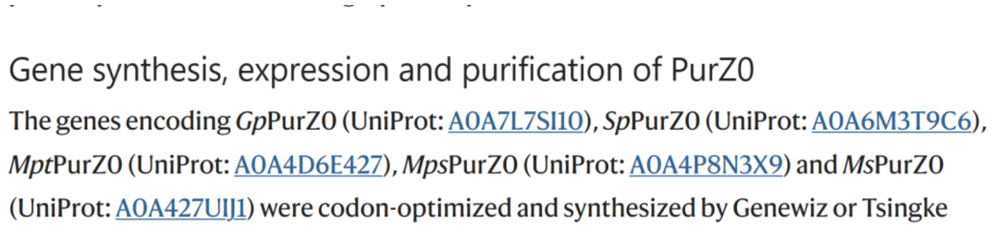
1.find PurZ0 uniprotid in
Alternative Z-genome biosynthesis pathway shows evolutionary progression from Archaea to phage | Nature Microbiology

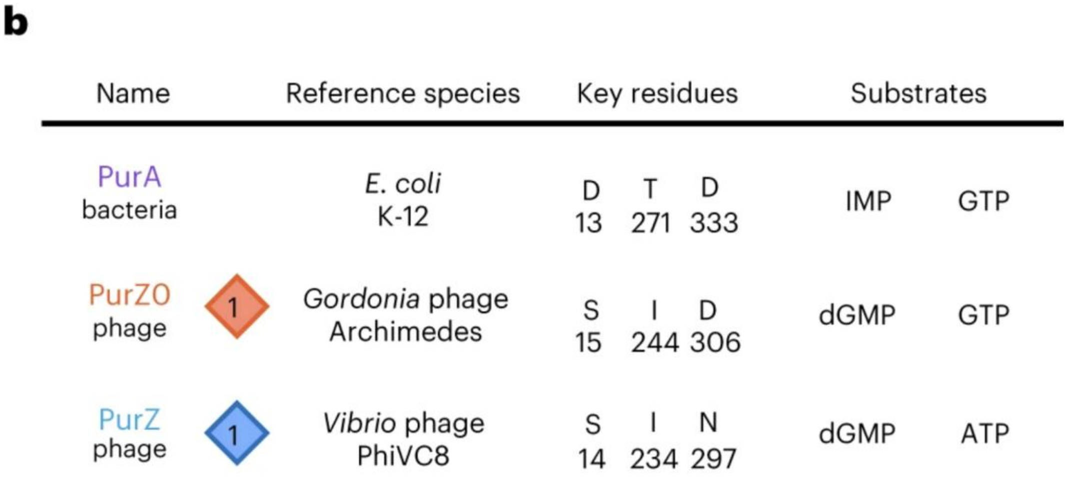


Gene synthesis, expression and purification of PurZ0

The genes encoding *Gp*PurZ0 (UniProt: A0A7L7SI10), *Sp*PurZ0 (UniProt: A0A6M3T9C6), *Mpt*PurZ0 (UniProt: A0A4D6E427), *Mps*PurZ0 (UniProt: A0A4P8N3X9) and *Ms*PurZ0 (UniProt: A0A427UIJ1) were codon-optimized and synthesized by Genewiz or Tsingke

2.

```
nohup blastp -query gppurz0.fasta -db /share/database/ncbi_nr/nr -out q2_gp_blast_results.txt -evalue 1e-5 -num_threads 10 -outfmt "6 sacc staxid qcovs pident evalue bitscore sseq qstart qend sstart send" -max_target_seqs 1000000 &
```



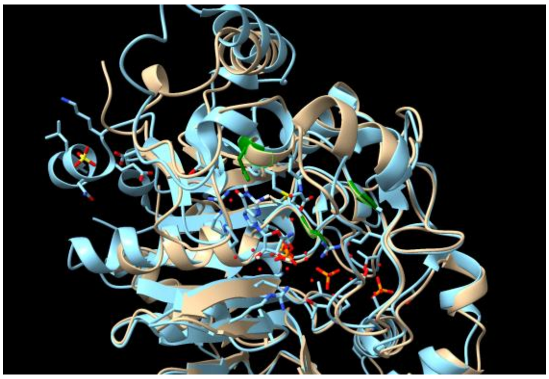
b

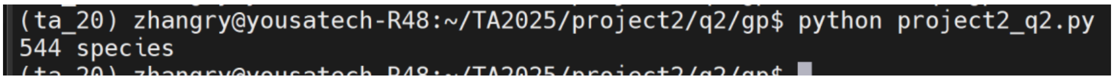
Name		Reference species	Key residues	Substrates	
PurA bacteria		*E. coli* K-12	D T D 13 271 333	IMP	GTP
PurZ0 phage	1	*Gordonia* phage Archimedes	S I D 15 244 306	dGMP	GTP
PurZ phage	1	*Vibrio* phage PhiVC8	S I N 14 234 297	dGMP	ATP

Collect complete sequences by blastdbcmd according to the ncbi accessions, the results of blastp.

Make msa of complete sequences and utilize three preserve residues to differ the truly matches from wrong ones.

# If you choose other sequence, please Match your structure predicted by the query sequence from AFDB to purz0 PDB：7vf6(structure of gppurz0). Then you will find preserve residue.





```
(ta_20) zhangry@yousatech-R48:~/TA2025/project2/q2/gp$ python project2_q2.py
544 species
(ta_20) zhangry@yousatech-R48:~/TA2025/project2/q2/gp$
```

751 real sequences of 73213 blastp results

3.

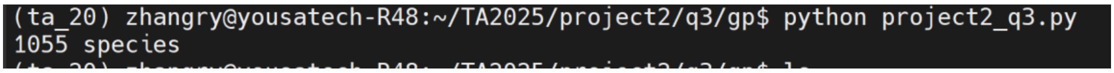
# Make database

makeblastdb -in /public/home/guest1/zry11/proj2/q3/IMG_VR/IMG_VR_2022-12-19_7/IMGVR_all_proteins.faa -dbtype prot -out /public/home/guest1/zry11/proj2/q3/IMG_VR/IMG_VR_db

# Process blastp in the database

blastp -query ../q2/q_purz0.fasta -db /public/home/guest1/zry11/proj2/q3/IMG_VR_db/IMG_VR_db -out q3_blast_results.txt -evalue 1e-5 -num_threads 10 -outfmt "6 sacc staxid qcovs pident evalue bitscore sseq"

process of filtering is similar to q2



```
(ta_20) zhangry@yousatech-R48:~/TA2025/project2/q3/gp$ python project2_q3.py
1055 species
(ta_20) zhangry@yousatech-R48:~/TA2025/project2/q3/gp$ ls
```

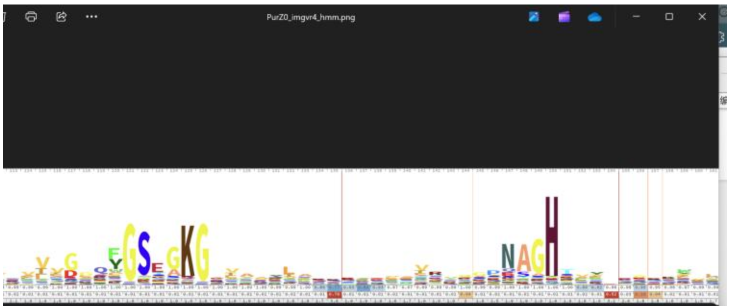
1423 real sequences of 11150 blastp results

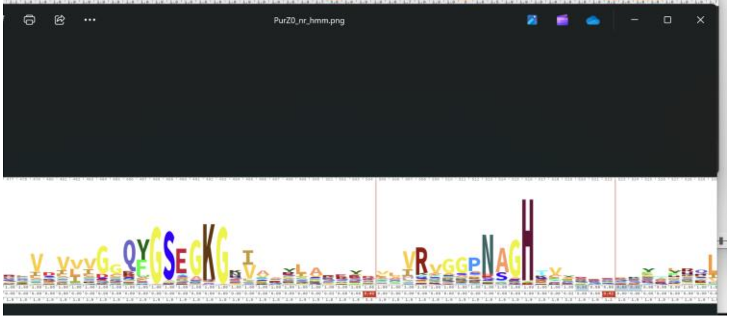
4.

In two pictures, we can find similar preserve residues pattern

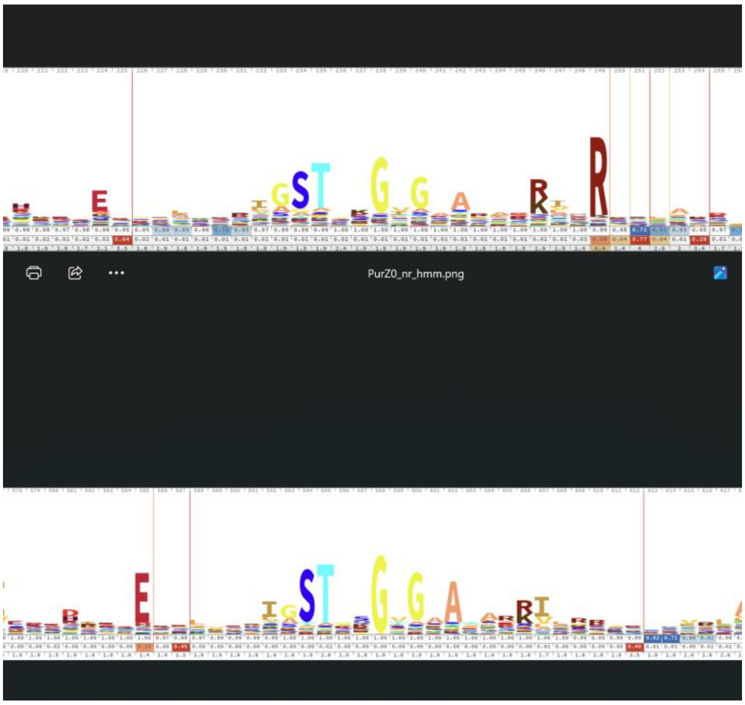
PurZ0_nr_hmm

PurZ0_imgvr4_hmm





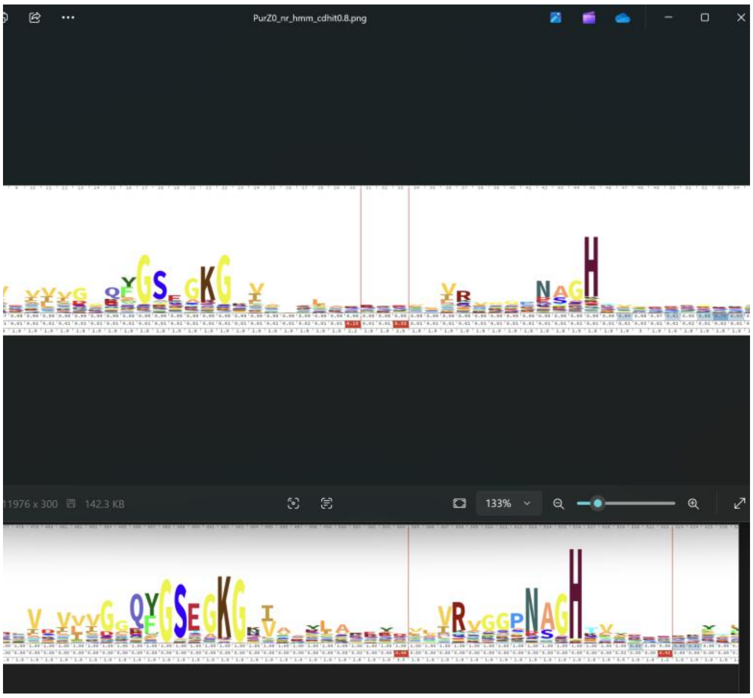
In the IMGVR sequences, the preservation of arginine (Arg) at specific positions exhibits some variation. This phenomenon is likely attributable to a bias in the redundant sequences at these positions.



5.

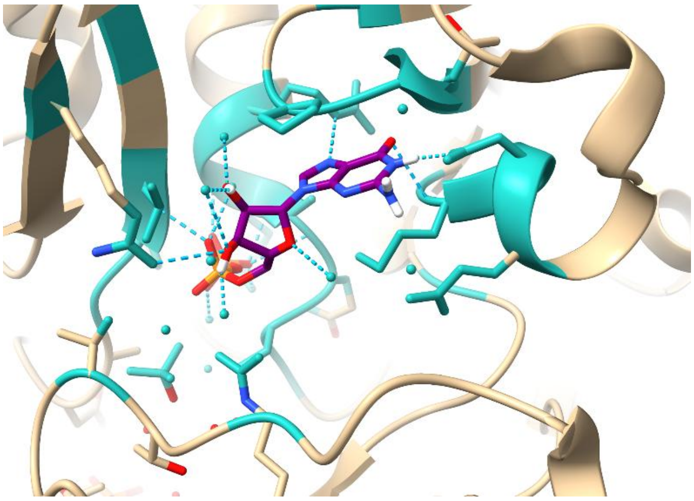
(1) There are fewer gaps at the beginning of the sequences, resulting in a more refined alignment.

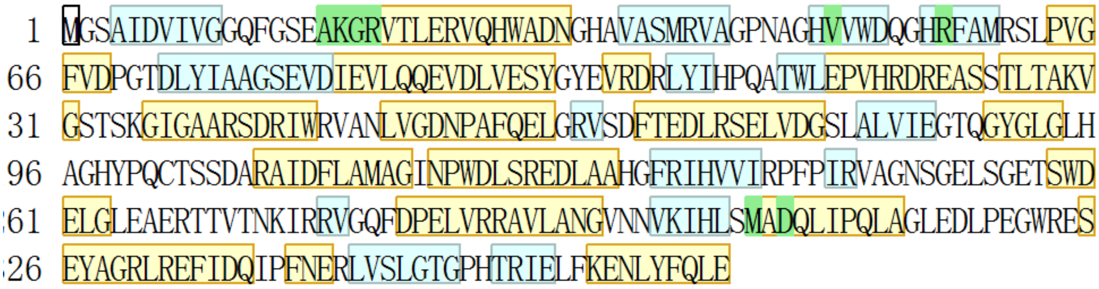
(2) The accuracy of positional preservation is enhanced by eliminating sequences with high identity, which reduces redundancy and improves clarity.



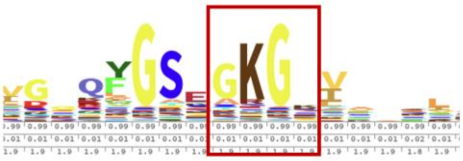
6.

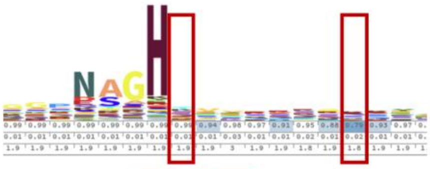
PurZ0

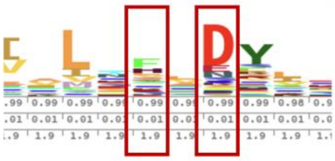




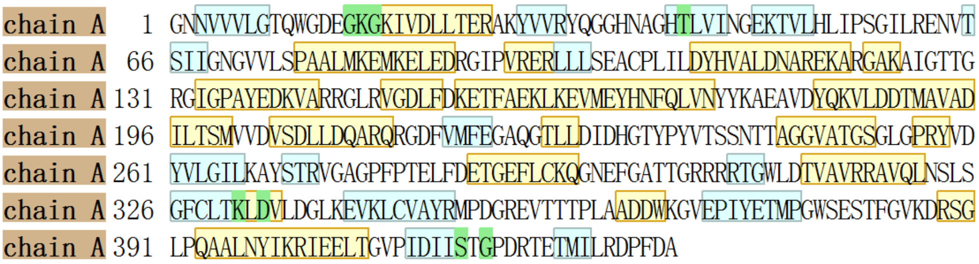
```
1  MGSAIDVIVGGQFGSEAKGRVTLERVQHWADNGHAVASMRVAGPNAGHVVWDQGHRFAMRSLPVG
66 FVDPGTDLYIAAGSEVDIEVLQQEVDLVESYGYEVRDRLYIHPQATWLEPVHRDREASSTLTAKV
31 GSTSKGIGAARSDRIWRVANLVGDNPAFQELGRVSDFTEDLRSELVDGSLALVIEGTQGYGLGLH
96 AGHYPQCTSSDARAIDFLAMAGINPWDLSREDLAAHGFRIHVVIRPFPIRVAGNSGELSGETSWD
61 ELGLEAERTTVTNKIRRVGQFDPELVRRAVLANGVNNVKIHLSMADQLIPQLAGLEDLPEGWRES
26 EYAGRLREFIDQIPFNERLVSLGTGPHTRIELFKENLYFQLE
```



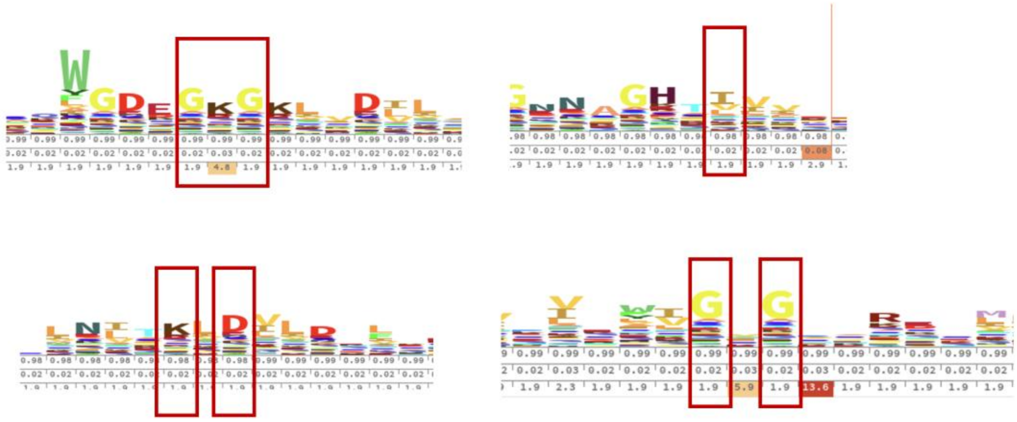




PurA



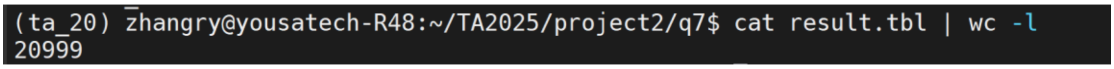
```
chain A   1 GNNVVVLGTQWGDEGKGKIVDLLTERAKYVVRYQGGHNAGHTLVINGEKTVLHLIPSGILRENVT
chain A  66 SIIGNGVVLSPAALMKEMKELEDRGIPVRERLLLSEACPLILDYHVALDNAREKARGAKAIGTTG
chain A 131 RGIGPAYEDKVARRGLRVGDLFDKETFAEKLKEVMEYHNFQLVNYYKAEAVDYQKVLDDTMAVAD
chain A 196 ILTSMVVDVSDLLDQARQRGDFVMFEGAQGTLLDIDHGTYPYVTSSNTTAGGVATGSGLGPRYVD
chain A 261 YVLGILKAYSTRVGAGPFPTELFDETGEFLCKQGNEFGATTGRRRRTGWLDTVAVRRAVQLNSLS
chain A 326 GFCLTKLDVLDGLKEVKLCVAYRMPDGREVTTTPLAADDWKGVEPIYETMPGWSESTFGVKDRSG
chain A 391 LPQAALNYIKRIEELTGVPIDIISTGPDRTETMILRDPFDA
```



7.

jackhmmer -N 5 -E 1e-5 --tblout result.tbl -o output.txt ../q2/q_purz0.fasta cleaned_file.faa

## sed 's/[^A-Za-z>\n]//g' IMGVR_all_proteins.faa > cleaned_file.faa (del the ‘-‘ in the faa)



```
(ta_20) zhangry@yousatech-R48:~/TA2025/project2/q7$ cat result.tbl | wc -l
20999
```

20999 results

8.

nohup mmseqs createdb imgvr_short.faa imgvr_v4_db &
mmseqs createdb gppurz0.fasta gppurz0_db
mmseqs search gppurz0_db imgvr_v4_db result_db tmp
mmseqs convertalis gppurz0_db imgvr_v4_db result_db result.m8

get 468 results.
More faster, less Answer.