

cd-hit

Opis

CD-HIT je široko korišten program za klasteriranje bioloških sekvenci kako bi se smanjila redundancija sekvenci i poboljšala izvedba drugih analiza sekvenci. CD-HIT je izvorno razvijen za klasteriranje proteinskih sekvenci za stvaranje referentnih baza podataka sa smanjenom redundancijom (Li, et al., 2001.), a zatim je proširen za podršku klasteriranju nukleotidnih sekvenci i usporedbu dva skupa podataka (Li i Godzik, 2006.). Web poslužitelj CD-HIT implementiran je 2009. godine, što korisnicima omogućuje grupiranje ili usporedbu sekvenci bez upotrebe CD-HIT-a iz naredbenog retka. Poslužitelj nudi interaktivno sučelje i dodatne alate za vizualizaciju.

Dostupne verzije

Verzija	Modul	Supek	Padobran
4.8.1	scientific/cd-hit/4.8.1		

Službena dokumentacija

- <https://github.com/weizhongli/cdhit/wiki>

Primjer

Primjer je napravljen po uzoru:

- <https://bioinformaticsreview.com/20190922/how-to-cluster-peptide-protein-sequences-using-cd-hit-software/>

```
#!/bin/bash

#PBS -N cd-hit-test
#PBS -l select=1:ncpus=12
#PBS -q cpu_180

cd $PBS_O_WORKDIR

module load scientific/cd-hit/4.8.1

cd-hit.sh cd-hit -i test.fa -o db100 -c 1.00 -n 5 -M 2000 -T $NCPUS
```