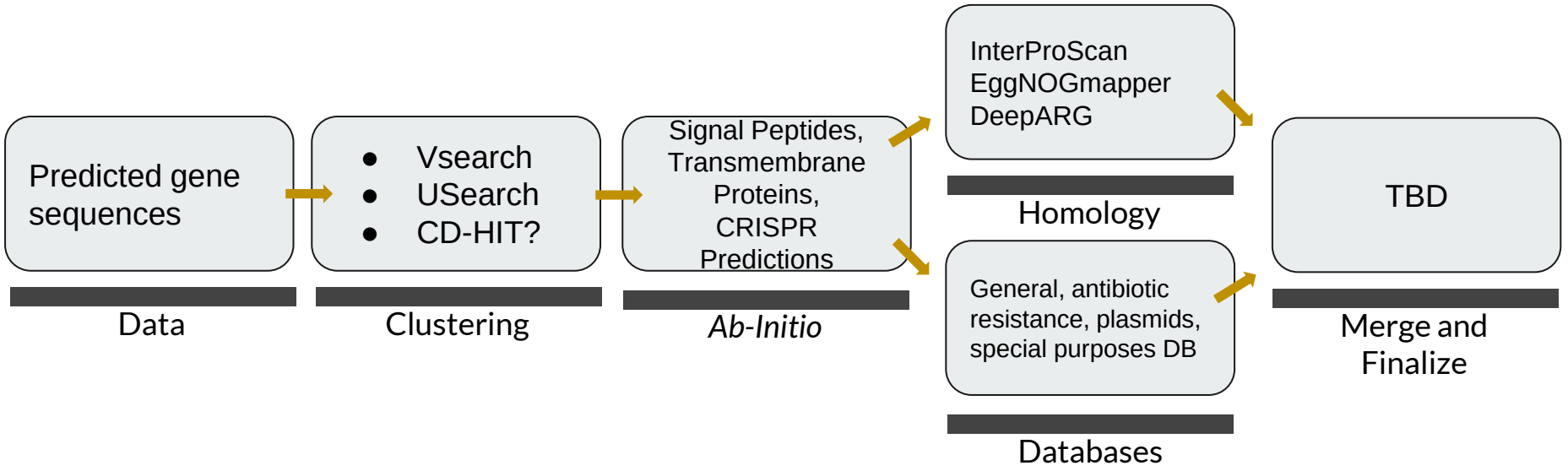
Functional Annotation: Background & Strategy

Team 1: Kenji Gerhardt, Manasa Vegesna, Shuheng Gan, Hyeonjeong Cheon, Maria Ahmad

Reasoning

- Produce annotations of genes supplied by the gene prediction team
  - RNA/Protein, location, function
  - Focus on determining virulence factors, plasmids, prophages, and ARGs
- Why?
  - Identifying the strain-distinguishing features of our microbe is key to determining its identity and phylogenetically classifying it
  - Horizontally transferred genes/plasmids are crucially relevant to strain differences
    - Conserved/shared genes are not as useful for this
  - Resistance genes are important to determining treatment approaches

Proposed Pipeline



Clustering - Goals and Measures

Goals:

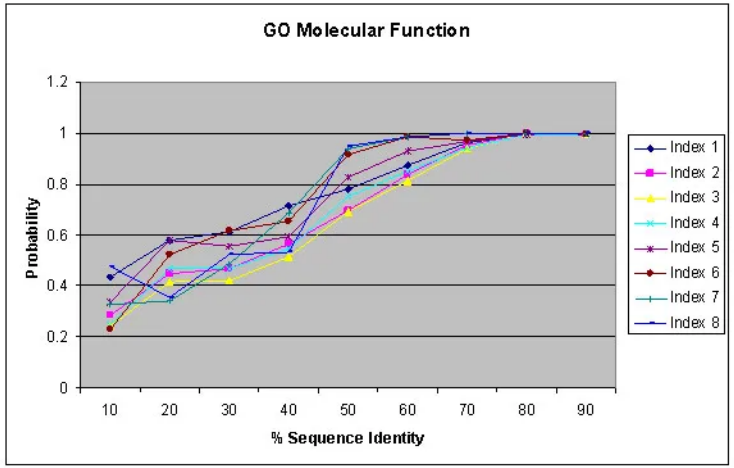
- Group similar objects together
  - Similar protein sequences
- Learn about features shared by the groups
  - Functional properties of proteins
- Use the groups/knowledge in some functional manner
  - Reduce the number of sequences that need annotated
    - increase speed of annotation

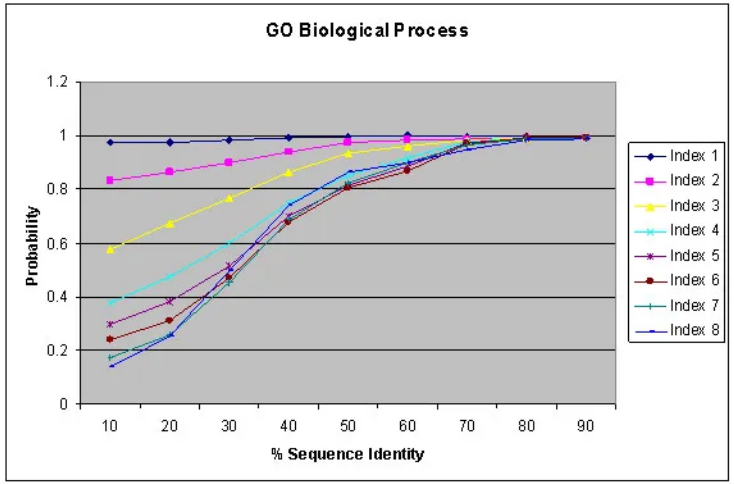
Considerations and Challenges:

- What is meant by “similar?”
  - Shared functional annotation
- How is similarity measured?
  - Sequence identity, similarity of annotation
- How are groups being made?
  - Depends on algorithm
- What defines the boundaries of a group?
  - Cutoffs/parameters
- How should the groups be used?
  - Annotate group representatives, and assign their function to the group they represent

Clustering - Proteins

- Functional properties are mostly determined by protein sequence
  - High sequence similarity ➡ functional similarity
    - *Or at least similar/identical annotations*
- Similarity of gene ontology annotations demonstrate this behavior well
  - GO terms annotate different aspects of protein location, role, and function
- Clustering reduces the number of sequences to annotate, but does not *necessarily* reduce the accuracy of annotations





Clustering - Choices

- Goal:
  - Produce cluster representatives to annotate
- Similarity measure: Sequence identity
  - Nucleotide level
  - Amino acid level
- Boundary: 90% Nucl. sequence identity makes a cluster
- Protein seq. ID.

- CD-HIT
  - Greedy-incremental
  - Sort sequences by length
  - Longest sequence is the representative of each cluster
  - Start a new cluster if the next sequence is too dissimilar
  - Very well cited, widely used
  - Scales poorly
  - Questionable representative selection
- USearch
  - Also greedy-incremental
  - Sequence order matches data; not sorted
  - Similar process from there
  - Centroid representative
- VSearch
  - Needleman-Wunsch alignment - not heuristic

Features of Prokaryotic Genome

We aim to annotate the following regions/features of the prokaryotic genome:

- **Protein-Coding Regions:**
  - Signal Peptides
  - Transmembrane Regions
  - Lipoproteins
  - Operons
- **Non-Coding RNA Regions:**
  - rRNA
  - tRNA
  - sRNA
  - CRISPR
- **Other important regions:**
  - Antibiotic Resistance
  - Virulence Factors
  - Prophage Genes

The *Ab-Initio* Approach

Ab-Initio Tools predict and annotate different regions of the prokaryotic genome using:

- Sequence composition
- Likelihoods within the gene models
- Gene content
- Signal Detection

Advantages of using *ab-initio* tools:

- No external data or evidence is needed for the prediction
- Used for finding new genes

Disadvantages:

- Presence of False Positives in the predicted data
- Over-predict small genes

Ab-Initio Tool Selection

Protein Subcellular Location Prediction:

Tool Name	Feature Prediction	Based on	Citations	Year	Overall Reason For Selection
PSORTb v 3.0	Subcellular location in Bacteria	SVMs, HMMs	1343	2010	Generates an overall prediction based on specific features and emphasizes precision. Specifically trained for gram-negative and gram-positive bacteria.

Signal Peptide Prediction:

Tool Name	Feature Prediction	Based on	Citations	Year	Overall Reason For Selection
SignalP v. 5.0	3 types of Signal Peptides	Deep Neural Networks	253	2019	Can differentiate between different types of signal peptides
LipoP	Lipoproteins and Transmembrane regions	HMM	1056	2003	4 classes of proteins are predicted and has been trained on gram-negative bacterial genomes
TatP	Tat Signal Peptides	Neural Networks	502	2005	Specialized for Tat and Sec Signal Peptide predictions

Transmembrane Protein Prediction:

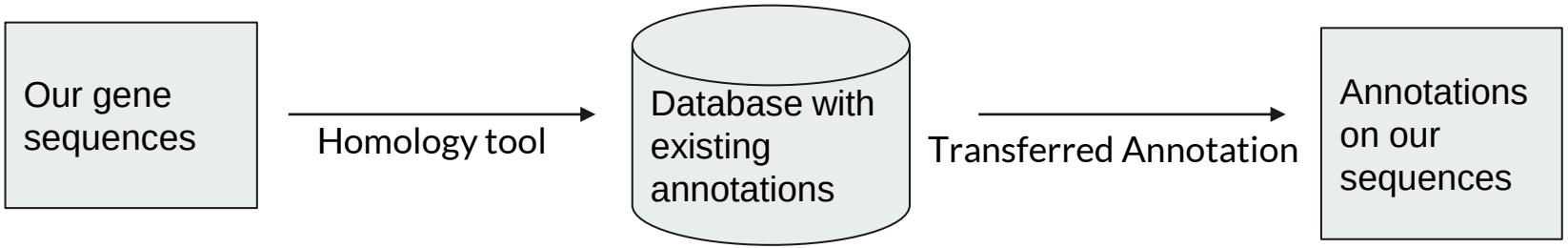
Tool Name	Feature Prediction	Based on	Citations	Year	Overall Reason For Selection
TMHMM	Transmembrane Regions	HMM	9687	2001	Also identifies soluble and membrane proteins with a high degree of accuracy. Easy to use and memory-efficient
Phobius	Transmembrane Regions and Signal Peptides	HMM	975	2007	Distinguishes N-terminal TM from signal peptides which reduces false positives

CRISPR Prediction:

Tool Name	Feature Prediction	Based on	Citations	Year	Overall Reason For Selection
CRISPRCasFinder	CRISPR regions and Cas Proteins	Maximal and Candidate repeats	95	2018	Identifies Cas Proteins along with repeat recognition. Designed for large datasets
PILER-CR	CRISPR regions	Identified repeats	1405	2007	High Sensitivity and Fast - completes a 5 Mb genome in 5 seconds
CRT	CRISPR regions	K-mer based approaches	430	2007	Fast and Memory Efficient, High Recall Rate and Quality. Faster for genomes containing larger number of repeats

Databases - Homology Introduction

- What is Homology?
  - Homology between genes means they share ancestry
  - Homologous genes that have recently diverged usually share function
    - By finding homologous genes, we're looking to transfer annotation on known genes to our predicted genes.
- Gene databases
  - Collections of annotated genes
    - Sometimes curated, sometimes not
    - Sometimes made for specific purposes
  - When we search a gene against a database, the search is looking for homology between our gene sequences and those in the database to determine what our genes' function will be



Databases - Reasonings & DB Topics

- Reliability and accuracy of functional annotation is highly dependent on reliability and accuracy of databases
- Too many databases, or too large databases = computationally $$
- Too small = ↑ probability of missing relevant annotation
    - Need *specific*, *quality* databases which *limit search size*
- Database specializations include:
    - prokaryotic operons
    - **virulence factors**
    - orthologous genes
    - CRISPR sites
    - **antibiotic resistance genes**
    - **prophage genes**
    - conserved regions
    - genetically mobile elements
    - **plasmids**
    - non-coding RNA
    - lipoproteins

Databases - Potential DB

General

- Swiss-Prot
  - Verified genomes, well-cited, 23,144 reviewed results for *E. coli* (functional information on proteins)
- Gene Ontology (GO)
  - Well-cited, 4391 genes for *E. coli*
- Others: DOOR2, EggNOG, EchoBase (specific for *E. coli*)

Antibiotic resistance

- DeepARG-DB
  - Contains the information from other databases as well: CARD, ARDB, and UnitPro
- VFDB
  - Virulence factors, 40 *E. coli* genomes, possibly not user-friendly
- Phaster
  - Prophage genes, > 14000 annotated bacterial genomes
- PlasmidSeeker
  - 8,514 plasmids from RefSeq

Homology-based tools - pros and cons

- Advantages of Homology-based tools
  - More accurate and reliable than *Ab-initio* tools
  - Can be targeted for specific purposes, e.g. antibiotic resistance genes

- Disadvantages of Homology-based tools
  - Dependent on existing annotations
  - Dependent on what databases are being searched

- General: BLAST, InterProScan, EggNOG-mapper, Prokka
- Specific: DeepARG

Homology-based tools - Tool selection

Name	Alignment	Database	Updated	Description	Advantages
InterProScan	multiple sequence alignments	HAMAP SUPERFAMILY PANTHER etc.	Nov 2019	- Combine 14 databases and 4 protein signature types - multiple sequence alignments	- powerful and sensitive - reduces redundancy - pathway information
eggNOG-mapper	HMMER DIAMOND	eggNOG (Orthologous Groups of proteins)	Jan 2019	- annotate large sets of sequences based on fast orthology assignments using precomputed clusters and phylogenies from the eggNOG database.	- infer fine-grained orthologs - faster than InterProScan (DOI: 10.1093/molbev/msx148, 2017)
Prokka	BLAST+, HMMER3	Core BLAST+ databases, HMM databases	Nov 2019	- Annotates prokaryotic genomes - Seems to predict genes	- uses a variety of databases - fast
DeepARG	DIAMOND	CARD ARDB	Sep 2017	- Annotates antibiotic resistance regions	- Machine learning solution - low false negative rate during predictions

Wrap-up

- Overall process
  - Cluster genes
  - Ab Initio prediction
  - Homology based prediction
    - Based on a selection of task-specific databases
- Merge results
  - Trust task-specific tools first
  - Agreement in homology tools second
  - Agreement in homology - ab initio third
    - Note on ab initio
- Questions?

Citations

- Clustering
  - Joshi, T., & Xu, D. (2007). Quantitative assessment of relationship between sequence similarity and function similarity. *BMC genomics*, *8*(1), 222.
  - Zou, Q., Lin, G., Jiang, X., Liu, X., & Zeng, X. (2020). Sequence clustering in bioinformatics: an empirical study. *Briefings in bioinformatics*, *21*(1), 1-10.
- Ab-Initio:
  - Nielsen, H., Tsirigos, K. D., Brunak, S., & von Heijne, G. (2019). A brief history of protein sorting prediction. *The protein journal*, *38*(3), 200-216.
  - Viklund, H., & Elofsson, A. (2004). Best α-helical transmembrane protein topology predictions are achieved using hidden Markov models and evolutionary information. *Protein Science*, *13*(7), 1908-1917.
  - Romine, M. F. (2011). Genome-wide protein localization prediction strategies for gram negative bacteria. *BMC genomics*, *12*(S1), S1.
  - Yu, N. Y., Wagner, J. R., Laird, M. R., Melli, G., Rey, S., Lo, R., ... & Brinkman, F. S. (2010). PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. *Bioinformatics*, *26*(13), 1608-1615.
  - Armenteros, J. J. A., Tsirigos, K. D., Sønderby, C. K., Petersen, T. N., Winther, O., Brunak, S., ... & Nielsen, H. (2019). SignalP 5.0 improves signal peptide predictions using deep neural networks. *Nature biotechnology*, *37*(4), 420-423.

Citations

- Juncker, A. S., Willenbrock, H., Von Heijne, G., Brunak, S., Nielsen, H., & Krogh, A. (2003). Prediction of lipoprotein signal peptides in Gram-negative bacteria. *Protein Science*, *12*(8), 1652-1662.
- Bendtsen, J. D., Nielsen, H., Widdick, D., Palmer, T., & Brunak, S. (2005). Prediction of twin-arginine signal peptides. *BMC bioinformatics*, *6*(1), 167.
- Käll, L., Krogh, A., & Sonnhammer, E. L. (2004). A combined transmembrane topology and signal peptide prediction method. *Journal of molecular biology*, *338*(5), 1027-1036.
- Krogh, A., Larsson, B., Von Heijne, G., & Sonnhammer, E. L. (2001). Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *Journal of molecular biology*, *305*(3), 567-580.
- Couvin, D., Bernheim, A., Toffano-Nioche, C., Touchon, M., Michalik, J., Néron, B., ... & Pourcel, C. (2018). CRISPRCasFinder, an update of CRISRFinder, includes a portable version, enhanced performance and integrates search for Cas proteins. *Nucleic acids research*, *46*(W1), W246-W251.

Citations

- Databases
  - Mao, F., Dam, P., Chou, J., Olman, V., & Xu, Y. (2009). DOOR: a database for prokaryotic operons. ***Nucleic acids research***, ***37***(Database issue), D459–D463. https://doi.org/10.1093/nar/gkn757
  - Chen, L., Yang, J., Yu, J., Yao, Z., Sun, L., Shen, Y., & Jin, Q. (2005). VFDB: a reference database for bacterial virulence factors. ***Nucleic acids research***, ***33***(Database issue), D325–D328. https://doi.org/10.1093/nar/gki008
  - eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. Jaime Huerta-Cepas, Damian Szklarczyk, Kristoffer Forslund, Helen Cook, Davide Heller, Mathias C. Walter, Thomas Rattei, Daniel R. Mende, Shinichi Sunagawa, Michael Kuhn, Lars Juhl Jensen, Christian von Mering, and Peer Bork.Nucl. Acids Res. (04 January 2016) 44 (D1): D286-D293. doi: 10.1093/nar/gkv1248
  - McArthur ***et al***. 2013. The Comprehensive Antibiotic Resistance Database. ***Antimicrobial Agents and Chemotherapy***, 57, 3348-57.
  - Nucleic Acids Res. 2009 Jan;37(Database issue):D443-7. doi: 10.1093/nar/gkn656. Epub 2008 Oct 2.
  - Arndt, D., Grant, J., Marcu, A., Sajed, T., Pon, A., Liang, Y., Wishart, D.S. (2016) PHASTER: a better, faster version of the PHAST phage search tool. Nucleic Acids Res., 2016 May 3.
  - Arango-Argoty, G., Garner, E., Pruden, A. ***et al.*** DeepARG: a deep learning approach for predicting antibiotic resistance genes from metagenomic data. ***Microbiome*** 6, 23 (2018). https://doi.org/10.1186/s40168-018-0401-z
  - Ashburner et al. Gene ontology: tool for the unification of biology. Nat Genet. May 2000;25(1):25-9.
  - Prediction of lipoprotein signal peptides in Gram-negative bacteria.A. S. Juncker, H. Willenbrock, G. von Heijne, H. Nielsen, S. Brunak and A. Krogh.Protein Sci. 12(8):1652-62, 2003

Citations

- Homology:
  - Huerta-Cepas, J., Forslund, K., Coelho, L. P., Szklarczyk, D., Jensen, L. J., Von Mering, C., & Bork, P. (2017). Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper. *Molecular biology and evolution*, *34*(8), 2115-2122.
  - Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. *Bioinformatics*, *30*(14), 2068-2069.
  - Arango-Argoty, G., Garner, E., Pruden, A., Heath, L. S., Vikesland, P., & Zhang, L. (2018). DeepARG: a deep learning approach for predicting antibiotic resistance genes from metagenomic data. *Microbiome*, *6*(1), 1-15.
  - Finn, R. D., Attwood, T. K., Babbitt, P. C., Bateman, A., Bork, P., Bridge, A. J., ... & Gough, J. (2017). InterPro in 2017—beyond protein family and domain annotations. *Nucleic acids research*, *45*(D1), D190-D199.
  - Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+
  - Buchfink, B., Xie, C., & Huson, D. H. (2015). Fast and sensitive protein alignment using DIAMOND. *Nature methods*, *12*(1), 59.