Home

Welcome to the knowledge pool for the danish EuroHPC competence center. Here you will find best practices, material about HPC, HPDA and AI software and hardware, Analysis of the danish national HPCs infrastructure, relevant resources and literature. All the material can be found in pdf form for offline consultation by clicking on the symbol on the top-right corner of this page.

This knowledge pool is meant to be available both to completely new users and advanced/expert users. If you are a new user, visit the section *HPC 101* to learn the basics about High Performance Computing, and the available systems and softwares you can use. Facilitation in accessing HPCs and dissemination within the project is offered - if any is needed, feel free to contact samuele@chem.au.dk.

If you desire a specific topic or material to be covered, or need any type of assistance related with the topics of the knowledge pool, contact the HPC facilitator at samuele@chem.au.dk.

This documentation is part of the EuroHPC Competence Center in Denmark, managed by DeiC.dk.

Best practices for HPC

This page lists some useful best practices to keep in mind when coding and running applications and pipelines on an HPC.

Code coverage, testing, continuous integration

Every time we code, testing is a concern and is usually performed by the coder(s) regularly during the project. One can identify some basic main types of test:

Regression test

Given an expected output from a specific input, the code is tested to reproduce that same output.

Unit test

Tests the smallest units of the software (e.g. single functions) to identify bugs, especially in extreme cases of inputs and outputs

Continuous integration

A set of tests for the software runs automatically everytime the software code is updated. This is useful to spot bugs before someone even uses the code.

More things one might need to test are the performance/scalability of the code, usability, response to all the intended types of input data.

Unit and regression test can be useful, but at some point not really feasible, since the code can scale to be quite large and complex, with a lot of things to control. It is thus a good practice to use continuous integration, and implement simple but representative tests that cover all the code, so that bugs can be spotted often before the final users do that. Code coverage tools to implement such tests exists for several programming languages, and also for testing code deployed on Github version control.

Links

Link	Description
pyTest	a package to test `python` code
Cmake	to test both `C`, `C++` and `Fortran code`
Travis CI	tool for continuous integration in most of the used programming languages. Works on Git version control.
covr	test coverage reports for R

Code styling

An important feature of a computer code is that it is understandable by other people reading it. To make this happen, a clean and coherent style of coding should be used in a project. Some languages have a preferred coding style, and in some GUI those styling rules can be set to be required. One can also use its own coding style, but it should be one easily readable by others, and it should be the same style over the whole project.

Links

Link	Description
styleguide	Google guide for coding styles of the major programming languages
awesome guidelines	A guide to coding styles covering also documentations, tools and development environments
Pythonic rules	Intoduction to coding style in python.
R style	A post on R coding style

Containerized applications

In this page we show the benefits of project and package managers, that are a way of organizing packages in separated environments. However, a higher degree of isolation than environments can be achieved by containerization. By containerizing, a user can virtualize the entire operating system, and make it ready to be

deployed on any other machine. One can for example deploy a container without the need of installing anything on the hosting machine! Note that containers are a different concept from Virtual Machines, where it is the hardware being instead virtualized.

Links

Link	Description
Docker	An open source widespread container that is popular both in research and industry
Docker course	A course to use Docker, freely hosted on youtube
Docker curriculum	Beginner introduction to docker
Docker basics	Intoduction tutorials to Docker from the official documentation page
Singularity	Singularity is another containerization tool. It allows you to decide at which degree a container interacts with the hosting system
Singularity tutorial	A well done Singularity tutorial for HPC users
Singularity video tutorial	A video tutorial on Singularity
Reproducibility by containerization	A video on reproducibility with Singularity containers

Documentation

When creating a piece of software, it is always a good idea to create a documentation explaining the usage of each element of the code. For packages, there are softwares that create automatically a documentation by using functions´ declarations and eventually some text included into them as a string.

Links

Link	Description
MkDocs	A generator for static webpages, with design and themes targeted to documentation pages, but also other type of websites. This website is itself made with MkDocs.
mkdocstrings	Python handler to automatically generate documentation with MkDocs
pdoc3	A package who creates automatically the documentation for your coding projects. It is semi automatic (infers your dependencies, classes, ... but adds a description based on your docstrings)
pdoc3 101	How to run pdoc to create an html documentation
Roxygen2	A package to generate `R` documentation - it can be used also with `Rcpp`
Sphinx	Another tool to write documentation - it produces also printable outputs. `Sphinx` was first created to write the `python` language documentation. Even though it is a tool especially thought for `python` code, it can be used to generate static webpages for other projects.

Documents with live code

Programming languages like `python` and `R` allows users to write documents that contain text, images and equations together with executable code and its output. Text is usually written using the very immediate markdown `language`. Markdown files for `R` can be created in the GUI `Rstudio`, while `python` uses `jupyter notebooks`.

Links

Link	Description
Introduction to Markdown	Markdown for `R` in `Rstudio`
Jupyter notebooks	create interactive code with `python`. You can write `R` code in a jupyter notebook by using the `python` package rpy2

Package/Environment management systems

When coding, it is essential that all the projects are developed under specific software conditions, i.e. the packages and libraries used during development (dependencies) should not change along the project's lifetime, so that variations in things such as output formats and new algorithmic implementations will not create conflicts difficult to trace back under development. An environment and package manager makes the user able to create separated frameworks (environments) where to install specific packages that will not influence other softwares outside the environment in use. A higher degree of isolation can be achieved through containers (see the related voice in this page).

Links

Link	Description
Conda	an easy to use and very popular environment manager
Getting started with conda	Introduction to `conda` setup and usage from the official documentation
Conda cheat sheet	Quick reference for `conda` usage
YARN	An alternative to `conda`

Many short jobs running

Everytime a job is submitted to the job manager (e.g. SLURM) of a computing cluster, there is an overhead time necessary to elaborate resource provision, preparation for output, and queue organization. Therefore it is wise to create, when possible, longer jobs. One needs to find the correct balance for how to organizing jobs: if these are too long and fail because of some issue, than a lot of time and resources have been wasted, but such problem can be overcome by tracking the outputs of each step to avoid rerunning all computations. For example, at each step of a job outputting something relevant, there can be a condition checking if the specific output is already present.

Massive STDOUT outputs

Try to avoid printing many outputs on the standard output STDOUT. This can be problematic when a lot of parallel jobs are running, letting STDOUT filling all the home directory up, and causing errors and eventual data loss. Use instead an output in software-specific data structures (such as `.RData` files for the `R` language) or at least simple text files.

Packaging a coding project

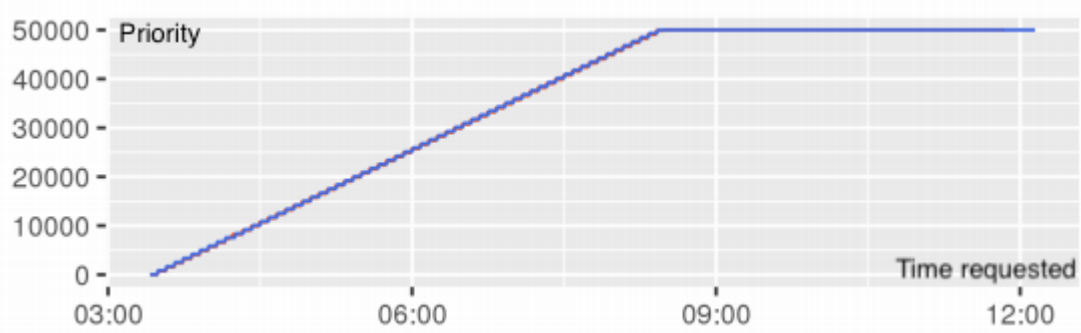
When coding a piece of software in which there are multiple newly implemented function, it can be smart to organize all those functions as a package, that can be reused and eventually shared with ease. Such a practice is especially easy and can be mastered very quickly for coding projects in `python` and `R`.

Links

Link	Description
pyPA	`python` packaging user guide
R package development	Develop an `R` package using `Rstudio`

Pipelining and submitting jobs in SLURM

SLURM is a job scheduler. It allows a user to specify a series of commands and resources requirements to run such commands. Slurm does consider the job submission on an HPC system together with all the other jobs, and prioritize them according to the resources requirement and the available computational power.



In figure above, the priority assigned to a SLURM job when the requested time increases, by keeping the memory and CPUs fixed. Decreased priority has higher values. Adapted from *A Slurm Simulator: Implementation and Parametric Analysis. Simakov et al 2017.*

The danish national HPCs, and most of the other EuroHPC supercomputers, use Slurm as job manager.

Links

Link	Description
SLURM example 1 and SLURM example 2	Some examples of how to make a Slurm script to submit a job from the danish HPC GenomeDK and from Princeton Research Computing.
Gwf, a simple python tool to create interdependent job submissions	Gwf, developed at the University of Aarhus, makes it easy to create Slurm jobs and organize them as a pipeline with dependencies, using the python language (you need python 3.5+). You get to simply create the shell scripts and the dependencies, without the complicating syntax of Slurm. The page contains also a useful guide.

Version control

Version control is the tracking of your development history for a project. This allows multiple people working on the same material to keep changes in sync without stepping over each other's contributions. Version control tools allow to commit changes with a description, set up and assign project objectives, open software issues from users and contributors, test automatically the code to find bugs before users step into them. Version control is useful for both teams and single users, and it is a good practice to have version control as a standard for any project.

Links

Link	Description
GitHub	the most used tool for version control
Github 101	quick introduction to get started on Github
GitLab and BitBucket	Two other popular alternatives to `Github`

Coding on HPC

This documentation is part of the EuroHPC Competence Center in Denmark, managed by DeiC.dk.

Software for HPC

This page contains useful links to tutorial, learning resources, benchmark and scripts for softwares usable on HPCs. Softwares are separated into broad scientific/application categories. Each tool has small numbers on its right side, indicating on which national HPC types those softwares are already installed. The symbol 🐍 means that the software is available through the conda package manager for installation, for example, on the GenomeDK Type2 HPC. Remember, you can always contact an HPC front office to inquire about the installation of packages that you need for your applications.

ML and AI

tensorflow 1️⃣ 🐍

`Tensorflow` is the Google's open source library used to develop, train and deploy machine learning models. Tensorflow is implemented for `python`, but now also available in `R`

Links

Link	Description
Tensorflow homepage	Tensorflow home page. Contains many tutorials and howtos.
Learning material	Books and courses in machine learning using tensorflow
R guide to tensorflow	How to install and use tensorflow in `Rstudio`

pyTorch 1️⃣ 🐍

`Pytorch` is another Machine Learning framework created mainly by Facebook.

Links

Link	Description
PyTorch homepage	The official `pyTorch` homepage, contains tutorials to learn how to use the framework
PyTorch book	A free book about deep learning with `pyTorch`
PyTorch for deep learning	A full video tutorial on `PyTorch`

Astrophysics and Cosmology

COSMOMC 1

CosmoMC is a Fortran 2008 Markov-Chain Monte-Carlo (MCMC) engine for exploring cosmological parameter space, together with Fortran and python code for analysing Monte-Carlo samples and importance sampling (plus a suite of scripts for building grids of runs, plotting and presenting results). The code does brute force (but accurate) theoretical matter power spectrum and Cl calculations with CAMB.

Links

Link	Description
CosmoMC homepage	The official homepage contains the documentation of `CosmoMC`
Introduction to CAMB and CosmoMC	A presentation to introduce `CosmoMC` and `CAMB`
Tutorial for CosmoMC	A lecture with usage example of `CosmoMC`
CosmoMC Paper	Reference paper of `CosmoMC`

Big data and HPDA

KNIME 1

A tool to easily read and transform data, model and visualize it, integrate data science practices and produce insights.

Links

Link	Description
Getting started	A practical example of data reading and manipulation
KNIME platform presentation	Slides listing the KNIME properties and tools
FORUM	Koin the community of users and ask/answer questions and problems
KNIME learning	A wide range of learning material to use KNIME proficiently

TRACER 1

Tracer is graphical tool for visualization and diagnostics of MCMC output. It can read output files from MrBayes and BEAST.

Links

Link	Description
TRACER documentation	Home page of `Beast2`, of which `tracer` is a tool. It contains tutorials and documentation.

Bioinformatics

ANGSD

A command line tool that integrates many different tools to analyze NGS data efficiently through native C++ implementation, multithreading support and calculation of genotype likelihoods instead of genotype calling. It is easily installed by downloading it from the `GitHub` repository and compilation of the code as per the instructions in the repository readme file.

Links

Link	Description
Github repository	Repository for download and installation
Documentation	Documentation for ANGSD

BEDtools 1

Collectively, the bedtools utilities are a swiss-army knife of tools for a wide-range of genomics analysis tasks. The most widely-used tools enable genome arithmetic: that is, set theory on the genome. For example, bedtools allows one to intersect, merge, count, complement, and shuffle genomic intervals from multiple files in widely-used genomic file formats such as BAM, BED, GFF/GTF, VCF. While each individual tool is designed to do a relatively simple task (e.g., intersect two interval files), quite sophisticated analyses can be conducted by combining multiple bedtools operations on the UNIX command line.

Links

Link	Description
Bedtools documentation	The documentation page. Contains well done tutorials and examples, both for beginners and advanced users.

Cellranger 1

The software pipeline used to align single cell sequencing data having Unique Molecular Identifier and coming as output from a 10X sequencing machine.

Links

Link	Description
What is cellranger	Description of the cellranger pipeline
Install cellranger	Installation guide
Use cellranger	User guide and tutorials

HOMER 1

HOMER (Hypergeometric Optimization of Motif EnRichment) is a suite of tools for Motif Discovery and next-gen sequencing analysis. It is a collection of command line programs for UNIX-style operating systems written in Perl and C++. HOMER was primarily written as a de novo motif discovery algorithm and is well suited for finding 8-20 bp motifs in large scale genomics data. HOMER contains many useful tools for analyzing ChIP-Seq, GRO-Seq, RNA-Seq, DNase-Seq, Hi-C and numerous other types of functional genomics sequencing data sets.

Links

Link	Description
HOMER homepage	Software homepage
Configure HOMER	Setup and configuration
Tutorial	Tutorial from the official home page

Kallisto 1

`Kallisto` is a program for quantifying abundances of transcripts from bulk and single-cell RNA-Seq data, or more generally of target sequences using high-throughput sequencing reads. It is based on the novel idea of pseudoalignment for rapidly determining the compatibility of reads with targets, without the need for alignment.

Links

Link	Description
Homepage	Software home page with tutorials and documentation
Kallistobus	Workflow to use kallisto to generate the expression data matrix
Reference	The Kallisto paper

MACS2 1

This tool identifies statistically significantly enriched genomic regions in ChIP- and DNase-seq data using the MACS2 algorithm (Model-Based Analysis of ChIP-Seq).

Links

Link	Description
Manual	MACS2 manual
Project page	Project page with installation and usage instructions
Tutorial	A tutorial for peak calling
Reference	Reference paper

Salmon

Highly-accurate & wicked fast transcript-level quantification from RNA-seq reads using selective alignment. Alevin, a tool integrated in Salmon, works on 3' tagged data, such as single cell data from 10X machines.

Links

Link	Description
Salmon	Documentation and tutorial page for Salmon
Alevin	Documentation and tutorial page for Alevin-Salmon
Homepage	Software homepage
Reference	Reference paper

SAMtools

Samtools is a suite of programs for interacting with high-throughput sequencing data. It consists of three separate repositories: - **Samtools**, Reading/writing/editing/indexing/viewing SAM/BAM/CRAM format - **BCFtools**, Reading/writing BCF2/VCF/gVCF files and calling/filtering/summarising SNP and short indel sequence variants - **HTSlib**, A C library for reading/writing high-throughput sequencing data

Links

Link	Description
Homepage	Documentation and tutorial page
Tutorial 1	Samtools tutorial
Tutorial 2	Blogpost on SAMtools

Scanpy

Scanpy is a scalable toolkit for analyzing single-cell gene expression data built jointly with anndata. It includes preprocessing, visualization, clustering, trajectory inference and differential expression testing. The Python-based implementation efficiently deals with datasets of more than one million cells.

Links

Link	Description
Homepage	Documentation page with tutorials, package documentation and access to example datasets ready to use

Seurat

Parallely to Scanpy in python, Seurat is the leading package for single cell data analysis in R.

Links

Link	Description
Homepage	Documentation page with tutorials, package documentation and access to example datasets ready to use

Seqtk

Seqtk is a fast and lightweight tool for processing sequences in the FASTA or FASTQ format. It seamlessly parses both FASTA and FASTQ files which can also be optionally compressed by gzip.

Links

Link	Description
Homepage	Documentation page with command usage
Repository and examples	Online repository with some usage examples

STAR 1

Alignment method utilizes to align sequence reads to the reference genome

Links

Link	Description
Homepage	STAR software repository
Tutorial	Online tutorial to use the STAR aligner

Computational fluid dynamics

openFOAM 1

OpenFOAM is the free, open source CFD software developed primarily by OpenCFD Ltd since 2004. It has a large user base across most areas of engineering and science, from both commercial and academic organisations. OpenFOAM has an extensive range of features to solve anything from complex fluid flows involving chemical reactions, turbulence and heat transfer, to acoustics, solid mechanics and electromagnetics.

Links

Link	Description
Homepage, openFOAM foundation	Software homepage
Forum	OpenFOAM users forum
Wiki	OpenFOAM (unofficial) wiki

ANSYS 1

Links

Link	Description
Tutorials for Undergraduate Mechanical Engineering Courses	Exercises intended only as an educational tool to assist those who wish to learn how to use ANSYS
Basics video tutorial	Tutorial for the ANSYS basics
Introduction to static structural	Video introduction to static structural with ANSYS
Simulation of 3d centrifugal pump	Video guide to simulate a centrifugal pump with ANSYS
Support resources for students	Support resources for students. Includes the possibility for a free download of the software for a tryout.
University of Alberta - ANSYS Tutorials	ANSYS tutorials from the UNiversity of Alberta

Computational chemistry

GROMACS 1

NWchem

LAMMPS 1

quantumESPRESSO 1

SIESTA 1

Development

Dash 1

Jupyter Lab 1

Matlab 1

Rstudio 1

Visual Studio code 1

Spark 1

Natural sciences and engineering

COMSOL 1

FEniCS 1

FreeCAD 1

Probabilistic programming

pyMC3

Social Sciences, Humanities

NetLogo 1

oTree 1

Literature

This documentation is part of the EuroHPC Competence Center in Denmark, managed by DeiC.dk.

Danish HPCs

This documentation is part of the EuroHPC Competence Center in Denmark, managed by DeiC.dk.

Hardware

This documentation is part of the EuroHPC Competence Center in Denmark, managed by DeiC.dk.