Program abstracts – version 1

---

1. SLU Bioinformatics Infrastructure and UPSC bioinformatics platform as NAISS power users

Nicolas Delhomme – email nicolas.delhomme@slu.se
SLU Bioinformatics Infrastructure, Umeå

SLUBI - the Swedish University for Agricultural Sciences Bioinformatics Infrastructure; and the UPSCb - Umeå Plant Science Center Bioinformatics Platform; are facilities providing local bioinformatics support to researcher, akin to the SciLifeLab's role nationally. After an introduction about the platforms history, members and portfolio, I will present our usage of the NAISS resources, issues we've encountered and reflect on what NAISS's further development we envision would be most helpful to us in the future.

---

2. Oxymoron? When HPC Drives Evolutionary Biology into the Real World

Thanat Chookajorn – email thanat.chookajorn@slu.se

Molecular Infection Medicine Sweden (MIMS), Umeå University

Here I will describe how HPC has transformed evolutionary biology and will present use cases in pathogen and wildlife surveillance. The speed and scale of HPC have shifted the field from retrospective reports to real-time responses, raising the question of how soon it will achieve reliable, real-world predictions.

---

3. HPC for 3D protein structure determination in Cryo-Electron Microscopy

Suresh Banjara – email suresh.banjara@umu.se
Tanvir Shaikh – email tanvir.shaikh@umu.se
Linda Sandblad – email linda.sandblad@umu.se
Department of Chemistry, Umeå University
Umeå Centre for Electron Microscopy (UCEM)

Cryo-electron microscopy is a powerful technique used to study how proteins function in their natural environment, within cells and how their structures may change between healthy and diseased states. In our research we focus on the Zona Pellucida (ZP) family of proteins to understand how they assemble into larger architectures and perform diverse biological roles. These include key functions in mammalian fertilization, defense against urinary tract infection (UTI), and kidney health maintenance.

To investigate these proteins, we prepare samples by flash-freezing them in cryogen at liquid nitrogen temperature, preserving their native conformations. Imaging is performed using state of the art transmission electron microscopes such as the Titan Krios operating at 300 kV. The images are recorded with a direct electron detector to acquire spatial resolution at atomic level. We collect images with automated software followed by advanced image processing. The average per specimen images recorded during 24-hour results in a dataset of 5-10 TB stored on a local server.

Image processing is a key downstream step in Cryo-EM, transforming raw 2D images into detailed 3D protein structures. Other than our local computing workstations, our work relies on high-performance computing resources, specifically the HPC2N cluster Kebnekaise. One of the software packages used is cryoSPARC, which efficiently parallelizes jobs across both CPUs and GPUs. This computational pipeline allows us to reconstruct near-atomic resolution structures from 2D images, providing essential insights into protein function in both healthy and diseased conditions.

---

4. Unveiling a Crystal-Invisible Active Enzyme State with HPC-Powered Molecular Dynamics Simulations

Mitul Srivastava[1], Andres Garrido Aparicio[2], Yuri Schwartz[2], Anna Linusson Jonsson[1]

Mitul Srivastava – email mitul.srivastava@umu.se
Anna Linusson Jonsson – email anna.linusson@umu.se

[1] *Department of Chemistry, Umeå University, Sweden*

[2] *Department of Molecular Biology, Umeå University, Sweden*

Crystal structures often serve as a valuable starting point for computational studies, providing an experimentally verified structure. However, proteins are dynamic, and crystal structure may not represent the biologically relevant conformation. In this regard, multiscale modeling coupled with molecular dynamics (MD) simulations provides an opportunity to elucidate conformational landscape of the proteins. In our current study, we are aiming to identify biologically significant non-histone targets of clinically relevant and evolutionary conserved lysine methyltransferases (KMTs), Ash1L/Ash1/Trx. Till date, several underlying processes such as enzyme's active/inactive states, molecular recognition, structural determinants and energetics governing the methylation process, remains unexplored. At first, we were determined to identify the enzyme’s active conformation. Hence, by using large-scale MD simulations (6-7 µs) on HPC2N/NAISS resources, we mapped the Ash1/Ash1L conformational landscape and discovered a stable active conformation absent from existing structural databases and not predicted by AI-based modeling. The active state reveals formation of a groove that provides accessibility for substrate methylation which is yet not reported. Furthermore, this active conformation is now utilized to systematically identify non-histone substrates of Ash1/Trx and gain molecular insights, which are otherwise challenging to record experimentally. Our results highlight the power of HPC-driven molecular simulations that go beyond the crystallographic studies.

---

5. Capturing Adenylate Kinase Dynamics with TR-XSS and Ensemble Modelling

Konstantinos Magkakis – email konstantinos.magkakis@umu.se
Magnus Andersson – email magnus.p.andersson@umu.se
Department of Chemistry, Umeå University

Combining time-resolved X-ray solution scattering (TR-XSS) and molecular dynamics (MD) simulations, we track the conformational changes of the enzyme adenylate kinase, a metabolic sensor that monitors cellular ATP levels. To interpret the scattering data generated from TR-XSS, we generate an extensive pool of conformations by running a combination of unbiased, targeted, and bias-exchange MD trajectories. Each simulation snapshot is used to compute a theoretical scattering profile, and the Ensemble Optimisation Method (a genetic-algorithm-based search) selects small ensembles of structures whose averaged profiles reproduce the experimental difference curves. The selected ensembles are then analysed by clustering on domain-angle coordinates, enabling us to track the evolution of protein conformations during the reaction.

---

6. Molecular dynamics simulations of Calcium-Transporting P-type ATPases in membranes with different phospholipid compositions

Fatemeh Sabzian – email fatemeh.sabzian@umu.se
Magnus Andersson – email magnus.p.andersson@umu.se
Department of Chemistry, Umeå University

We use molecular dynamics simulations with GROMACS to study the structural dynamics of calcium-transporting P-type ATPases, key membrane proteins involved in calcium regulation in muscle and heart cells. We simulate these proteins in membranes with different phospholipid compositions to investigate how lipid–protein interactions influence their structural dynamics. The simulation data will be combined with synchrotron-based X-ray scattering for structural refinement.

---

7. Swedish Biodiversity in Time and Space (SweBITS)

Johan Stenberg – email johan.stenberg@umu.se
Chemistry department, Umeå University

This project leverages a unique archive of air filters, collected weekly for 60 years from multiple sites across Sweden. Originally part of the Swedish Defence Research Agency's (FOI) nuclear surveillance program, each filter processed massive volumes of air (>100,000 m³), unintentionally capturing environmental DNA (eDNA) from across the region surrounding the sampling sites. We have shown that this eDNA originates from a vast array of organisms, including birds, mammals, fungi, and even fish [1]. This archive, and airborne eDNA analysis overall, present an immense resource for a plethora of applications, such as biodiversity monitoring, invasive species detection, discovering ecological interactions, and even as a tool for investigating antagonistic biological threats.

To analyze this archive, we extract and shotgun sequence the DNA from each filter, resulting in approximately 500 million short (150 base pair) DNA sequences per filter. With 900 filters sequenced to date, our raw dataset comprises 440 billion sequences, totalling 30 TB of compressed data. The first critical step for downstream analysis is taxonomic classification: determining the organism of origin of each sequence. This is done by matching hashes of short sub-sequences from each sequence against a compact in-memory reference database of known genomes.

We are building this reference database using Kraken 2, the *de facto* standard for large-scale taxonomic classification. The database, constructed from 280,000 reference genomes, totals 15 TB and would require 9 TB of RAM to build using default settings. The largest compute nodes available to us (the "largemem" nodes on Kebnekaise) have 3 TB of RAM. To fit on the node, we were forced to compromise on the database's density, reducing its information content by 70%. While this compromise also speeds up the computation significantly, the job will still take 10 days to complete. Evidently, we are currently computationally bounded, a challenge that will only continue to grow as the number of reference genomes expand exponentially.

Beyond this initial taxonomic classification, we will continue to require the computational resources of NAISS and HPC2N for downstream analyses. These will include validation of the classification using computationally taxing alignment of the sequences against their reference genomes, time series and atmospheric transport modelling, and population-level genetic variant detection.

[1] https://www.biorxiv.org/content/10.1101/2023.12.06.569882v1

8. Scalable single-cell metagenomic analysis with Bascet and Zorn

Hadrien Gourlé, Iryna Yakovenko, Jyoti Verma, Julian Dicken, Florian Albrecht, Linas Mažutis, Johan Normark, Nongfei Sheng, Nicklas Strömberg, Tommy Löfstedt, Laura M. Carroll, **Johan Henriksson[1]****

Johan Henriksson – email johan.henriksson@umu.se
[1]Department of Molecular Biology, Umeå University, Sweden

Single-cell metagenomic sequencing (scMetaG) is the sequencing of individual bacterial genomes, making it possible to give detailed answers to what bacteria are present, how their DNA mutates, and how bad (and good) genes spread. Recent advances in microfluidics now enables up to a million genomes to be sequenced in a single experiment.

As a typical approach (using NextFlow) would result in millions of small files being created, computational infrastructure suffers: (1) SLURM job arrays are not designed for this amount; (2) Disk I/O is not designed for many small files; (3) typical R and Python scripts are too slow to process the data. Furthermore, sparse genome sequencing (as opposed to RNA sequencing) requires large amount of working memory or out-of-core approaches.

We present our solution Zorn/Bascet (http://zorn.henlab.org/) which overcomes these solutions using novel file formats and Rust implementation. We will show how specialized hardware (TPUs) may be especially suited for comparing large numbers of genomic sequences.

[1] https://www.biorxiv.org/content/10.1101/2025.06.20.660799v1

9. The energy transfer from the solar wind to the Martian ionosphere

Xiao-Dong Wang – email wang@irf.se
Institutet för rymdfysik

Mars has no global dipole magnetic field like the Earth. Therefore, the solar wind, a supersonic, magnetized and fully ionized plasma stream, can directly interact with Mars' ionized upper atmosphere---the ionosphere. This interaction manifests as the formation of an induced magnetosphere standing off the solar wind. However, the energy carried by the solar wind can be transferred to the planetary ions in the ionosphere, causing the latter to accelerate and escape from the planet. This energy transfer process drives the loss of oxygen ions from the planet, contributes to the loss of water into space, and shapes the evolutionary history of Mars.

We use Amitis, the first GPU-based hybrid (ions as particles, electrons as a fluid) particle-in-cell code to simulate solar wind interaction with Mars' ionosphere. The code is written in CUDA and follows the typical hybrid plasma equations. It runs on NVIDIA GPU nodes of Kebnekaise, HPC2N. A typical run consists 500 M particles and 500 K steps, corresponding to ~40 GB memory and 2-10 GPU-days depending on the time resolution.

10. Disease Aware Parameter-Efficient Fine Tuning of Vision Language Models

Filippo Ruffini[1] – email filippo.ruffini@unicampus.it
Paolo Soda[2] – email paolo.soda@umu.se
[1]Unit of Artificial Intelligence and Computer Systems, Università Campus Bio-Medico di Roma, Italy,
[2]Department of Diagnostics and Intervention, Umeå University, Sweden

Our research project investigates parameter-efficient fine-tuning (PEFT) strategies for adapting large vision-language models (VLMs) to the task of automated radiology report generation from chest X-rays. By incorporating disease-specific knowledge, we aim to make themodels optimized to generate clinically accurate reports from the CXRs.

We apply methods such as LoRA to fine-tune small sized VLM models (i.e., Gemma3) on medical multimodal radiology datasets while maintaining low computational overhead. To accommodate the high memory requirements of processing high-resolution images alongside language model components, we employ multi-node distributed training using DeepSpeed ZeRO-3 and DistributedDataParallel across A100/A40 GPU nodes . The implementation, based on the Hugging Face ecosystem and deployed on NAISS infrastructure (Alvis), enables scalable training workflows through optimized memory management and distributed coordination, addressing core computational challenges in multimodal medical AI.

11. Solving Partial Differential Equations on an HPC cluster

Balaje Kalyanaraman[1] – email balaje.kalyanaraman@umu.se
Siyang Wang[2] – email siyang.wang@umu.se

[1]Department of Computing Science, Umeå University
[2]Department of Mathematics and Mathematical Statistics, Umeå University

Numerical methods are popular techniques to obtain approximate solutions to partial differential equations since closed-form exact solutions are difficult to compute, particularly in real-life applications. Finite element methods are a class of successful numerical techniques to solve a variety of partial differential equations. Typically, these methods rely on functions called the basis functions, and a combination or a blend of these basis functions expresses the approximate solution. Depending on the application, we calculate the basis functions using a different set of governing equations, which often leads to a method with good approximation properties combined with computational efficiency. In this talk, I will introduce two illustrating examples and then discuss potential challenges associated with implementing the methods on an HPC cluster.

---

12. Mixed and lower precision algorithmic design for CFD and other applications

Roman Iakymchuk – email roman.iakymchuk@umu.se

Department of Computing Science, Umeå University

I am working on mixed and lower precision algorithmic design for CFD and other applications. This work gives us not only faster time to solution but even better energy footprint of applications. Therefore, I am interested in a possibility to measure energy consumption on Kebnekaise and other NAISS clusters (currently possible on Dardel) as well as a possibility to change frequency of CPU and memory (currently not possible at all).
