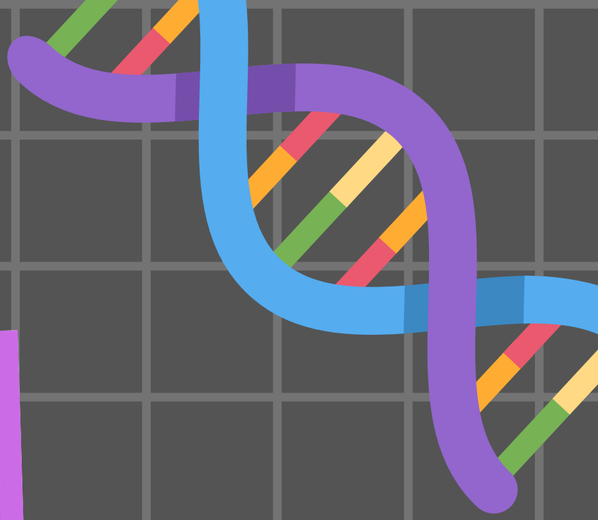


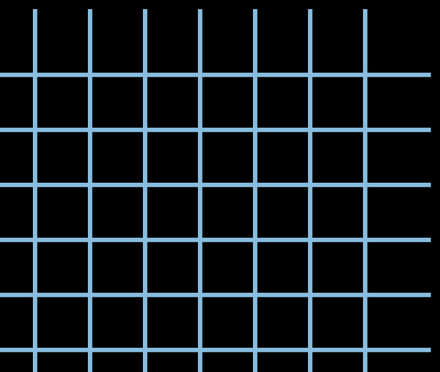
HIDDEN PATTERNS IN PROTEIN SEQUENCES

TEAM MOTIF HUNTERS

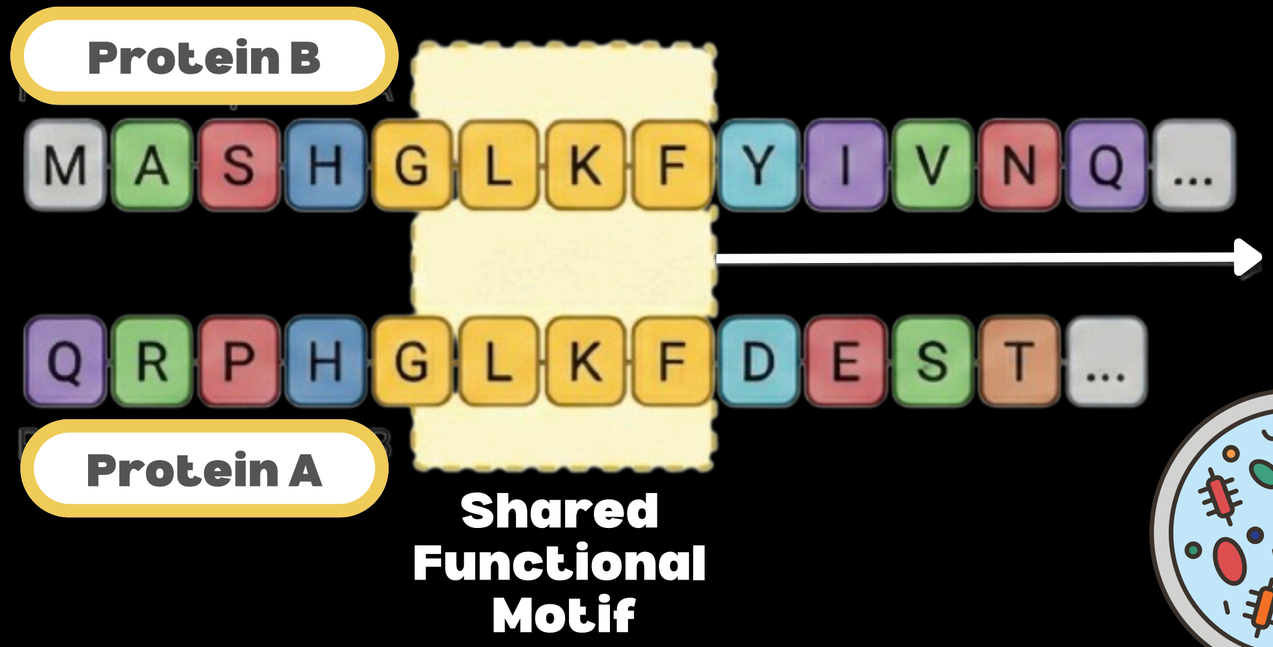
UDAY REDDY

YASH PHATAK

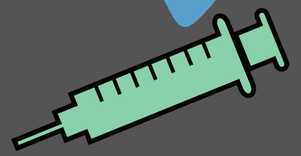
AAYUSH YADAV



Motivation



Motif Identification & Matching



Identifying unknown pathogens

Accelerating vaccine discovery

Understanding gene fucntion

Datasets

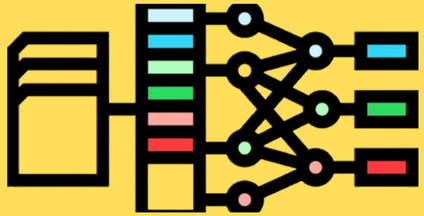
Manually curated bacterial protein sequences with high-quality annotations

Databases containing known functional motifs and protein domains

Used as the source dataset from which we sampled 100,000 proteins for analysis

Used to validate the motifs discovered from each cluster

Data Mining Pipeline



Feature Extraction

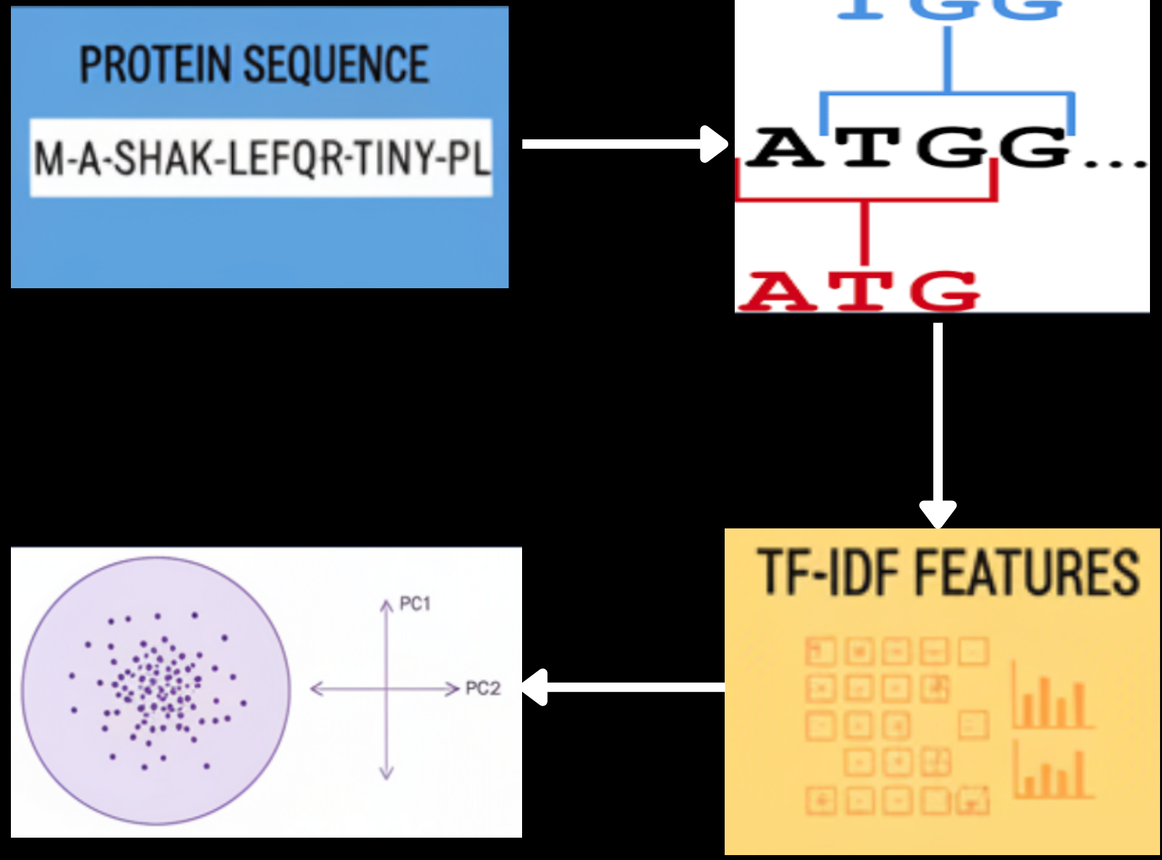
Dimensionality Reduction

Clustering

Motif Mining

Validation

Data Preprocessing



1. **Feature Extraction:**

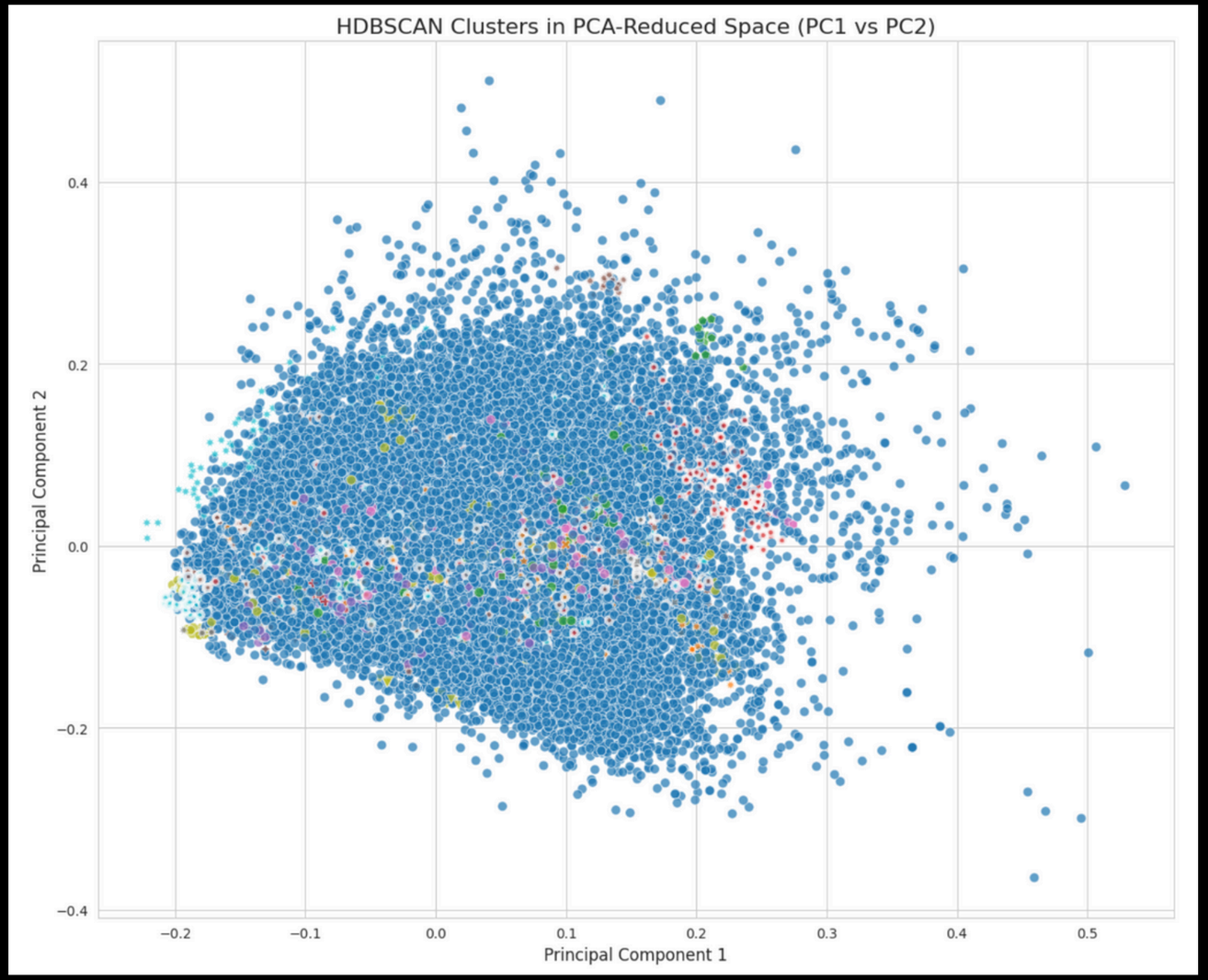
- **3-mers Preserves local biochemical context**
- **Found 9138 unique 3-mers**

2. **Compute TF-IDF scores**

3. **Dimensionality Reduction**

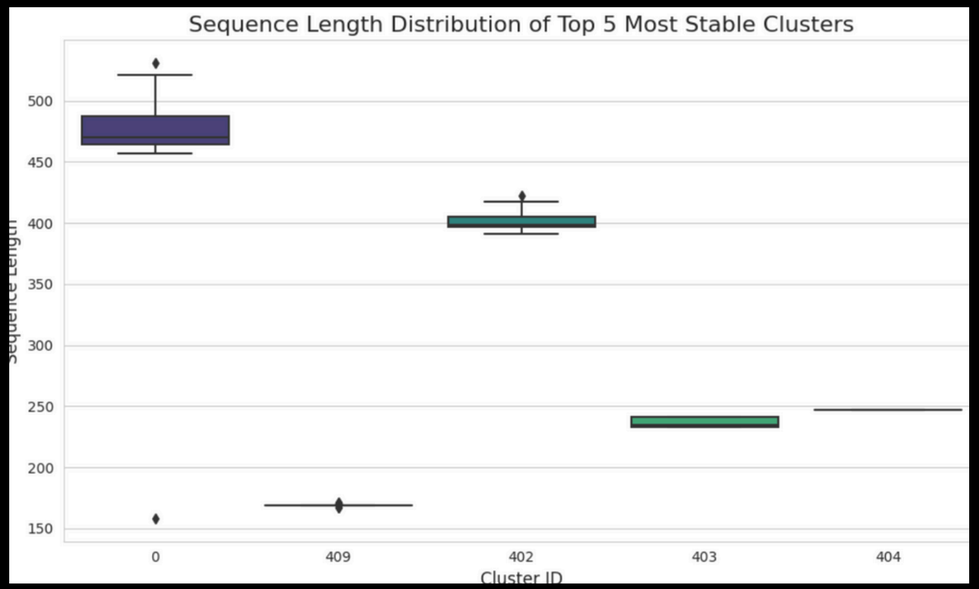
- **PCA (50 components)**
- **Captured 6% variance**

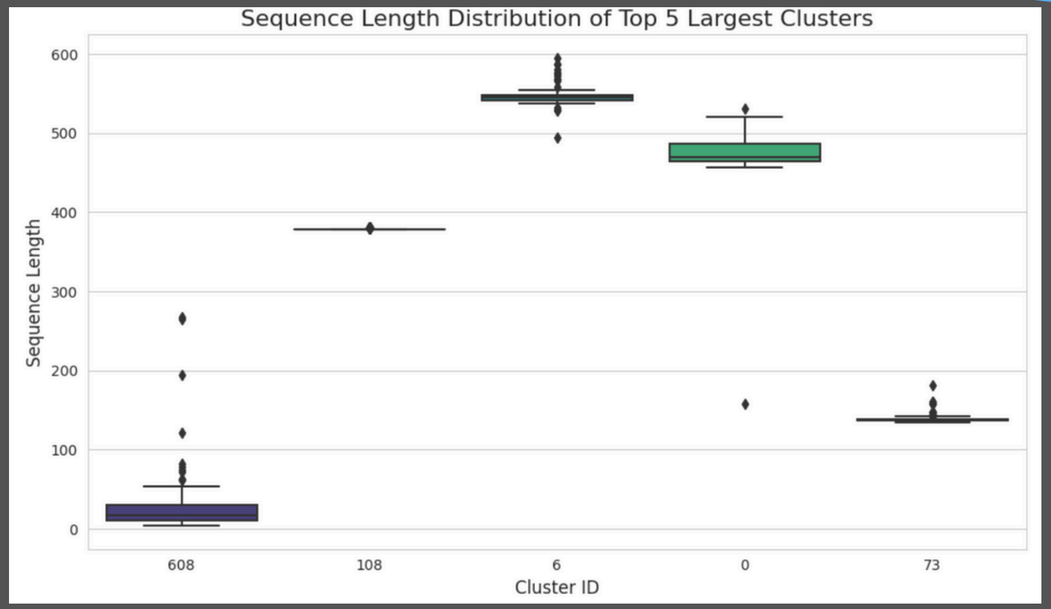
Unsupervised Clustering

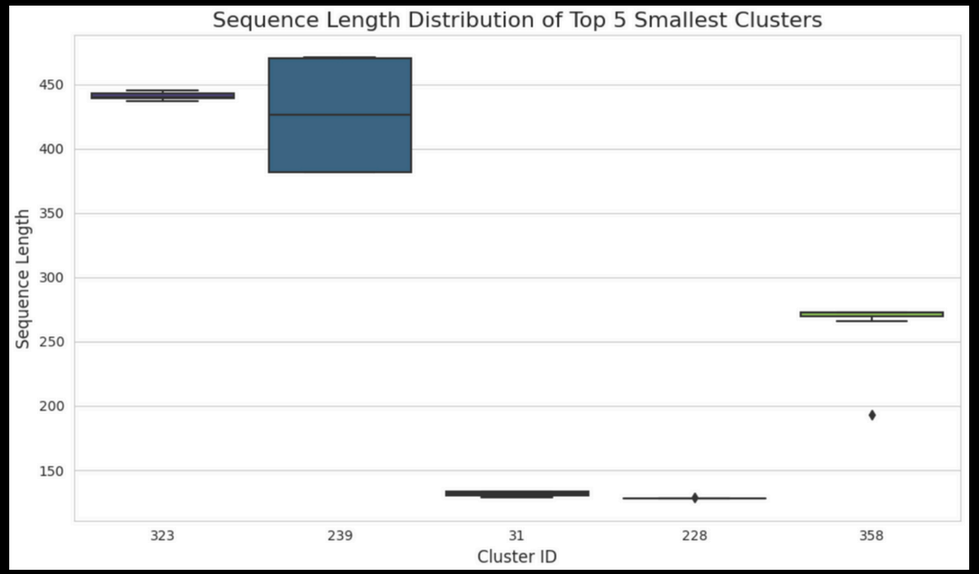


- **Used HDBSCAN to cluster proteins**
- **Discovered 609 clusters**
- **High noise proportion**

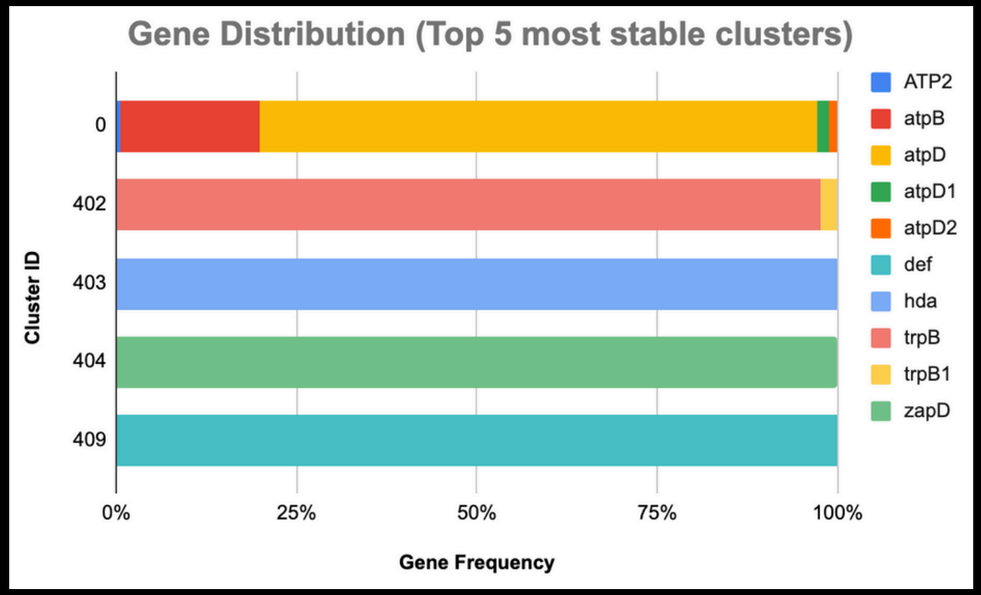
Unsupervised Clustering



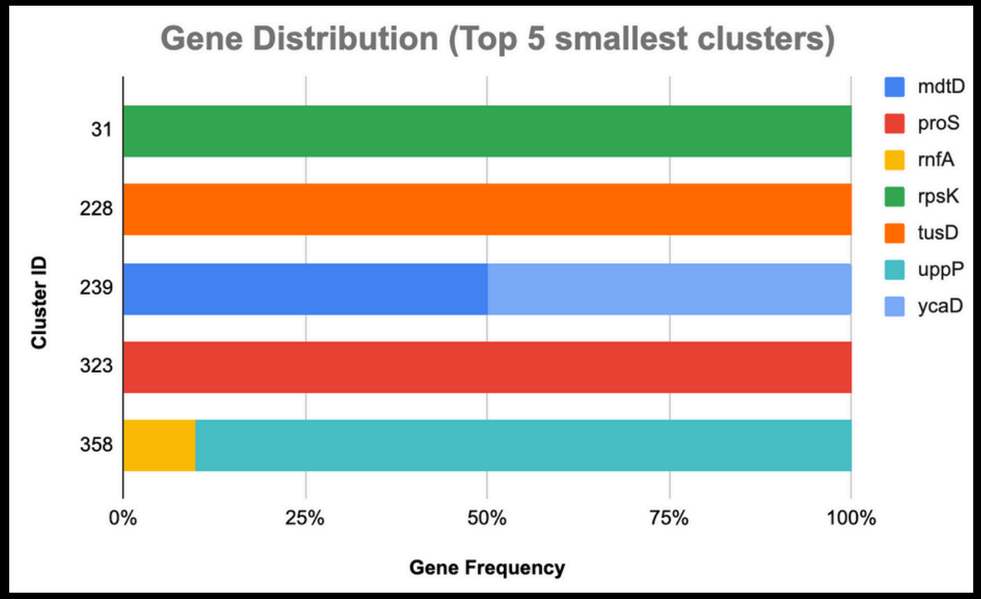


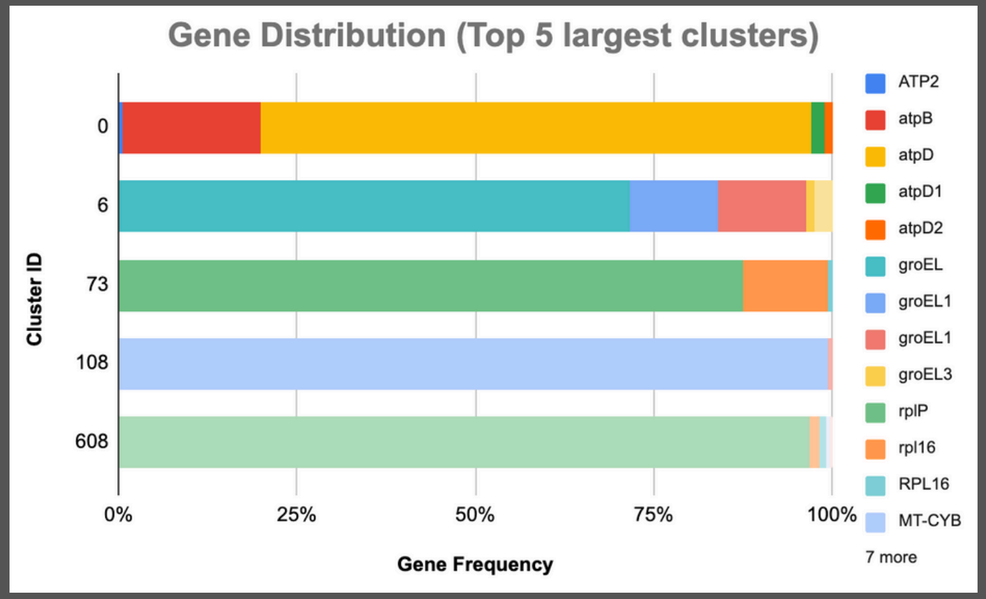


- **Protein sequence lengths in each cluster have tight bounds**
- **Sequence lengths across clusters do not overlap, making them a distinguishing factor in grouping proteins**



- **Stable clusters represent housekeeping genes essential for survival**
- **Large clusters also correspond to highly conserved protein families**
- **Small clusters are dominated by single gene implying close relationships**





Multiple Sequence Alignment

Cluster 0

Consensus Sequence (Length 599):

MVKAVMAKAATPKTTAAAEAKPAAKAPAKKAAPMRIAANPTTSTTGPAVSTLEEM
NTGKIVQVIGPVVDVEFAPPGKELMLPSVKIYNALEVKGRDTGGQKINEAECNGLT
LEVQQHLGDNVVRTIAMGSTDGLVRGMEVIDTGAPISVPVGKATLGRIFNVLGEPI
DLELGKPVGTEQSDDVERWPIHRSAPAFEELSTKTEILETGIKVIDLLAPYAKG**GKIG**
LFGGAGVGKSTVIMELINNIAKEHGGYSVFAGVGERTREGNDLYHEMKESGVINEK
NIAEESMEGGGWDLSKVDKNELKKSKVALVYGQMNEPPGARMRVALTGLTMAEY
FRDGVGEGEKDENGKGQDVLLFIDNIFRF**TQAG**SEVSALLGRMPSAVGYQPTLAT
EMGVLQERITSTKKGSITSIQAVYVPADDLTDPAPATTFAHLDATTVLSRQIASLGIY
PAVDPLDSTSRMLDPLIVGEEHYEVARGVQQTLQRYKELQDI
IAILGM**DELSEED**KLTVARARKIQRFLSQPFFVAEVFTGSPGKYVPLKETIRGFKGIL
EGEYDHLPEQAFYMVGSIDEAIEKAKKLEAEVSESKLKKLLEKARVAQASS

DELSEED

Walker A / P-loop

Walker B

Cluster 608

Consensus Sequence (Length 267):

MAAAAAALALSSFAFAGQALALSPSAIEGNGNGRIRKRAAKAKPAPSGSPW**XX**
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXSXXXX
XXXTSXKG**XXXXXXXXXXXXX**ECCCGICCGCCCCGEECCQGXFGVSCCDNGFC
RFKWCVCYRNRGCCCKYRCAGGPLCCVKDCLCPGGICDCLGLACDPEAFAEL
KVKEIKNGRLAMFSMFGFFVQAIVTGKGPLENLADHCADPVNNNAWAFATNFV
PGK

hydrophobic stretches
(typical in transporters))

Cluster 239

Consensus Sequence (Length 477):

MTELRDYTRPVLLILAFGFFLLSLAIAILNTALPLMAA**QEGESPLHMGMVS**
SSYFLGNAVGLLASGWLADKIGFNNIFFLAIFIFAAGSLFCALSGTLNQLV
LARVLQ**GLGGAMMVPVGRLTVMKIVPRAQYMAAMTFVTLPGQIGPLLG**
PALGGLLIEFASWLAWIFLAGIGCAIIGAVAESALMCNGTSRNRGFDLAA
FLLLYAVGTFLAQLLLAKDGSFLMGVLPWTVTGLTAAGILALLFTYLLHAQ
AENH**ALFSITSLFKLRQARLGLNGCFAGGIGLGMLPFLMPLFLNHGLGF**
SNAHAGFMMAPLVLAGIGGKRIIGQLADKFGYRRLLVATTLGLQLFVLIFG
SIALLGQAAMAPLALFILGAAGFTLFPSAMALACEDLEDHLASAGNQALL
LIMQLGMLIGPSFAAMLLGNFGDNHLFIDIASTSFIFLYTWLCMAVIIALPA
LIFARNAGDTPKNVAHS**RRKR**SLQ

Validation Using Pfam

Cluster 0 - ATP Synthase

	Library	Accession	Name	Domain_Description	Start_Res	End_Res	Coverage	E-value	Score
0	PFAM	PF00006	ATP-synt_ab	ATP synthase alpha/beta family, nucleotide-bin...	208	458	COMPLETE	0.0	222.7
1	SUPERFAMILY	0043650	N-terminal domain of alpha and beta subunits o...	None	55	146	N/A	0.0	N/A
2	SIGNALP_GRAM_POSITIVE	SignalP-TM	SignalP-TM	None	1	18	N/A	N/A	N/A
3	GENE3D	6fkfB03		None	461	584	COMPLETE	0.0	223.4
4	PFAM	PF22919	ATP-synt_VA_C	C-terminal domain of V and A type ATP synthase	466	553	INCOMPLETE	0.0	65.0
5	PFAM	PF02874	ATP-synt_ab_N	ATP synthase alpha/beta family, beta-barrel do...	60	144	COMPLETE	0.0	62.0

Cluster 608 - Signal Peptide (Prominent in Insects)

	Library	Accession	Name	Domain_Description	Start_Res	End_Res	Coverage	E-value	Score
0	SIGNALP_GRAM_POSITIVE	SignalP-TM	SignalP-TM	None	1	21	N/A	N/A	N/A
1	SIGNALP_EUK	SignalP-noTM	SignalP-noTM	None	1	21	N/A	N/A	N/A
2	PHOBIUS	NON_CYTOPLASMIC_DOMAIN	Non cytoplasmic domain	Region of a membrane-bound protein predicted t...	22	267	N/A	N/A	N/A
3	PANTHER	PTHR21649	CHLOROPHYLL A/B BINDING PROTEIN	None	175	267	C_TERMINAL_COMPLETE	0.0	159.9
4	PHOBIUS	SIGNAL_PEPTIDE	Signal Peptide	Signal peptide region	1	21	N/A	N/A	N/A
5	GENE3D	5xnl301	Chlorophyll a/b binding protein domain	None	153	260	COMPLETE	0.0	105.1

Cluster 239 - MFS Transporters

35	PANTHER	PTHR23521:SF2	TRANSPORTER MFS SUPERFAMILY	None	155	443	INCOMPLETE	0.0	213.1
36	PHOBIUS	NON_CYTOPLASMIC_DOMAIN	Non cytoplasmic domain	Region of a membrane-bound protein predicted t...	357	361	N/A	N/A	N/A
37	PHOBIUS	NON_CYTOPLASMIC_DOMAIN	Non cytoplasmic domain	Region of a membrane-bound protein predicted t...	218	228	N/A	N/A	N/A
38	TMHMM	TMhelix	None	Region of a membrane-bound protein predicted t...	50	69	N/A	N/A	N/A
39	NCBIFAM	NF007799	PRK10504.1	multidrug transporter subunit MdtD	1	476	N_TERMINAL_COMPLETE	0.0	565.6
40	TMHMM	TMhelix	None	Region of a membrane-bound protein predicted t...	303	322	N/A	N/A	N/A
41	PROSITE_PROFILES	PS50850	MFS	Major facilitator superfamily (MFS) profile.	11	458	N/A	N/A	N/A
42	PHOBIUS	TRANSMEMBRANE	Transmembrane region	Region of a membrane-bound protein predicted t...	102	123	N/A	N/A	N/A
43	GENE3D	2gfpA00	Multidrug resistance protein D	None	11	186	COMPLETE	0.0	137.6

Consensus Sequence

↓

PFAM Hidden Markov Model

↓

Matched Protein Families

Key Takeaways

Discovering uncharacterized proteins

Cluster 608 (Signal Peptide) demonstrates that our workflow can reveal the biological identity of “unknown” proteins by grouping them with well-characterized hormone families.

Discovering relationships between proteins

Cluster 239 (MFS Transporter) groups MdtD and YcaD together, which are homologs within the same MFS Transporter subfamily as confirmed by Pfam.

Drug Design Applications

Cluster 402 comprises of HGGAHKT, GGGSNAIGM, VIGVEPAGKGIETGMEHGA, and NQVLGQAL as the prominent motifs confirming the presence of TrpB enzyme (Drug Target for Latent TB).

THANK YOU!

ANY QUESTIONS??