sanna.gudmundsson@scilifelab.se)
Sanna Gudmundsson 1,2,3, Moriel Singer-Berk 2, Sarah L. Stenton 2,3, Nicholas A Watts 2, Julia K. Goodrich 2, Michael Wilson 2, Genome Aggregation Database Consortium, Samantha Baxter 2, Heidi L. Rehm 3,4,5, Daniel G. MacArthur 2,6,7, Anne ODonnell-Luria 2,3,4
1. Science for Life Laboratory, Department of Gene Technology, KTH Royal Institute of Technology, Stockholm, Sweden
2. Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA
3. Division of Genetics and Genomics, Boston Childrens Hospital, Boston, MA
4. Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA
5. Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA, USA
6. Centre for Population Genomics, Garvan Institute of Medical Research and UNSW
Sydney, Sydney, New South Wales, Australia
7. Centre for Population Genomics, Murdoch Childrens Research Institute, Melbourne,
Australia

Large collections of human genomic sequencing data enable us to study sequence variant architecture and rare phenomena in the general population. The Genome Aggregation Database (gnomAD) is the most widely used open-access variant database with aggregated data from 200,000 individuals. Investigating presumably unaffected individuals in gnomAD with a variant associated with the disease provides an opportunity to identify mechanisms of incomplete penetrance and increase understanding of variant effects. Focusing on the predicted loss of function (pLoF) variants (nonsense, frameshift, and essential splice site variants) associated with severe, highly penetrant, pediatric phenotypes not expected in individuals in gnomAD, and access to over 76,000 genomes allowed us to study hundreds of these cases. Our studies have increased our understanding of pathogenic variants' rescue and error modes and improved our ability to interpret the human genome.

Keywords:
pLoF, population data, gnomAD, variant interpretation, genome sequencing

10. U-FISH: a universal deep learning approach for accurate FISH spot detection across diverse datasets

Weize Xu, AI Cell Lab, KTH (vet.xwz@gmail.com)
Weize Xu (College of Veterinary Medicine, Huazhong Agricultural University, Wuhan, China; State Key Laboratory of Agricultural Microbiology, Huazhong Agricultural University, Wuhan, China) Huaiyuan Cai (College of Veterinary Medicine, Huazhong Agricultural University, Wuhan, China; State Key Laboratory of Agricultural Microbiology, Huazhong Agricultural University, Wuhan, China) Qian Zhang (College of Informatics, Huazhong Agricultural University, Wuhan, China) Gang Cao (College of Veterinary Medicine, Huazhong Agricultural University, Wuhan, China; State Key Laboratory of Agricultural Microbiology, Huazhong Agricultural University, Wuhan, China; College of Informatics, Huazhong Agricultural University, Wuhan, China; Bio-Medical Center, Huazhong Agricultural University, Wuhan, China)

In the ever-advancing landscape of fluorescence in situ hybridization (FISH) technologies, there exists a significant need for sophisticated, yet adaptable, spot detection algorithms. This study introduces U-FISH, a deep-learning-based solution that sets new benchmarks in accuracy and generalizability. Our model is trained and validated on a comprehensive dataset comprising over 4,000 images and more than 1.6 million manually annotated spots, from both real-world experiments and simulations. The algorithm's versatility eliminates the need for laborious manual parameter tuning and allows for application across a wide range of datasets and formats. Moreover, U-FISH is designed for high scalability, supporting large and complex data storage formats and extending its applicability to 3D FISH data. To foster community adoption and accessibility, a user-friendly API, command-line interface, Napari plugin, and web application are provided. The complete dataset is publicly available, serving as a robust foundation for future research in this domain.

Keywords:
FISH spot detection; deep learning

11. Model based DNA basecallers

Joakim Jaldén, KTH Royal Institute of Technology (jalden@kth.se)
Xuechun Xu, Joakim Jaldén

Basecalling, a pivotal step in Nanopore genome sequencing, involves transforming measured current signals into nucleotide sequences, i.e., basecalling. Current methods employing CTC-DNN architectures with tens of millions of parameters are efficient but costly. This poster presents our ongoing work on developing large scale graphical model based basecallers and hybrid basecallers for nanopore sequencing, and computationally efficient inference algorithms for these models. This includes a million-state HMM specifically adapted for GPU implementation, coupled with a custom decoding algorithm. This tailored approach accelerates basecalling while preserving cost-effectiveness, making it an appealing choice for broader applications.

Keywords:
DNA, Nanopore sequencing, Hidden Markov models, deep learning, GPU acceleration

12. Detecting resin wood in Scots pine with laboratory-bespoke X-ray computed tomography imaging

Sheng Joevenller, Luleå University of Technology (sheng.joevenller@associated.ltu.se)
Sheng Joevenller; Department of Mathematics and Engineering, Luleå University of Technology.
Fredrik Nysjö; NBIS-SciLifeLab, Uppsala University.
Dick Sandberg; Department of Mathematics and Engineering, Luleå University of Technology.

The increasing challenge of pine forests confronting notably Scots pine blister rust (SPBR), raises concerns in Norrland Sweden. SPBR, a fungal pathogen targeting tree trunks, triggers heightened resinous canker. We collected eight SPBR-infected pine trees and two healthy pine trees to develop an imaging segmentation for identifying resin wood based on pixel intensity variations.The method draws inspiration from ambient occlusion in computer graphics and visualization for lights. Instead of assessing ray visibility, we focus on discerning differences between the sampled image intensity profile along a ray and the ideal heartwood intensity profile encircled by resin wood. These distinctions are cumulatively computed for each pixel and adjusted by a normalized distance map to prevent misclassification of resin wood pixels near the bark as heartwood. Our current setup employs eight rays per pixel in a given slice. Further improvement to the heart wood segmentation is necessary to resolve sampled from empirical data.

Keywords:
Computed Tomography, Image Segmentation, Scots pine blister rust, resinous wood.

13. Inferring dissipation from cell-membrane fluctuations

Sreekanth K Manikandan, Department of Chemistry, Stanford University (sreekm@stanford.edu)
K Manikandan, Sreekanth - Wallenberg Postdoctoral Fellow, Department of Chemistry, Stanford University, USA
Ghosh, Tanmoy - PhD student, Indian Institute of Science Education and Research Kolkata, India
Mandal, Tithi - PhD student, Indian Institute of Science Education and Research Kolkata, India
Arikta Biswas - PhD student, Indian Institute of Science Education and Research Kolkata, India
Bidisha Sinha - Assistant Professor, Indian Institute of Science Education and Research Kolkata, India
Dhrubaditya Mitra - Assistant Professor,
NORDITA, KTH Royal Institute of Technology and
Stockholm University, Sweden

We present a novel scheme that can be used to infer the energy dissipation rate from flickering data of active cell-membranes. Our method distinguishes active cell membranes from ATP-depleted states and resolves activity at the µm scale. We demonstrate its effectiveness using Interference Reflection Microscopy data from HeLA cells. This approach combines recent findings from non-equilibrium statistical physics with machine learning, yielding model-agnostic dissipation estimates.

Keywords:
Inference of dissipation, cell-flickering data, Machine-learning for inference

14. An Advanced Deep Learning Pipeline for Genome-Wide Imaging Screen Analysis Uncovering Cell Death Regulators

Salma Kazemi Rashed, lund university (salma.kazemi_rashed@med.lu.se)
Salma Kazemi Rashed ,Klara Esbo ,Mariam Miari, Malou Arvidsson , Liu J, Cowley K, Simpson K, Johnstone R, Sonja Aits

In many image-based cell studies, a plethora of unstructured information is embedded in cell objects, which must be accurately extracted before further utilization in linking to underlying biological mechanisms explored in the original study. Often, extracting this information poses a significant challenge due to the large and intricate data, as well as the need for highly accurate tools. One effective technique for this purpose is instance segmentation of cell objects. In our study, which we explore cell death and associated regulatory networks through a genome-wide microscopic screen, the dataset is extensive and challenging to manually curate or annotate. To address this, we trained U-net-based nuclei segmentation models using a small but meticulously annotated dataset to ensure high accuracy, achieving an F1-score of 89%. Subsequently, we selected the superior U-Net model to segment entire objects in the genome-wide screen, comprising nearly 5 million image frames. By taking the extracted features from the nuclei channel as biological metrics to explore cell death, the downstream analysis has revealed a correlation between the extracted features from screen and several well-known genes associated with established cell death pathways such as apoptosis.

Keywords:
Cell death, Deep learning, Genome-wide screen, Nuclei Segmentation, U-Net

15. Beam search decoder for enhancing sequence decoding speed in single-molecule peptide sequencing data

Javier Kipen, KTH (kipen@kth.se)
Javier Kipen, Joakim Jaldén. KTH, Stockholm, Sweden.

Next-generation single-molecule protein sequencing technologies have the potential to accelerate biomedical research significantly. These technologies offer sensitivity and scalability for proteomic analysis. One auspicious method is fluorosequencing, which involves: cutting naturalized proteins into peptides, attaching fluorophores to specific amino acids, and observing variations in light intensity as one amino acid is removed at a time via Edman degradation. The original peptide is classified from the sequence of light-intensity reads, and proteins can subsequently be recognized with this information. Even though a framework (Whatprot) has been proposed for the peptide classification task, processing times remain restrictive due to the massively parallel data acquisition system. We propose a new beam search decoder with a novel state formulation that obtains much lower processing times with slightly higher accuracies than Whatprot. Furthermore, we explore how our novel state formulation may lead to even faster decoders in the future.

Keywords:
Signal Processing, Protein sequencing, Fluorosequencing

16. BIIF - Support on BioImage Analysis in Sweden

Anna Klemm, SciLifeLab (anna.klemm@it.uu.se)
BioImage Informatics Unit, SciLifeLab and NBIS

The BioImage Informatics Unit provides the support and training to perform state-of-the-art analyses on your image data. Our experts can help you deploy computational methods using computer vision, machine learning, and bioinformatics to analyze your images.
SERVICES

- Advice on best-practice and guidance on overall experimental design (staining, sample preparation, and image acquisition) for research involving microscopy imaging and quantitative data analysis.
- Guidance on image analysis assay development, including image processing algorithm development and software engineering to address challenging project goals.
- Advice on best-practice and guidance on high throughput/large-scale image processing using computing clusters, including data transfer and storage during the activity of the project.
- Guidance on large-scale data analysis and visualization.
- Dissemination of bioimage analysis knowledge in courses and workshops.

Contact: biif@scilifelab.se
Web: https://www.scilifelab.se/units/bioimage-informatics/

Keywords:
bioimage informatics. image analysis

17. Single-cell RNA sequencing-based program-polygenic risk scores associated with pancreatic cancer risks in the UK Biobank cohort

Yelin Zhao, KI (yelin.zhao@ki.se)
Yelin Zhao[1], Martin Smelik[1]*, Xinxiu Li[1], Oleg Sysoev[2], Firoj Mahmud[1]*, Dina Mansour Aly[1]*, Mikael Benson[1]**
[1] Medical Digital Twin Research Group, Department of Clinical Science, Intervention and Technology (CLINTEC), Karolinska Institutet, Stockholm, Sweden.
[2] Division of Statistics and Machine Learning, Department of Computer and Information Science, Linköping University; Linköping, Sweden.

Background: Partitioned PRS using single-cell RNA sequencing (scRNA-seq) may provide a valuable tool to assess increased or decreased risk of pancreatic cancer.
Methods: ScRNA-seq data from PC tumor and normal tissues were. Pathway enriched with differentially expressed genes (DEGs) were clustered into programs based on gene similarity. Program PRSs (pPRSs) were created by mapping PC associated genetic variants to each program and were evaluated in UK Biobank participants (1,310 PDAC and 407,473 controls). The Cox regression analysis was performed to determine associations of pPRSs and PC risk.
Results: Twenty-four pPRSs were generated. Four distinct PRSs were significantly associated with an increased risk of developing PC. Among these, P6 exhibited the greatest hazard ratio (adjusted HR[95% CI] = 1.67[1.14-2.45]). In contrast, P10 and P4 were associated with lower risk of developing PC (0.58[0.42-0.81] and 0.75[0.59-0.96]).
Conclusion: ScRNA-seq-based pPRSs may be used to assess increased or decreased risk of PDAC.

Keywords:
Pancreatic cancer, polygenic risk score, scRNA-seq, UK Biobank

18. Deep Learning with Big Data for Genetic Epidemiology

Max Kovalenko, Uppsala University (max.kovalenko@it.uu.se)
Kovalenko Max, Department of Information Technology, Science for Life Laboratory, Uppsala University, Sweden
Nettelblad Carl, Department of Information Technology, Science for Life Laboratory, Uppsala University, Sweden
Johansson Åsa, Department of Immunology, Genetics and Pathology, Science for Life Laboratory, Uppsala University, Sweden

The poster lays out a project that aims to capture a concise representation of human genetic variation, which will then be used to predict traits, and ultimately -- to identify their causal variants. The work is based on a prototype model -- a convolutional autoencoder called GCAE, originally developed for dimensionality reduction of human SNP data. Dimensionality reduction is useful both for eliminating confounding by population structure and for reducing the number of features for deep learning. GCAE will be trained on the UK Biobank dataset, containing an unprecedented amount of genetic and health information; this represents serious computational challenges, which are currently being addressed. In addition, GCAE is to be supplemented with a trait prediction module. The intended application is prediction of genetic disease risk and discovery of genetic variants that contribute to it.

Keywords:
deep learning, autoencoder, genetic epidemiology

19. Spatial transcriptomics and GWAS data identify putative causal tissue structures for complex traits

Linda Kvastad, SciLifeLab (KTH) (linda.kvastad@scilifelab.se)
L. Kvastad[1], A. Kollotzek[1], C. Chang[1] and T. Lappalainen[1-3]
[1] Science for Life Laboratory, Department of Gene Technology, KTH Royal Institute of Technology, Stockholm, Sweden.
[2] New York Genome Center, New York, NY, USA.
[3] Department of Systems Biology, Columbia University, New York, NY, USA.

Pinpointing the causes and discovering effective treatment targets for complex diseases remains challenging. However, it has been estimated that selecting genetically supported drug targets could increase new drugs clinical development success rate. The Open Targets platform provides a systematic drug-target identification and prioritization resource, integrating publicly available datasets and scoring target-disease associations. Here we integrate these data with published spatial transcriptomics datasets of healthy tissues from humans and mice to identify and prioritize spatial structures of interest for complex disease traits. Comparing disease-associated genes per trait to a random set of genes, we found that in the colon, genes associated with ulcerative colitis were enriched towards the top of the villus, whereas for Crohns disease impacted submucosal lymphoid follicles. In brain, schizophrenia-associated genes were enrichment in the cortex and hippocampus, with ongoing analyses of diverse diseases and traits linked to the heart, lung, spinal cord, kidney, and liver.

Keywords:
Spatial transcriptomics, GWAS, Complex diseases

20. Bacterial vaginosis: Understanding the effect of antibiotic-free treatment on the vaginal microbiome

Emilia Lahtinen, Karolinska Institutet, CTMR/MTC (emilia.lahtinen@ki.se)
Lahtinen Emilia (presenter), Centre for Translational Microbiome Research, Department of Microbiology, Tumor and Cell Biology (MTC), Karolinska Institutet, Stockholm, Sweden & Gedea Biotech AB, Medicon Village, Sweden
Edfeldt Gabriella, Centre for Translational Microbiome Research, Department of Microbiology, Tumor and Cell Biology (MTC), Karolinska Institutet, Sweden. W. Hugerth Luisa, Science for Life Laboratory, Department of medical biochemistry and microbiology, Uppsala University, Sweden
Strevens Helena, Gedea Biotech AB, Medicon Village, Sweden
Kornfält Sten, Gedea Biotech AB, Medicon Village, Sweden
Säfholm Annette, Gedea Biotech AB, Medicon Village, Sweden
Engstrand Lars, Centre for Translational Microbiome Research, Department of Microbiology, Tumor and Cell Biology (MTC), Karolinska Institutet, Sweden
Schuppe Koistinen Ina, Centre for Translational Microbiome Research, Department of Microbiology, Tumor and Cell Biology (MTC), Karolinska Institutet, Sweden

Bacterial vaginosis (BV) results from an imbalance in the vaginal microbiome and impacts nearly 30% of reproductive-age women globally. The molecular mechanisms underlying BV are unknown leading to treatment challenges. Current approaches rely on antibiotics, however recurrence rates can reach 50% within a year. Repeated treatments also increase the burden of antibiotic resistance genes. Gedea Biotech AB has developed an antibiotic-free intra-vaginal tablet, pHyph, to treat BV. pHyph aims to disrupt the biofilm supporting harmful bacteria, restore favorable vaginal pH and promote the growth of Lactobacilli to treat and prevent recurrence of BV. Preliminary results indicate that pHyph inhibits the growth of bacteria involved in functions contributing to BV formation, such as biofilm formation. These functions were primarily associated with members of the Prevotella genus. Further data analysis will investigate factors contributing to differences in vaginal microbiome composition, success of the treatment with pHyph, and the recurrence of BV.

Keywords:
vaginal microbiome, women’s health, bacterial vaginosis, antibiotic-free

21. SciCommander Track provenance of any shell command

Samuel Lampa, Karolinska University Hospital (samuel.lampa@scilifelab.se)

There exist a multitude of pipeline tools for bioinformatics [1]. As using a pipeline tool is more complex than just writing shell scripts [2], a lot of bioinformatics work happens in a more ad-hoc fashion, with individual shell commands executed to run analyses. This makes it much harder to keep a full audit log of the analyses, since it is easy to miss to document some steps. It is also often not clear afterwards which output files were created by which command. Additionally, shell scripts are missing functionality to re-use already finished intermediary output files to resume cancelled runs. SciCommander is a tool that addresses these limitations by allowing to track the produced output files of almost any shell command and avoiding to re-run already run commands, by only slight changes in the commands, prepending the command itself and the inputs, outputs and parameters of the command with special markers. Using this information, SciCommander checks if any of the output files already exist, and if so skips that command. Secondly, it produces an audit log for each output file in JSON format. It includes a command to convert this file into an HTML report with a graphical visualization of all the steps needed to produce this particular file. All in all, this functionality allows providing full provenance at the individual file level, as well as allows resuming interrupted runs even for simple shell commands. SciCommander is open source, can be installed via the Python Package Index, or from GitHub: https://github.com/samuell/scicommander

References

[1] Wratten L, Wilm A & Göke J (2021) Reproducible, scalable, and shareable analysis pipelines with bioinformatics workflow managers. Nat Methods 18: 11611168.
[2] Spjuth O, Bongcam-Rudloff E, Hernández GC et al. (2015) Experiences with workflows for automating data-intensive bioinformatics. Biol Direct 10 (43).

Keywords:
Reproducibility, Provenance

22. Kidney automatic segmentation and cystic renal occupying lesions classification system based on the Deep Learning.

Hui Li, Chalmers University of Technology (lihui_lois@163.com)

Kidney disease is a major concerns in adults, with globally increasing incidence and mortality rates. Artificial intelligence (AI) machine learning models are promising tools for earlier disease detection and diagnosis, which can increase the chances for successful treatment, compared to current standards.
Using abdominal CT image data from 1 084 patients, we built an automatic segmentation and cystic renal occupying lesions classification system, with deep learning 2D and 3D models to segment (2D) and classify (3D) the renal lesions. Different combinations of phases were used to predict the risk of lesions and provide interventional treatment strategies to support radiologists and surgeons.
AI models were evaluated using the Dice coefficient for segmentation and the accuracy for classification. Results confirm that our AI models have comparable performance to the current histopathological standard and as such are currently in use across multiple hospitals.

Keywords:
Automated Diagnosis, Deep Learning, Image Segmentation, Lesion Classification, Kidney

23. Metagenomics of insect bulkDNA for population genetics

Samantha López Clinton, Swedish Museum of Natural History (samantha.lopezclinton@nrm.se)
López Clinton Samantha, Swedish Museum of Natural History, Centre for Palaeogenetics, Stockholm University, Sweden, Jin Chenyu, Swedish Museum of Natural History, Centre for Palaeogenetics, Stockholm University, Sweden, Miraldo Andreia, Swedish Museum of Natural History, Sweden, Iwaszkiewicz-Eggebrecht Elzbieta, Swedish Museum of Natural History, Sweden, Ronquist Fredrik, Swedish Museum of Natural History, Sweden, van der Valk Tom, Swedish Museum of Natural History, Centre for Palaeogenetics, Sweden

Monitoring biological and genetic diversity, which includes measures related to population genetics, plays a vital role in conservation efforts by furnishing insights into the conditions and potential threats facing species and ecosystems. In this study, we delve into the application of extensive databases storing terabytes of data, along with specialized software for taxonomic classification and subsequent sequence alignments. Our focus is on genome-wide shotgun sequencing data obtained from assorted arthropod samples collected throughout Sweden as part of the Insect Biome Atlas Project. We explore the possibilities and constraints of employing a genome-wide approach to not only ascertain species composition but also population genomics parameters among populations collected from diverse locations. Through the exploration of novel metagenomics bioinformatic techniques applied to mixed and environmental DNA samples, we contribute to enhancing the accessibility and efficacy of conservation strategies rooted in genomics, thereby setting a significant precedent for conducting metagenomic-scale population genomics research.

Keywords:
Metagenomics, population genetics, bioinformatics, biodiversity, arthropoda

24. Multimodal signal recordings in neuroscience: modeling strategies

Melisa Maidana Capitan, Linköping University (melisa.maidana.capitan@liu.se)
Melisa Maidana Capitan, Linköping University
Alejandra Alonso, Radboud University
Francesco Battaglia, Radboud University
Annumita Samanta, Radboud University
Adrian Aleaman, Radboud University
Lisa Genzel, Radboud University
Fredrik Lindsten, Linköping University
Marcin Szczot, Linköping University

The field of systems neuroscience is experiencing significant growth, driven by advancements in in-vivo registration technologies that enable simultaneous recording of multiple data modalities in complex experimental setups. These modalities encompass time series signals of both brain activity, behavioral readouts, and external stimuli.
We will show how standard machine learning tools have been employed to extract pertinent features for describing brain activity in two examples. The first focuses on classifying electrophysiological recordings into different oscillatory types, while the second centers on predicting free-ranging behavior using calcium imaging signals. These examples show the importance of refining modeling techniques to understand the relationship between neural activity and behavior.
As we introduce a third example for head-fixed mice, we will describe the benefits of developing specific models tailored to these datasets which take into account the specific nature of biological and behavioral signals.

Keywords:
systems neurocience, memory, hippocampus

25. National Genomics Infrastructure (NGI) Next Generation Sequencing and Genotyping for Swedish Research

Tom Martin, National Genomics Infrastructure (NGI) (tom.martin@medsci.uu.se)
National Genomics Infrastructure (NGI)

An overview of the services provided by NGI

Keywords:
NGS, Genotyping, Single Cell, Proteomics, Spatial

26. Multi-metabolic signature of controlled modification of dietary carbohydrate quality

Cecilia Martinez Escobedo, Chalmers University of Technology (escobedo@chalmers.se)
Cecilia Martinez Escobedo,
Rikard Landberg
Clemens Wittenbecher

The metabolic impact of high glycemic index (GI) carbohydrates (CHO) is well documented and has plausible links to cardiometabolic disease (CMD). Deriving multi-metabolite signatures (MMS) of dietary GI holds potential in elucidating the metabolic adaptations to CHO quality and understanding the long-term effects on CMD risk.
This study aims to develop a MMS that captures the long-term metabolic adaptation to CHO quality. We generated LC-MS based untargeted metabolomics data from pre- and post-intervention plasma samples in the MedGICarb study, a 12-week RCT with 135 participants in two arms: low-GI and high-GI. We used cross validated elastic net regression (CV-ENR) to train a multi metabolite model that predict the dietary intervention group. The CV-ENR selected 15 metabolite features that were significantly associated with high vs low GI of dietary carbohydrates in the validation set (p-value=0.01). A weighted MMS based on these features effectively predicted the dietary intervention group (accuracy=0.725). Our results indicate that a multi-metabolite signature captures the metabolic adaptation to dietary carbohydrate quality.

Keywords:
glycemic index, carbohydrates, cardiometabolic disease, untargeted metabolomics, multi-metabolite signature

27. From Genes to Causal Maps: A Benchmark for Gene Regulatory Network Inference

Mariia Minaeva, KTH Royal Institute of Technology (mariia.minaeva@scilifelab.se)
Minaeva Mariia, Science for Life Laboratory, Department of Gene Technology, KTH Royal Institute of Technology, Stockholm, Sweden
Scherrer Nino, Independent, Switzerland
Bauer Stefan, Technical University of Munich, Germany; TUM School of Computation, Information and Technology, Helmholtz AI, Munich, Germany
Lappalainen Tuuli, Science for Life Laboratory, Department of Gene Technology, KTH Royal Institute of Technology, Stockholm, Sweden, New York Genome Center, New York, NY 10013, USA

Gene regulatory networks (GRNs) are essential for understanding cellular biology, often represented as graphical models. The challenging task of inferring GRNs is commonly tackled using correlations-based or causal structure learning approaches. However, performance evaluation remains challenging due to simplistic synthetic benchmarks that do not match real-world data structure, the lack of biological ground truth GRNs, and hard-to-interpret performance metrics. To address these issues, we introduce a novel GRN-inference benchmark, GRN-Bench, designed to accommodate both method categories. By incrementally increasing data complexity, from noise-free linear to noisy kinetic-driven simulations, we assess the impact of data characteristics on performance through a set of tailored evaluation metrics. Notably, our study reveals substantial variation in predictive potential among different data generation mechanisms, emphasising method limitations in real-world scenarios, particularly regarding missing data. This benchmark provides critical insights into GRN inference quality and offers a practical tool for advancing gene regulatory inference.

Keywords:
Causal Discovery; Benchmark; GRN

28. Identification of metabolomic networks linked with incident heart failure.

Jakub Morze, SGMK Copernicus University / Chalmers University of Technology (jakub.morze@sgmk.edu.pl)
Morze, Jakub, MD (a,b,c)
Wittenbecher, Clemens, PhD (b)
Guasch-Ferré, Marta, PhD (d,c)
Rynkiewicz, Andrzej, MD, PhD (a)

a. SGMK Copernicus University, Poland
b. Chalmers University of Technology, Sweden
c. Harvard T.H. Chan School of Public Health, USA
d. University of Copenhagen, Denmark

High-throughput metabolomics holds significant promise in elucidating the underlying mechanisms of heart failure (HF). However, to date, only a limited number of prospective cohort studies have examined it in the context of incident HF. In this study, we aimed to identify metabolomic networks associated with the risk of HF. We employed Gaussian Graphical Models to construct metabolomic networks based on NMR Nightingale platform markers analyzed in baseline plasma samples from 93,922 participants (including 1,638 incident HF cases) from the UK Biobank. Out of the ten identified network clusters, seven showed associations with incident HF after multivariable adjustment. The strongest positive association with HF risk was observed for the ”Triglycerides, Fatty Acids, and Other Lipids” cluster, while the strongest inverse association was found for the ”Albumin-Amino acids” cluster. This study provides observational evidence that may contribute to the identification of novel metabolic biomarkers for HF.

Keywords:
Metabolomics; Heart failure; NMR; Metabolomic networks

29. Predicting Protein-Protein Interactions using Machine Learning

Sarah Narrowe Danielsson, Stockholm University/SciLifeLab (sarah.narrowe@scilifelab.se)
Sarah Narrowe Danielsson[1], Arne Elofsson[1]

[1] Department of Biochemistry and Biophysics, Stockholm University, Science for Life Laboratory, Box 1031, 17121 Solna, Sweden

Proteins, the workforce and fundamental building blocks in humans, consist of amino acids arranged into complex chains, often working collaboratively to perform their function. Understanding protein-protein interactions (PPIs) is pivotal in order to fully understand their function. Experimental methods for investigating PPIs are available but suffer from high false positive rates and conflicting results. To address these challenges, computational approaches have emerged, notably AlphaFold2 (AF2) by Deepmind, which excelled in predicting protein structures, including multi-chain complexes, and subsequently, PPIs. Despite its success, efforts continue to improve upon its limitations. AF2 relies on time-consuming multiple sequence alignments (MSAs) to extract co-evolutionary information. This benchmark study includes alternative methods such as OmegaFold and ESMFold, which do not rely on MSAs. Additionally, methods like D-SCRIPT offer probability-based PPI predictions. While AF2 outperforms other methods in accuracy, alternatives may be valuable for faster results. Future research aims to shift from benchmarking to novel PPI prediction method development.

Keywords:
Protein-Protein Interactions, Machine Learning, Bioinformatics

30. Recalibrating differential gene expression analysis by variance in gene dosage

Philipp Rentzsch, SciLifeLab (philipp.rentzsch@scilifelab.se)
Rentzsch Philipp, SciLifeLab Solna, Sweden
Kollotzek Aaron, SciLifeLab Solna, Sweden
Mohammadi Pejman, University of Washington, USA
Lappalainen Tuuli, SciLifeLab Solna & New York Genome Center, Sweden & USA

Differential expression testing is a prevalent method for identifying genes that are functionally relevant for an observed phenotype. However, gene selection based on expression fold-change tends to favor variable genes. This limits the application of differential expression approaches in the study of mechanisms where dosage sensitivity is a contributing factor.

Addressing this limitation, we have developed a method to recalibrate expression fold-change per gene based on variance in the human population. The recalibrated metric ranks genes not by nominal fold-change, but in comparison to natural dosage variation. With that, our method adjusts the expression bias observed for known disease genes when using nominal gene expression and highlights pathways and biological processes related to metabolic and regulatory activity.

The provided, novel view on differential expression bridges the gap between statistical and biological significance. We believe that our approach will improve the identification of disease candidate genes and enhance therapeutic target discovery.

Keywords:
differential expression, statistics, RNAseq,

31. Mapping the genetic architecture of sarcoidosis across populations

Natalia Rivera, Karolinska Institutet (natalia.rivera@ki.se)

Keywords:
sarcoidosis; autoimmune diseases

32. The Swedish Childhood Tumor Biobank

Johanna Sandgren, KI (johanna.sandgren@ki.se)
Teresita Díaz de Ståhl1, Elisa Basmaci1, Gabriela Prochazka1, Praveen Raj Somarajan1, Katarzyna Zielinska-Chomej1, Maxime Garcia1,2, Christopher Illies3, Karim Katkhada1, Markus Mayrhofer4, Jenny von Salomé, Susan Pfeifer5, Monica Nistér1, Gustaf Ljungman5 and Johanna Sandgren1,3
1Department of Oncology-Pathology, Karolinska Institutet, Stockholm, Sweden
2National Genomics Infrastructure, Science for Life Laboratory, Stockholm, Sweden
3Clinical Pathology and Cancer Diagnostics, Karolinska University Hospital, Stockholm, Sweden
4National Bioinformatics Infrastructure Sweden, Uppsala University, Uppsala, Sweden
5Department of Women´s and Children's Health, Uppsala University, Uppsala, Sweden.

In Sweden 340 children are diagnosed with cancer each year. Deeper biological knowledge on these malignancies is essential for improved survival and quality of life. The aim of The Swedish Childhood Tumor Biobank (Barntumörbanken, BTB) is to increase the understanding of pediatric solid tumors by providing infrastructure resources, biological samples and molecular genetic/genomic data for research.
BTB has a nation-wide collaboration with the six university hospitals that treat and operate pediatric cancer patients. Fresh frozen tumors and blood samples are collected, with informed consent. BTB prepares and stores the biobank samples as well as performs whole genome sequencing, whole transcriptome sequencing and methylation array profiling. Bioinformatic pipelines are co-developed with SciLifeLab.
2100 cases are now registered and 26 000 samples collected. More than 1300 cases have been genomically characterized were BTB are responsible of the generated data, including from the GMS Barncancer study. BTB samples and data have been distributed to several research projects after formal application processes. BTB is also assisting clinical studies with sample logistics and data analysis/interpretation.

Keywords:
genomic, pediatric cancer, biobank

33. Targeted Proteomics of Blood Plasma from the hPOP Cohort

Thanadol Sutantiwanichkul, KTH (thanadol@kth.se)
Thanadol Sutantiwanichkul 1, Fredrik Edfors 1
KTH Royal Institute of Technology and Science for Life Laboratory, Sweden

Plasma targeted proteomics becomes an accessible technology to quantitatively and precisely used for precision medicine. During the HUPO conferences from 20162018, several specimens have been collected to explore with integrative techniques worldwide. Using Proteomics RECombinaint isotopic Standard (PRECiS) technology developed by the Targeted Proteomics group at KTH, 71 plasma protein targets from 321 samples were quantified at transitional levels. At the peptide level, batch effect correction has been considered from the coefficient of variation (CV) of top ten reproducible peptides with a linear regression approach. We found this calibration technique can robustly cover plasma targets with a high dynamic range. In addition, plasma peptide discovery was also reproducible among the cohort. With sample metadata, a few peptides were able to stratify the target groups. In summary, we quantified targeted plasma proteins precisely with mass spectrometry. We could also correct a number of cohort samples with a robust and reproducible approach.

Keywords:
Precision medicine, Proteomics, Mass spectrometry, Sytems biology

34. Barcode-free prediction of cell lineages from scRNA-seq datasets

Marcel Tarbier, KI / SciLifeLab (marcel.tarbier@scilifelab.se)
Almut S. Eisele (1,,)*
Marcel Tarbier (2,) - presenting author
Alexia A. Dormann (1)
Vicent Pelechano (2)
David M. Suter (1,)*

1: Ecole Polytechnique Fédérale de Lausanne, School of Life Sciences, Institute of Bioengineering; Lausanne, Switzerland
2: Science for Life Laboratory, Department of Microbiology, Tumor and Cell Biology, Karolinska Institute; Solna, Sweden
**: Corresponding authors*
: These authors contributed equally to this work

Assigning single cell transcriptomes to cellular lineage trees by lineage tracing has transformed our understanding of differentiation during development, regeneration, and disease. However, lineage tracing is technically demanding and most scRNA-seq datasets are devoid of lineage information. Here we introduce Gene Expression Memory-based Lineage Inference (GEMLI), a computational tool allowing to robustly determine cell lineages solely from scRNA-seq datasets. GEMLI allows to study heritable gene expression, to discriminate symmetric and asymmetric cell fate decisions and to reconstruct individual multicellular structures from pooled scRNA-seq datasets. In human breast cancer biopsies, GEMLI revealed previously unknown gene expression changes at the onset of cancer invasiveness. The universal applicability of GEMLI allows studying the role of cell lineage trees in a wide range of physiological and pathological contexts. GEMLI is available as an R package on GitHub (https://github.com/UPSUTER/GEMLI).

Keywords:
bioinformatics, computational proxies, lineage predictions

35. Deep Learning for Time Series Classification of Parkinsons Disease Eye Tracking Data

Gonzalo Uribarri, EECS & SciLifeLab, KTH (uribarri@kth.se)
Gonzalo Uribarri, EECS & SciLifeLab, KTH, Sweden
Erik Fransén, EECS & SciLifeLab, KTH, Sweden

Eye-tracking is an accessible and non-invasive technology that provides information about a subject's motor and cognitive abilities and is a valuable resource in the study of neurodegenerative diseases such as Parkinson's disease. However, to date, no single eye movement biomarker has been found that can conclusively differentiate patients from healthy controls. In the present work, we investigate the use of state-of-the-art deep learning algorithms to perform Parkinson's disease classification using eye tracking. We implement two different time series classification models, InceptionTime and ROCKET. We find that the models are able to learn to classify unseen subjects from □1.5 s long fixation intervals, achieving 78% and 88% accuracy, respectively. We also developed a novel method for pruning the ROCKET model to improve its interpretability and generalizability, achieving 96% accuracy. Our results suggest that fixation data are suitable for use with machine learning in the discovery of disease-relevant biomarkers.

Keywords:
Eye-tracking, Deep Learning, Parkinson, Time Series

36. Benchmarking orthologous clustering programs for proteins a case study in Pseudomonas aeruginosa

Virág Varga, Department of Life Sciences (LIFE), Division of Systems and Synthetic Biology, Chalmers University of Technology (virag.varga@chalmers.se)
Varga Virág (Division of Systems and Synthetic Biology, Department of Life Sciences, Chalmers University of Technology, Gothenburg, Sweden), Bengtsson-Palme Johan (Division of Systems and Synthetic Biology, Department of Life Sciences, Chalmers University of Technology, Gothenburg, Sweden)

Orthologous clustering programs utilize a variety of algorithms to sort proteins into presumed gene families, aiming to determine which genes may be descended from a common ancestral gene. Many orthologous clustering tools are available, utilizing a variety of parameters and statistical models, and they allow the user to fine-tune their search in different ways. However, relatively little research has been targeted at benchmarking such tools against one another, in order to suggest ideal use cases for different tools. Here, we present the results of a statistical benchmarking process comparing the efficacy of the CD-HIT, Diamond, MMseqs2 and USEARCH programs. As a case study, we use them to cluster proteins in the bacterial opportunistic pathogen, Pseudomonas aeruginosa. We recommend the statistical tests used here to other researchers intending to evaluate the quality of orthologous clustering programs when used on their own data.

Keywords:
orthologous clustering, benchmarking

37. Predicting plastic-degrading potential of the Baltic Sea: a data-driven approach with taxonomic information

Máté Vass, Chalmers University of Technology, Department of Life Sciences, Division of Systems and Synthetic Biology (mate.vass@chalmers.se)
Máté Vass, Johan Bengtsson-Palme
Chalmers University of Technology, Department of Life Sciences, Division of Systems and Synthetic Biology, Science for Life Laboratory

This study presents a data-driven approach for predicting plastic pollution and its degradation potential in the Baltic Sea. Building upon a recently published machine-learning algorithm, we incorporated taxonomic information through label encoding to enhance the algorithm's performance. The modified algorithm is designed to determine whether an environmental sample possesses the capacity to degrade plastics and, if so, to identify the specific plastic types involved. The primary objective of this research is to leverage this methodology to effectively monitor plastic pollution trends in the Baltic Sea, offering valuable insights into environmental management and mitigation strategies.

Keywords:
machine-learning, environmental pollution, microplastics, Baltic Sea

38. Predicting Preterm Birth Using Machine Learning

Nicole Wagner, Karolinska Institutet (nicole.wagner@ki.se)
Wagner, Nicole, Karolinska Institutet, Sweden
Hugurth, Luisa, Uppsala Universitet, Sweden
Gudnadottir, Unnur, Karolinska Universitet, Sweden
Fransson, Emma, Uppsala Universitet, Sweden
Boulund, Fredrik, Karolinska Institutet, Sweden
Schuppe-Koistinen, Ina, Karolinska Institutet, Sweden
Engstrand, Lars, Karolinska Institutet, Sweden

The Swedish Maternal Microbiome Project (SweMaMi) is a comprehensive study investigating the link between maternal microbiomes during pregnancy and pregnancy outcomes and child health. This research encompasses 3100 participants whose pregnancies concluded between 2018-2021. Participants completed detailed questionnaires and provided vaginal, fecal, and saliva samples at three intervals: first trimester, second trimester, and post-delivery. Microbiome taxonomic annotations were generated using Metaphlan4. Utilizing a nested case-control framework, distinguishing preterm from term births, we employed PERMANOVA and regression analyses to identify potential confounding variables influencing the microbiome. Focusing on fecal and vaginal taxonomic profiles along with confounding variables, we developed predictive models using classifiers such as Random Forest and K-Nearest Neighbors (KNN). The predictive models were rigorously assessed on the SweMaMi cohort and a separate cohort of women to gauge their prediction accuracy. This research offers valuable insights into the complex relationship between maternal microbiomes and pregnancy outcomes and child health.

Keywords:
Preterm birth, Machine learning models, Metagenomics

40. Navigating the Toolbox: A Comparative Analysis of Metagenomic Tools for Taxonomic and Resistance Gene Identification

Marcus Wenne, Chalmers univeristy of technology (marcus.wenne@chalmers.se)
Wenne, Marcus, (1,2), Sweden
Bengtsson-Palme, Johan, (1,2,3) Sweden
1. Division of Systems and Synthetic Biology, Department of Life Sciences, SciLifeLab, Chalmers University of Technology, Gothenburg, Swede
2. Centre for Antibiotic Resistance Research in Gothenburg (CARe), Gothenburg, Sweden
3. Department of Infectious Diseases, Institute of Biomedicine, The Sahlgrenska Academy, University of Gothenburg, Gothenburg, Sweden

In the rapidly evolving field of metagenomics, accurately identifying the composition of taxonomic and antibiotic resistance genes is crucial for a wide range of scientific fields.However, the selection of the most suitable computational tool remains a challenge, often dictated by the specific research question and available computational resources available. This study presents a comparative analysis of metagenomic tools, evaluating their performance in terms of accuracy and resource requirements when it comes to evaluating the taxonomic and resistance gene composition of a sample. Preliminary results suggest a wide range of resource requirements, as well as precision versus recall corves when comparing tools. This benchmarking data will be very valuable for future studies in our group, and hopefully for other researchers as well.

Keywords:
Metagenomics, bioinformatics, AMR, benchmarking

40. SciLifeLab Data Platform

Liane Hughes, SciLifeLab/UU (liane.hughes@scilifelab.uu.se)
Liane Hughes, Katarina Öjefors Stark, Senthilkumar Panneerselvam, Hanna Kultima, and Johan Rung

SciLifeLab Data Centre, SciLifeLab, Sweden.

The SciLifeLab Data Platform comprises a technical environment with a web interface. The main aim of the Platform is to accelerate data-driven life science research by supporting those involved in the research.

The technical environment offers hosting for, among other things, tools and databases related to data-driven life science. The web interface (data.scilifelab.se) displays these alongside other, similar tools and databases. Collectively, we refer to these are services. Users can explore the list of services to find those most relevant to them. The web interface also includes multiple other elements useful for data-driven life science. For example, it includes data-centric articles (highlights) detailing recent research sharing data/code, a resources section consisting of collections of information on topics e.g. compute and storage resources in Sweden, and sections showing relevant jobs, events, and funding opportunities.

Those interested in having their work promoted or hosted on the Platform are invited to reach out!

Keywords:
Data Platform, Services, Data Centre

41. SciLifeLab Serve - enabling sharing of machine learning models and applications

Arnold Kochari, SciLifeLab (arnold.kochari@scilifelab.uu.se)
Kochari Arnold, SciLifeLab Data Centre, Sweden

SciLifeLab Serve is a platform developed by the SciLifeLab Data Centre, funded by the DDLS programme. It allows life science researchers to serve their trained machine learning models and allow custom input predictions by the research community and general public. In addition, it offers hosting of applications with user interfaces for ML models (e.g., Streamlit, Gradio) as well as other types of applications (e.g., Shiny, Dash, Flask). SciLifeLab Serve is free to use for all life science researchers in Sweden and their collaborators.

Keywords:
AI, machine learning, ML applications, data service

42. Services and support from SciLifeLab Data Centre

Katarina Öjefors Stark, SciLifeLab Data Centre (katarina.ojefors.stark@scilifelab.uu.se)