Genome 540 discussion

February 13th, 2025 Joe Min



Sequence alignment at scale: sequence clustering

Utilizing sequence databases: making MSAs

Sequence clustering

MMseqs2 overview

We can align 2 or 3 sequences at a time

For >3 sequences we need to be more efficient for each pairwise sequence comparison

MMseqs2 achieves "sensitivities better than PSI-BLAST at more than 400 times its speed"

How is MMseqs2 so fast?



How is MMseqs2 so fast?

"The critical insight was to combine the double-match criterion with making k-mers as long as possible ... This effectively bases our decision on up to 2 × 7 = 14 residues instead of just 2×3 in **BLAST**"





How is MMseqs2 so fast?

Steps 2 and 3 are things we've done!

- Ungapped alignment
- Smith-Waterman gapped alignment



Step 1 (preprocessing/prefiltering) is the big improvement in efficiency

Previously clustered sequence databases exist

UniProt Reference Clusters (UniRef) made using CD-HIT

- UniRef100 Combines identical UniProtKB sequences with 100% sequence identity into common entries
- UniRef90 clusters UniRef100 sequences that have
 >90% identity and 80% length overlap
- UniRef50 clusters together UniRef90 sequences with at least 50% sequence identity and 80% length overlap

Making MMseqs2-clustered databases







(3,4) Compare each sequence in group only with center



... not with all sequences in the group



(5) Sequences are recruited by center sequences into clusters



Putting it all together:

- k-mers make rough, overlapping groups
- Pick the longest sequence to represent overlapping k-mer groups (red dot)
- Cluster boundaries form where pairwise sequence identity to the representative falls below some threshold (e.g., 90%)

Making MMseqs2-clustered databases

- Due to higher sensitivity, MMseqs2 can cluster down to 30% identity, resulting in Uniclust30
- Due to better use of functional annotations, Uniclust90 and Uniclust50 clusters show higher functional consistency scores than their UniRef counterparts
- **Result:** Uniclust databases are collections of functionally similar sequences of thresholded sequence similarity

Making MSAs

What are MSAs?

Multiple Sequence Alignments (MSAs) are aligned sets of sequences

Sequences tend to be evolutionarily related

Annotations can help find alignable sequences; alignments can help make further annotations Q5E940 BOVIN ------MPREDRATWKSNYFLKIIGLLDDYPKCFIVGADNVGSKOMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE 76 76 76 76 76 76 76 76 76 75 75 76 RLAO SULAC -----MICLAVITIKKIAKWKVDEVAELTEKIKI KIIIANIEGFPADKIHEIRKKIRGK-ADIKVIKNNIFNIALKNAG-----YDIK RLAO SULTO ----MRIMAVITOERKIAKWKIEEVKELEOKIREYHTIIANIEGFPADKIHDIRKKMRGM-AEIKVIKNNIFGIAAKNAG-----DVS 79 RLAO^{_}SULTO ----MRIMAVITQERKIAKWKIEEVKELE<mark>O</mark>KLREYHTIIIANIEGFPADKLHDIRKKMRGM-AEIKVTKNTLFGIAAKNAG-80 RLAO⁻SULSO ----MKRLALALKORKVASWKLEEVKELTELIKNSNTILIGNLEGFPADKLHEIRKKLRGK-ATIKVTKNTLFKIAAKNAG-80 - IDIE RLAO AERPE MSVVSIVGOMYKREK<mark>PIPEWK</mark>TIMIRELE<mark>EIFSKHRVVIFADITGTPT</mark>FVVORVRKKIWKK-YPMMVAKKRIIIRAMKAAGIE---IDDN 86 RLAO PYRAE -MMLAIGKRRYVRTRQYPARKVKIVSEATELLQKYPYVFLFDLHGLSERILHEYRYRLRRY-GVIKIIKPTLFKIAFTKVYGG---IPAE 85 RLAO METAC -----MAEERHHTEHIPOWKKDEIENIKELIOSHKVFGMVGIEGILATKMOKIRRDLKDV-AVLKVSRNTLTERALNOLG----ETIP 78 RLAO^{_}METMA -----MAEERHHTEHI<mark>P</mark>QWKKDEIENIKELIQSHKYF<mark>G</mark>MVRIEGILATKI<mark>Q</mark>KIRRDLKDY-AVLKVSRNTLTERALNQLG----ESIP 78 RLAO ARCFU -----MAAVRGS---PPEYKVRAVEEIKRMISSKPVAIVSFRNVPAGOMOKIRREFRGK-AEIKVVKNTLLERALDALG-----GDYL 75 RLAO MAYKAKGOPPSGYEPKVAEWKRREVKELKELMDEYENVGLVDLEGIPAPOLOEIRAKLRERDTIIRMSRNTLMRIALEEKLDER--PELE 88 RLAO METTH -----MAHVAEWKKKEVQELHDLIKGYEVYGIANLADIPARQLQKMRQTLRDS-ALIRMSKKTLISLALEKAGREL--ENVD 74 RLAO METTL -----MITAESEHKIAPWKIEEVNKLKELLKNGOIVALVDMMEVPAROLOEIRDKIR-GTMTLKMSRNTLIERAIKEVAEETGNPEFA 82 RLAO METVA -----MIDAKSEHKIAPWKIEEVNALKELLKSANVIALIDMMEVPAVQLQEIRDKIR-DQMTLKMSRNTLIKRAVEEVAEETGNPEFA 82 RLA0 METJA -----METKVKAHVAPWKIEEVKTLKGLIKSKPVVAIVDMMDVPAPOLOEIRDKIR-DKVKLRMSRNTLIIRALKEAAEELNNPKLA 81 RLAO PYRAB ------MAHVAEWKKKEVEELANLIKSYPVIALVDVSSMPAYPLSOMRRLIRENGGLLRVSRNTLIELAIKKAAGELGKPELE 77 RLAO PYRHO ------MAHVAEWKKKEVEELAKLIKS YPVIALVDVSSMPAYPLSOMRRLIRENGGLLRVSRNTLIELAIKKAAKELGKPELE 77 RLA0 PYRFU ------MAHVAEWKKKEVEELANLIKSYPVVALVDVSSMPAYPLSQMRRLIRENNGLLRVSRNTLIELAIKKVAQELGKPEL 77 76 RLAU HALMA ----MSAESERKTETIPEWKOEEVDAIVEMIESYESYGVYNIAGIPBROLOMRROLHGT-AELRVSRNTLLERALDDVD--79 ---DGLE RLAO HALVO ----- MSESEVRQTEV IPOWKREEVDELVDFIES YESVGVVGVAGIPS ROLOSMRRELHGS-AAVRMSRNTLVN RALDEVN 79 79 72 72 72

Why are MSAs useful?

Huge sources of evolutionary information

- Which amino acids are used in which residue positions
- Interpositional dependencies
- Sequence/domain conservation analyses



Making MSAs

The situation:

• We have a protein sequence but don't know what it is or what it's similar to

A solution:

• Find homologous sequences in Uniclust30 to which our sequence aligns well

Making MSAs

- First, HHblits creates an HMM profile given our protein seq
- It then compares this to the

Table 1. Statistics of Uniclust databases			
Database	Clusters	Singletons	Average cluster size
Uniclust90	30.9 M	23.8 M	2.0 (5.4)
Uniclust50	13.5 M	9.6 M	4.6 (13.4)
Uniclust30	9.7 M	7.0 M	6.3 (19.8)

- HMM profiles of the representative sequences for each database cluster to find clusters of potential homology
- If pairwise alignment of any sequence in the cluster to the query exceeds a threshold, add it to the MSA

ML models use MSAs as evolutionary information

EVE uses MSAs to predict the effect of mutations



ML models use MSAs as evolutionary information

AlphaFold generates MSAs to predict structure





Reminder:

Homework 5 is due Sunday, February 16th at 11:59pm!