GWDG
Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen

Alphafold on the SCC

An application of software containers in HPC

Stefanie Mühlhausen, Pavan Kumar Siligam



16/01/2023

A protein structure is at the heart of many biological questions

Proteins are the workhorse of a cell

- Involved in every biological process in every living organism
- Amino acid sequence defines the structure of the protein
- 3D structure determines the protein's function

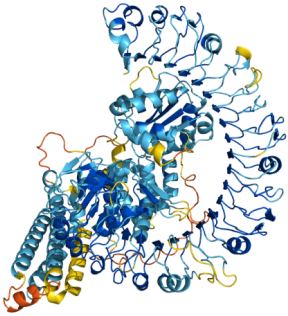
Protein folding problem

Predicting how a protein folds is one of the grand challenges in biology

AI based solutions to the protein folding problem

AlphaFold[1]

- Open source deep learning application to predict protein structures
- Regularly achieves an accuracy competitive with experiment
- AlphaFoldDB[2] contains over 2 million predicted structures



Source: https://alphafold.ebi.ac.uk/

[1] https://www.deepmind.com/research/highlighted-research/alphafold
[2] https://alphafold.ebi.ac.uk/

AlphaFold at GWDG

Software stack

- Singularity container
- Project directory
  - Required generic (sequence) databases (2TB)
  - Sample sbatch script
- GPU compute resources significantly reduce runtime

Community support

- Regular GWDG service
- RocketChat user group
- GöHPC coffee session in December (recorded)



Your next steps to use AlphaFold at GWDG

- Scientific compute cluster open to all researchers of the Max Planck Society and University of Göttingen
- Write a ticket to support@gwdg.de to get access to AlphaFold project directory

A complete sbatch example

Adjust file paths to your needs

```
#!/bin/bash
#SBATCH -job-name=alphafold
#SBATCH -partition=gpu
#SBATCH -nodes=1
#SBATCH -ntasks=10
#SBATCH -time=2-00:00:00
#SBATCH -mem=100GB
#SBATCH -G v100:2
FASTA_PATHS="/usr/users/smuehlh/alphafold_test/ins.fasta"
OUTPUT_PATH="/usr/users/smuehlh/alphafold_test"
/scratch/projects/alphafold/scripts/run_alphafold.sh -fasta_paths $FASTA_PATHS \
-output_dir $OUTPUT_PATH
```