ABSTRACTS

Matchmaking WASP-DDLS

Updated 230328

Information about the call

PIs and Co-PIs need to fulfil the criteria stated under “Who can apply?” to receive funding (see call webpages)

Webpage SciLifeLab
Webpage WASP

Support and questions

ddls-calls@scilifelab.se

info@wasp-sweden.org

Abstracts
Matchmaking WASP - DDLS

Find your collaborator for the joint call 2022

Abstracts listed in incoming order

1. Mark Clements
2. Erik Fransén
3. Wojcech Chacholski
4. Roland Nilsson
5. Erik Sonnhammer
6. Leila Ghalebani
7. Rene Kaden
8. Natalia Rivera
9. Jenny Hansson
10. Lukas Käll
11. Wen Zhong
12. Alexandros Sopasakis
13. Walker Jackson
14. Priyantha Wijayatunga
15. Laura Carroll

Mark Clements

Email: mark.clements@ki.se
Title: Docent
Organization: Karolinska Institutet

1

Research interest

Life Science community
Area: Precision medicine and diagnostics, AI/Math, Software
Categories: Have an open project or idea that could benefit from a collaboration

Abstract

Keywords: Precision health, multi-state models, ordinary differential equations, discrete event simulation, time-to-event models

Multi-state models in precision health with ordinary differential equations and discrete event simulation

The application domains in precision health are (a) cost-effectiveness analysis of AI-assisted breast histopathology and(b) cost-effectiveness of individualised cancer screening.

We are interested in both Markov models solved using ordinary differential equations (ODEs) and models with multiple time scales (e.g. attained age and time in state) using discrete event simulation (DES). For both analytical models, we would like to use time-to-event regression models for the transition intensities, and represent the uncertainty (variance) for different predictions. Predictions under different treatment or screening strategies could include: state occupation probabilities; life expectancy; quality-adjusted life-years; and costs. For the variance calculations, we could use the bootstrap or, more elegantly, using sensitivity equations and the delta method.

As a proof of concept for the ODEs, we have a very slow implementation in R (https://cran.r-project.org/web/packages/rstpm2/vignettes/multistate.pdf). Is it possible to make this more flexible and to make this several orders of magnitude faster? We are open to the choice of language (e.g. Julia, C++, a DSL or Modelica - https://github.com/mclements/modelica_markov).

As a proof of concept for the DES, we have an implementation that uses R for pre- and post-processing and C++ for the discrete event simulation (https://CRAN.R-project.org/package=microsimulation). Is it possible to make this more flexible and to make this scale across computational nodes?

For the time-to-event models, we have several classes of flexible parametric survival models (https://CRAN.R-project.org/package=rstpm2). We would also like to implement phase-type distributions, particularly to represent time-in-state for the Markov models.

Short bio

Mark Clements is an academic biostatistician at KI with interests in computational aspects for biostatistics and health economics, including time-to-event models, multi-state models, cancer screening and cost-effectiveness.

https://staff.ki.se/people/mark-clements
https://github.com/mclements
https://scholar.google.se/citations?user=l6xk21wAAAAJ&hl=en

Erik Fransén

Email: erikf@kth.se
Title: Professor
Organization: KTH

2

Research interest

Life Science community
Area: Cell and molecular biology, Precision medicine and diagnostics
Categories: Have a research tool or approach and want to find a collaborator with an idea or data to apply it to

Abstract

Keywords: machine learning, temporal data, feature extraction, temporal signatures

Machine learning analysis of temporal biomedical data

Many experimental approaches generate temporal biomedical data, including time-lapse microscopy, repeated biochemical analysis (temporal sampling/screening) of protein levels, etc. Implicit or explicit it is assumed that the sequence contains important information, cell states, how proteins interact in intracellular signaling networks, which can be used for classification or prediction, and more generally to contribute to the understanding of the biological system. We use machine learning methods to analyze biomedical data (currently magnetoencephalography data and eye movement data from Parkinson's patients) aiming to extract temporal signatures differentiating the patients from healthy controls. We are seeking experimental partners generating temporal data with an interest to collaborate with a partner who can contribute to temporal data analysis.

Short bio

Erik Fransén a professor of Computer Science at KTH Royal Institute of Technology, Sweden. He is also a visiting professor at Edinburgh University, Centre for Clinical Brain Sciences and is a StratNeuro faculty at Karolinska Institutet. He leads research projects in Computational Neuroscience and Computational Biology. He holds a BSc in physics from Uppsala university 1987, received his PhD in computer science at Stockholm University 1996 and was a postdoctoral fellow at Harvard University with Michael Hasselmo 97-98 after which he obtained a tenure-track position at KTH in 1998. He has served as vice department head, head of PhD education, member of the scientific council of Stockholm Brain Institute, and member of the OCNS board and was the main conference organizer for CNS2011.

Wojcech Chacholski

3

Email: wojtek@kth.se
Title: Professor
Organization: KTH

Research interest

WASP/ Life Science community
Area: AI/Math
Categories: Looking for open problems/questions to create new collaborations, Have a research tool or approach and want to find a collaborator with an idea or data to apply it to

Abstract

Keywords: Topological data analysis, geometry, subgrouping

Geometry and data

Topological Data Analysis (TDA) is the merging of data analysis and topology, where simplifying summaries are extracted from data while retaining meaningful information for the task at hand. TDA has been successful in various applications, particularly in Life Sciences when dealing with data that has a lot of variability, noise, or uncertainty in measurements. An example of this is a study our group as part of published in Nature Neuroscience, where TDA was used to identify sex-dependent differences in microglial morphology from development to degeneration. Our TDA methods are also effective in identifying subgroups of data based on geometry, subgroups which in a growing number of examples have interesting clinical characteristics.

Short bio

As a professor at the Mathematics Department of KTH, I lead the Topological Data Analysis group. Our research focuses on utilizing topological methods to examine spaces generated from data. We work closely with partners from Life Sciences, including the dBrain project in Digital Futures and Sandra Siegert's group at IST. Additionally, our group is a partner in the MultipleMS consortium, which is supported by the Horizon 2020 program. We also collaborate with the Brummer & Partner MathDataLab at KTH and are active members of the Altogelis network, an international collaboration between institutions such as EPFL, KTH, MPI Leipzig, Oxford, and MIT.

Roland Nilsson

Email: roland.nilsson@ki.se
Title: Senior researcher
Organization: Karolinska Institutet

4

Research interest

Life Science community
Area: Cell and molecular biology, AI/Math
Categories: Have an open project or idea that could benefit from a collaboration, Have a dataset and want to see if there is more that can be done with it

Abstract

Keywords: metabolic networks, optimization, probability theory, inverse problems, parameter estimation

Learning metabolic fluxes from large-scale isotope tracing data

Cellular metabolism, the molecular machinery that converts nutrients into useful biological products, is central to common human disorders such as diabetes, obesity and cardiovascular disease, but also to cancer and inflammatory disorders. The key variables in metabolism are metabolic fluxes, that is, biochemical reaction rates. The gold standard method for measuring fluxes is isotope tracing, where nutrients containing heavy isotopes are fed to cells or tissues, and isotope incorporation into metabolic products is measured with mass spectrometry. Metabolic fluxes can then be inferred by fitting such isotope data to a metabolic network model that quantitatively describes the relationship between fluxes and isotopes. These data sets may contain 1,000's of biomolecules and 10,000's of state variables, providing a wealth of potential information on metabolism.

However, estimating fluxes based on isotope tracing data is still difficult and time-consuming. A key obstacle is that parameter estimation in metabolic network models is poorly understood. It is not clear what optimization strategies to use, whether local optima exist, or even how to best define the objective (how to measure model error). Current software simply applies generic black-box optimization methods such as gradient descent on squared errors, with no guarantees on correctness. As a result, flux estimation is fraught with pitfalls, difficult to scale up to large networks, and difficult to automate, which limits the wider adoption of isotope tracing in metabolism research. Resolving these problems would open up possibilities of measuring human metabolism with unprecedented resolution, which would be of great value for diagnostics and drug development in metabolic disorders.

A potential WASP-DDLS project could consist of (1) developing better parameter estimation methods along the lines described above, and (2) applying these methods in proof-of-principle metabolism studies. There are several possible technical advances that could be explored. For example, while the metabolic network structure consists of layers of bilinear systems, methods for solving such systems have not been considered; and while the state variables are probability distributions, established metrics such as mutual information or divergences are not used. Our research group can provide several large isotope tracing data sets suitable for proof-of-principle studies. For example, ongoing isotope studies of cultured human liver tissue could can be investigated with these methods, aiming to better understand liver pathology in common metabolic disease.

Short bio

I am a biotechnology engineer by training with a PhD in computational biology. My research focus is on human cellular metabolism and advanced methods for measuring metabolism. Our research group at Karolinska Institutet has made contributions to one-carbon metabolism, cancer metabolism, and method development in isotope tracing and mass spectrometry data analysis. I am also a senior advisor to Sapient Bioanalytics LLC (San Diego, US), which develops next-generation mass spectrometry methods. I am broadly interested in the basic science of cellular metabolism and maintain a collaborative network of biomedical researchers to apply our methods across various fields.

Erik Sonnhammer

5

Email: erik.sonnhammer@scilifelab.se
Title: Professor of Bioinformatics
Organization: Stockholm University

Research interest

Life Science community
Area: Cell and molecular biology, Evolution and biodiversity, Software, Bioinformatics
Categories: Looking for open problems/questions to create new collaborations, Have an open project or idea that could benefit from a collaboration, Have a research tool or approach and want to find a collaborator with an idea or data to apply it to

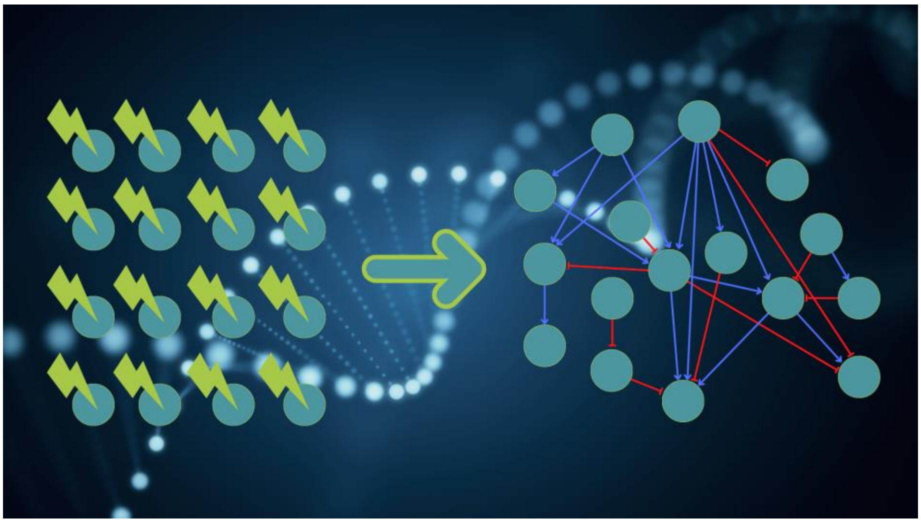
Abstract

AI models to improve data-driven inference of gene regulatory networks

Understanding the human genome requires knowledge of the regulatory circuitry that governs how encoded components are expressed and orchestrated to form living and functional cells and organisms. Only when the regulatory circuitry is fully understood can we control and avert the key components that lead to poor health.

The Sonnhammer group (life science partner) has developed a number of such methods for traditional bulk perturbation data, measured when batches of cells have reached steady state after a gene knockdown. Recent advances in gene expression control and single-cell technology have opened up new possibilities to generate higher quality and gene expression data and at a larger scale. The CRISPR technology can induce transcriptional repression or activation of a gene with CRISPRi/a. With the Perturb-seq technique such gene expression perturbations can be carried out massively parallel in single cells, from which the transcriptomic response in each cell can be obtained. This new type of data however poses new challenges, both in terms of handling the special properties of single-cell data and in terms of the larger scale which can be millions of cells in which ten thousand or more genes are perturbed.

We are looking for a WASP expert in AI and/or machine learning to collaborate with on this project. We will use either public data or generate new data with SciLifeLab's single-cell core facility. By employing AI methods we aim to develop novel algorithms to infer GRNs from single-cell perturbation data that are significantly more reliable than existing methods, which are mostly based on regression or random forests. We also aim to develop tools for simulating single-cell Perturb-seq data from a known GRN in order to benchmark the newly developed algorithm and compare it to alternative approaches and assess their predictiveness.



Short bio

Erik Sonnhammer is Professor of Bioinformatics at Stockholm University, and previously had the same position at Karolinska Institutet, Stockholm. He did a Ph.D. in bioinformatics at the Sanger Institute in Cambridge, England. His research interests are in network and systems biology as well as protein evolution and function. See http://sonnhammer.org/

Leila Ghalebani

6

Email: leila.ghalebani@gmail.com
Title: Researcher
Organization: Freelanser

Research interest

Life Science community
Area: Cell and molecular biology, Evolution and biodiversity, Precision medicine and diagnostics, Epidemiology and biology of infections
Categories: Have a research tool or approach and want to find a collaborator with an idea or data to apply it to

Abstract

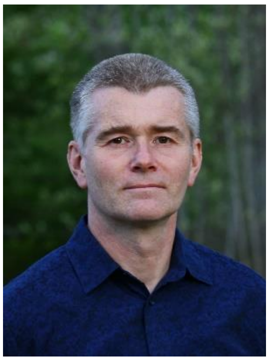
Keywords: Data-driven, Unsupervised approaches, Predictive-biomarkers, Improve-accuracy, Multi omics, subgrouping, Dementia-types, Cancer-types

Improving diagnostic accuracy, a data-driven approach

Combining multi-omics approaches, such as genomics, transcriptomics, proteomics, and metabolomics, with data-driven subgrouping in epidemiology and medicine has the potential to identify new biomarkers that improve the prediction sensitivity and specificity of diagnostic methods. Current diagnostic methods for dementia are often inconclusive and lack accuracy. An innovative combination of different unsupervised approaches, developed with over a decade of experience working with various datasets, greatly benefits clinical research efforts in identifying patterns, subgroups, and biomarkers associated with specific disease types and conditions. This is especially useful in neurodegenerative diseases and cancer, where the number of available samples for the study is often limited, and traditional methods may be less effective.

Short bio

Driven and energetic physical chemist with a lifelong passion for scientific discovery.
More than two decades of experience in research and development.
Strong computational skills in NMR relaxation & NMR Metabolomics as well as MS metabolomics.
Expert in Biological applications of ML Within Clinical & toxicological science. (Multivariate Data analytics and Multi-Omics).



Rene Kaden

7

Email: rene.kaden@medsci.uu.se
Title: Asscoc. Prof.
Organization: Uppsala University, Uppsala University Hospital

Research interest

WASP/ Life Science community
Area: Evolution and biodiversity, Precision medicine and diagnostics, Epidemiology and biology of infections
Categories: Have an open project or idea that could benefit from a collaboration

Abstract

Keywords: AI, collaboration, Method development, validation, QS

Collaboration Platform "Myrstack"

We aim to establish a project for development of an online platform that connects scientists of all ages and locations to collaborate and develop workflows for challenging methods and procedures in scientific research. The platform aims to provide a space for sharing expertise and solving complex problems across different fields and disciplines, while also promoting sustainable working practices. By facilitating cross-disciplinary collaboration, the platform helps to advance scientific research in a sustainable way and provides support for researchers new to a particular method or technique. Additionally, the platform could also promote work-life balance for scientists by allowing for more flexible collaboration and remote working. The ultimate goal of this project is to create a valuable resource for the scientific community that promotes collaboration, knowledge-sharing, and the advancement of science in a sustainable way.

Short bio

Assoc. Prof. Medical Microbiology, Uppsala University

SciLife Group leader "Epidemiology Taxonomy and Evolution"

Gullstrand fellow, Uppsala University Hospital

Coordinator GMS- Infectious Diseases Region Uppsala

Work package leader “Microbiology”, Clinical Genomics Uppsala, SciLife Laboratory

Natalia Rivera

Email: natalia.rivera@ki.se
Title: Assistant Professor
Organization: Karolinska Institutet

8

Research interest

Life Science community
Area: Precision medicine and diagnostics
Categories: Looking for open problems/questions to create new collaborations, Have an open project or idea that could benefit from a collaboration

Abstract

Keywords: biomarkers; genomics; complex diseases; genetics

Clinical biomarkers for sarcoidosis based on genomic signatures

Sarcoidosis is a complex systemic disease of unknown etiology. The hallmark of the disease is noncaseating granulomas in the affected organ. At present, there are no biomarkers for diagnosing the disease or predicting its disease course.

The MESARGEN – Multi-Ethnic Sarcoidosis Genomics Consortium is an international initiative that constitutes multiple population cohorts dedicated to genomic studies of sarcoidosis. Population cohorts in MESARGEN are well-characterized with enriched clinical data, large sample sizes, and diverse ancestries. Using this resource, we aim to identify biomarkers for diagnosis and prognosis of the disease and elucidate disease mechanisms by integrating different molecular phenotypes using different omics technologies.

Short bio

My PhD is in molecular medicine and genetic epidemiology. During my doctoral studies, I gained extensive experience working with different consortia for investigating complex diseases and traits using genome-wide association studies and meta-GWAss. At KI, I am a junior faculty, and my research focus is on the genetics and epigenetics of immune-mediated diseases, particularly sarcoidosis and rheumatoid arthritis.

Jenny Hansson

9

Email: jenny.hansson@med.lu.se
Title: Senior Lecturer
Organization: Lund University

Research interest

Life Science community
Area: Cell and molecular biology, Precision medicine and diagnostics
Categories: Looking for open problems/questions to create new collaborations, Have an open project or idea that could benefit from a collaboration, Have a research tool or approach and want to find a collaborator with an idea or data to apply it to

Abstract

Keywords: Proteomics, Mass spectrometry, Single cell

Using deep proteomics data to the maximum for single cell proteomics

One idea I have is to try to tackle the hurdle of data coverage of single cell proteomics by taking advantage of AI in some way.

This would be highly valuable in many life science sub-fields. In my lab we could apply it to the field of haematopoiesis and leukaemia (I have leukaemia mouse models that we can use), where it has the chances to resolve a current major hurdle for the development of novel therapeutics tailored to the biology of the disease (precision medicine).

I am also interested in other projects that could use deep proteomics data to answer other questions of life science.

Short bio

I have a PhD in proteomics (from 2011) and have since my postdoc applied advanced mass spectrometry-based proteomics to different important questions relating to blood cell development and particularly leukaemia. I started my group at LU in 2016 and am senior lecturer since Aug 2022.

I would contribute with expertise in advanced mass spectrometry-based proteomics and its application to life science questions. My group can generate high quality proteomics data from tissue, primary cells, blood, etc. We have the expertise and infrastructure to isolate single primary mouse cells by FACS, to prepare the samples for proteomic analysis, and to generate and analyse proteomics data.

Lukas Käll

Email: lukas.kall@scilifelab.se
Title: Professor
Organization: KTH/CBH & SciLifeLab

10

Research interest

WASP/ Life Science community
Area: Cell and molecular biology, Precision medicine and diagnostics, AI/MLX, Software
Categories: Looking for open problems/questions to create new collaborations, Have an open project or idea that could benefit from a collaboration, Have a dataset and want to see if there is more that can be done with it, Have a research tool or approach and want to find a collaborator with an idea or data to apply it to

Abstract

Keywords: Proteomics, Matrix Factorisation, Autoencoders, Algorithmics

Decomposition of quantitative mass spectrometry data to infer proteoforms

A human cell is thought to contain about 20 thousand protein-coding genes, however, each gene generates a large variety of proteoforms, either by splicing events or post-translational modifications (PTMs). It has been notably difficult to obtain a consensus picture of the importance of various proteoforms, as can be seen by wide variation in the estimates of the number of proteoforms which range from about a hundred thousand to several millions.

Mass spectrometry-based proteomics identifies peptides, but the identity of the peptides themselves cannot uniquely identify the causative proteoforms. However, given enough peptides appearing under a sufficiently large number of conditions, their co-appearance can be used to identify the combination of proteoforms appearing in a particular sample. The approach can be further improved by using quantitative abundance information.

We have previously designed a graphical model, Triqler, that combines the error rates from the identification process of spectra with their quantitative differences across conditions to determine gene-product level differences in protein concentrations.

However, it is possible to identify the co-appearance of causative proteoforms of a peptide mixture by decomposing peptide abundances across their genes using traditional principal component analysis. We propose implementing a system for inferring proteoforms from the quantitative measurements of peptides in shotgun proteomics data. Each peptide is encoded by its measured intensity in a particular experiment and its experimentally-independent

evolutionary conservation from the network. For the project, we can access about 3 billion spectra through our collaborative partners at EMBL-EBI.

Short bio

After a M.Sc. Engineering Physics from Uppsala University, 1994, I spent a couple of years in the industry as a programmer. I obtained a Ph.D. in Bioinformatics from Karolinska Institutet, 2006, and made a postdoctoral stay at the University of Washington, Department of Genome Sciences 2006-08. After a couple of years as Assistant Professor at Stockholm University, I joined KTH in 2011, and since 2018 I am a full professor.

My machine learning-based software for analyzing mass spectrometry-based proteomics data, Percolator, is provided together with the majority of all sold mass spectrometry equipment sold for proteomics usage.

Wen Zhong

11

Email: wen.zhong@liu.se
Title: DDLS fellow, Asst. Prof.
Organization: Linköping University

Research interest

Life Science community
Area: Precision medicine and diagnostics
Categories: Looking for open problems/questions to create new collaborations, Have an open project or idea that could benefit from a collaboration

Abstract

Keywords: precision medicine, omics, systems medicine, risk stratification

Molecular Diagnosis and Risk Stratification for Precision Medicine

Recent advances in high-throughput technologies have allowed for the simultaneous measurement of multiple molecular levels in biological systems, including genomics, transcriptomics, proteomics, and metabolomics. This has led to the development of integrative multi-omics approaches, which allow for a holistic understanding of the complex interactions between different molecules and their roles in health and disease. The main objective of this proposal is to use integrative multi-omics to understand human molecular fingerprints and develop new molecular diagnostic biomarkers for precision medicine based on data-driven strategies using the integration of in-house and publicly available datasets. This will involve the simultaneous analysis of genetics, plasma proteome, and plasma metabolome from various human health and disease cohorts.

Short bio

My research mainly focuses on the integration of multi-omics, the interplay between genetics and phenotypes, and the development of data-driven strategies/tools for precision medicine. The aim is to investigate the molecular biomarkers for the estimation of disease risks, early diagnosis of disease, stratification of drug treatment response, disease progression monitoring and the stratification of patients.

Alexandros Sopasakis

Email: alexandros.sopasakis@math.lth.se
Title: Docent
Organization: Lund University/ Mathematics

12

Research interest

WASP
Area: Epidemiology and biology of infections, AI/Math, Software
Categories: Looking for open problems/questions to create new collaborations, Have a research tool or approach and want to find a collaborator with an idea or data to apply it to

Abstract

Keywords: Mask-RCNN, SORT, YOLO, object detection

Image segmentation and object tracking

Looking for a collaborator within scilifelab. Can develop advanced machine learning methods to segment and track complex objects. Specialized loss functions can be considered depending on project needs. Typically we apply known knowledge or otherwise called "physics" about the problem together with data in order to improve the learning for the machine learning algorithm. Alternatively including such "physics" can make up for lack of sufficient data while still maintaining good accuracy for the algorithm. If interested please email.

Short bio

Mathematician, PhD from Texas A&M Univ., USA. Worked in research in mathematics at Georgia Tech, UC Berkeley, New York Univ. Research within machine learning, stochastic particle systems, traffic modeling, SIR models etc.

Walker Jackson

Email: walker.jackson@liu.se
Title: Senior Lecturer
Organization: Linköping University

Research interest

Life Science community
Area: Cell and molecular biology
Categories: Have an open project or idea that could benefit from a collaboration, Have a dataset and want to see if there is more that can be done with it

Abstract

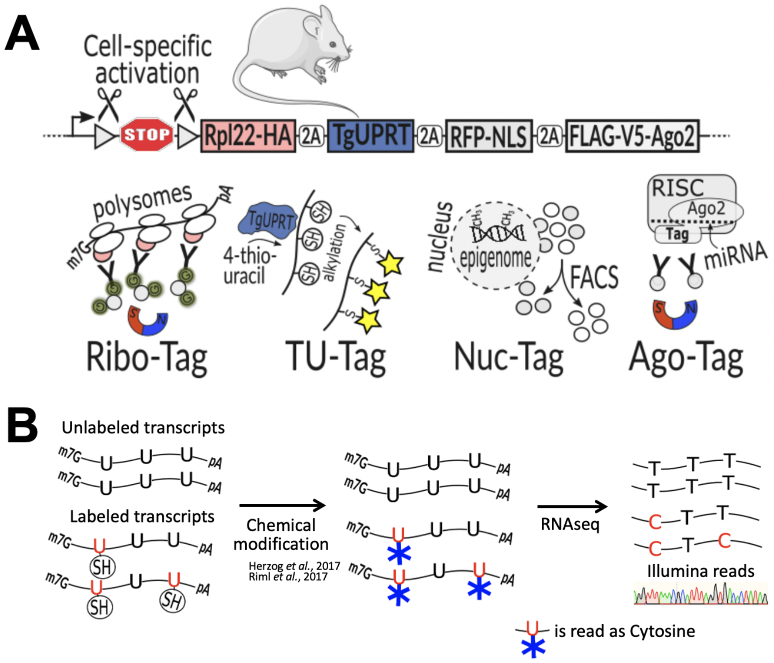
Keywords: pain, data integration, RNA dynamics, transcription, translation, RNA silencing, miRNA, epigenetics

Dissecting gene expression dynamics of pain sensing with a combination of omics and biorthogonal labeling of RNA

We recently developed a tool called Tagger [1] to aid gene expression analysis in vivo. Tagger is a transgenic mouse line expressing enrichment ”handles” for studying multiple domains of gene expression in specific cells of the body. This tool allows us to analyze multiple types of nucleic acids from specific cells, and metabolically label RNA to study rapid transcription boosts using SLAM-seq [2].

We are now applying Tagger to study pain perception, specifically to identify and functionally characterize new pain receptor cells in the peripheral nervous system. Artificial Intelligence methods are being sought to estimate synthesis and degradation rates for each transcript in the transcriptome. We are also looking for a partner to jointly develop a computational framework for data integration and analysis. Methods developed in this project may also be useful for single-cell/nucleus experiments involving metabolically labeled RNA. I am currently a co-applicant on a KAW grant to characterize a novel class of pain-sensing neurons and the project proposed here will be a tremendous contribution to that.

Figure caption: A, Tagger mice express four separate proteins in cell types of interest to enable the study of (left to right) translating mRNAs, metabolically labeled RNAs, nuclei, and miRNAs. B, Applying SLAM-seq chemistry to metabolically labeled RNA results in the conversion of U to a molecule that is read as C in NGS data. We have established a robust pipeline to read real conversions while filtering out most of the errant conversions.

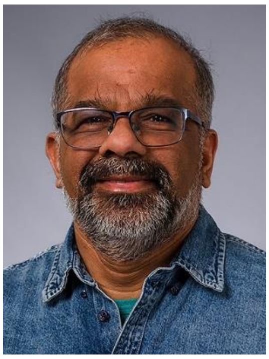


References

1. Kaczmarczyk L, Bansal V, Rajput A, Rahman R, Krzyżak W, Degen J, et al. Tag- gera swiss army knife for multiomics to dissect cell typespecific mechanisms of gene expression in mice. PLOS Biology 2019;17:e3000374. https://doi.org/10.1371/journal.pbio.3000374.
2. Herzog VA, Reichholf B, Neumann T, Rescheneder P, Bhat P, Burkard TR, et al. Thiol-linked alkylation of RNA to assess expression dynamics. Nature Methods 2017;14:1198–204. https://doi.org/10.1038/nmeth.4435.

Short bio

We study how specific cells in the nervous system responds to neurodegenerative diseases, sleep, and more recently, pain. We mostly develop and use tools that enable us to affinity purify specific types of nucleic acids from cell types of interest. I was born and educated in the US, started my research lab in Germany in 2011, and moved to Sweden in 2018.



Priyantha Wijayatunga

Email: priyantha.wijayatunga@umu.se
Title: Senior Lecturer
Organization: Umeå University

14

Research interest

WASP, Life Science community
Area: AI/Math, Probabilistic and statistical methods
Categories: Looking for open problems/questions to create new collaborations, Have a research tool or approach and want to find a collaborator with an idea or data to apply it to

Abstract

Keywords: AI, probabilistic prediction, multiple sequential data streams, sparse data, enhanced accuracy, missing data, uncertainty quantification

General Probabilistic Prediction Model for multiple Data Streams

Most of the real world phenomena have sequential data, such as time series data or spatially ordered data or ordered data in some other sense, that can be used for predictions and inference tasks; e.g., a surgeon in an intensive care unit may be interested in predicting elevated intracranial blood pressure in the brain of a patient with a severe/traumatic brain injury within next 15-minute interval, the next-next interval, etc. using patient's heart rate, respiration, ECG, etc. variations in the immediate past few hours so that he/she can administer required medical interventions in time to avoid any secondary damages or death of the patient. We have developed a simple and explainable, yet effective Bayesian probabilistic prediction model for such tasks. The model can combine many features (explanatory variables), both dynamic and static, in order to increasing the prediction accuracy of desired events. It can be argued the proposed model is useful for the clinical, biomedical, etc. Big Data. It can utilize, not only raw data directly, but also some measures, thresholds, statistics, etc. in them. Since it is a probabilistic model, it can be used even without any subject domain knowledge, i.e., it is data driven. But it can also include domain expert knowledge, if any, through Bayesian parameterization. Furthermore, it can be used for predictions with partial input data. And the model can be built from sparse and missing data due to its probabilistic nature. Also, it allows to quantify uncertainties in its predictions. A preliminary implementation of the model is shown here:

https://doi.org/10.1038/s41598-022-13732-x This model can be used in new applications in any bio-medical, etc. Big Data contexts.

Short bio

I graduated with a PhD Degree in Mathematical and Computational Sciences from the Tokyo Institute of Technology, Japan in 2007. Thereafter I worked as a postdoctoral researcher there before I joined with Umeå University in 2008. My research interests are probabilistic prediction and classification, causal inference, statistical measures of dependences, etc. I work with prediction enhancement methods, uncertainty quantification, feature extraction, etc. I am looking forward to working with scientists who wish to have statistical and computational expertise.

Laura Carroll

Email: laura.carroll@umu.se
Title: Assistant Professor
Organization: Umeå University

15

Research interest

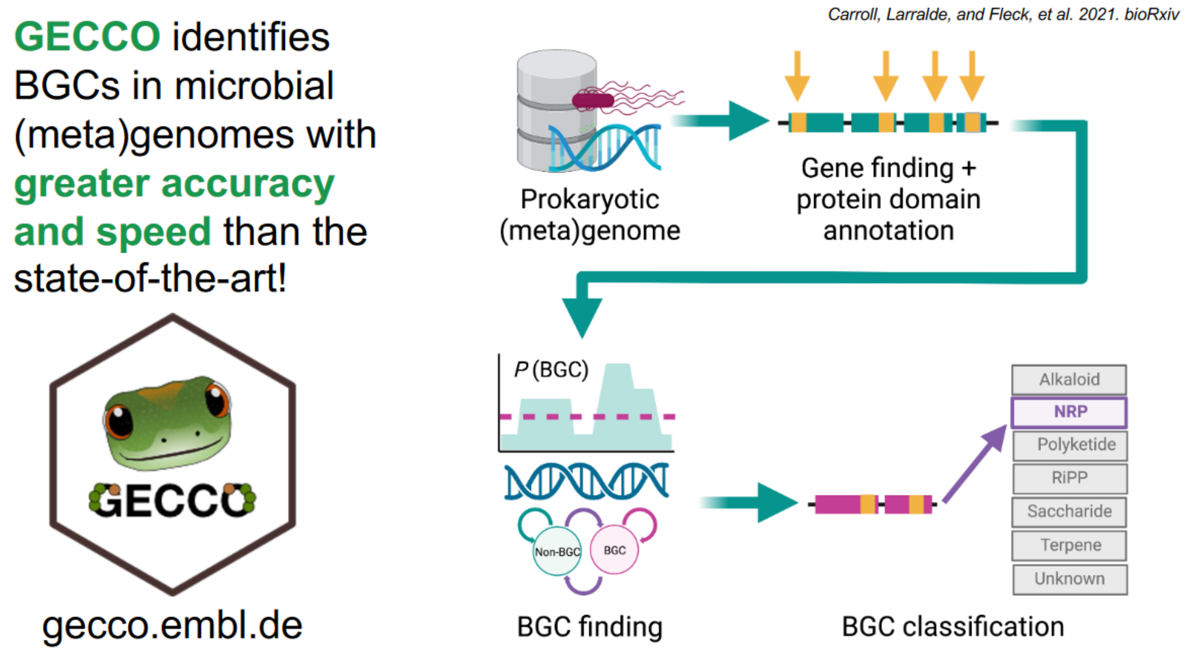
Life Science community
Area: Evolution and biodiversity, Epidemiology and biology of infections
Categories: Have an open project or idea that could benefit from a collaboration, Have a dataset and want to see if there is more that can be done with it, Have a research tool or approach and want to find a collaborator with an idea or data to apply it to

Abstract

Keywords: machine learning, natural language processing, biosynthetic gene cluster, natural product, microbiome, genomics, metagenomics, bacteria

Machine learning approaches for novel biosynthetic gene cluster discovery

Microbial biosynthetic gene clusters (BGCs) are enticing targets for (meta)genomic mining efforts, as they may be responsible for the production of novel, specialized metabolites with potential uses in medicine and industry (e.g., novel antimicrobials, anticancer agents) or roles in human health (e.g., novel toxins, carcinogens, pathogen virulence factors). I developed GECCO (GEne Cluster prediction with COnditional random fields; https://gecco.embl.de), a high-precision, scalable method for identifying novel BGCs in microbial (meta)genomic data using conditional random fields (CRFs). Based on an extensive evaluation of de novo BGC prediction, GECCO is both more accurate and faster than other state-of-the-art, machine learning-based BGC detection approaches; however, like all currently available BGC detection approaches, GECCO struggles at accurately predicting BGC boundaries (i.e., the start and end point of a BGC in a genome). This issue represents a severe bottleneck when natural product chemists use GECCO and other tools for BGC discovery, particularly in settings where GECCO is applied to large (meta)genome sets (e.g., millions of genomes), as experimentalists are forced to manually refine the boundaries of each predicted BGC before testing it in the laboratory. Thus, BGC mining approaches would benefit enormously from sequence segmentation models, which can improve cluster boundary accuracy.



GECCO identifies BGCs in microbial (meta)genomes with **greater accuracy and speed** than the state-of-the-art!

gecco.embl.de

Carroll, Larralde, and Fleck, et al. 2021. bioRxiv

Short bio

Laura Carroll is an Assistant Professor and Data-Driven Life Science (DDLS) Fellow in the Department of Clinical Microbiology at Umeå University. Her research group develops and utilizes bioinformatic approaches to monitor and combat the spread of bacterial pathogens.