CD-HIT
RRID:SCR_007105
Type: Tool

Proper Citation
CD-HIT (RRID:SCR_007105)

Resource Information

URL: http://weizhong-lab.ucsd.edu/cd-hit/

Proper Citation: CD-HIT (RRID:SCR_007105)

Description: Software program for clustering biological sequences with many applications in various fields such as making non-redundant databases, finding duplicates, identifying protein families, filtering sequence errors and improving sequence assembly etc. It is very fast and can handle extremely large databases. CD-HIT helps to significantly reduce the computational and manual efforts in many sequence analysis tasks and aids in understanding the data structure and correct the bias within a dataset. The CD-HIT package has CD-HIT, CD-HIT-2D, CD-HIT-EST, CD-HIT-EST-2D, CD-HIT-454, CD-HIT-PARA, PSI-CD-HIT, CD-HIT-OTU and over a dozen scripts. * CD-HIT (CD-HIT-EST) clusters similar proteins (DNAs) into clusters that meet a user-defined similarity threshold. * CD-HIT-2D (CD-HIT-EST-2D) compares 2 datasets and identifies the sequences in db2 that are similar to db1 above a threshold. * CD-HIT-454 identifies natural and artificial duplicates from pyrosequencing reads. * CD-HIT-OTU cluster rRNA tags into OTUs The usage of other programs and scripts can be found in CD-HIT user’s guide. CD-HIT was originally developed by Dr. Weizhong Li at Dr. Adam Godzik’s Lab at the Burnham Institute (now Sanford-Burnham Medical Research Institute).

Resource Type: Resource, software resource, software application, data processing software, source code

References: PMID:20053844, PMID:16731699

Keywords: cluster, protein, sequence, classification, domain, analysis, nucleotide sequence, dna, protein sequence, bio.tools
**Parent Organization:** Google Code, University of California at San Diego; California; USA

**Funding Agency:** NCRR

**Website Status:** Last checked down

**Abbreviations:** CD-HIT

**Resource Name:** CD-HIT

**Resource ID:** SCR_007105

**Alternate IDs:** nif-0000-30240, biotools:cd-hit


**Old URLs:** http://bioinformatics.ljcrf.edu/cd-hi/

---

**Ratings and Alerts**

No rating or validation information has been found for CD-HIT.

No alerts have been found for CD-HIT.

---

**Data and Source Information**

**Source:** [SciCrunch Registry](https://scicrunch.org/registry)

---

**Usage and Citation Metrics**

We found 1518 mentions in open access literature.

**Listed below are recent publications.** The full list is available at [FDI Lab - SciCrunch.org](https://scicrunch.org).


Kalantar KL, et al. (2020) IDseq-An open source cloud-based pipeline and analysis service
for metagenomic pathogen detection and monitoring. GigaScience, 9(10).


Yang H, et al. (2020) Quyushengxin Formula Causes Differences in Bacterial and Phage
Composition in Ulcerative Colitis Patients. Evidence-based complementary and alternative medicine: eCAM, 2020, 5859023.