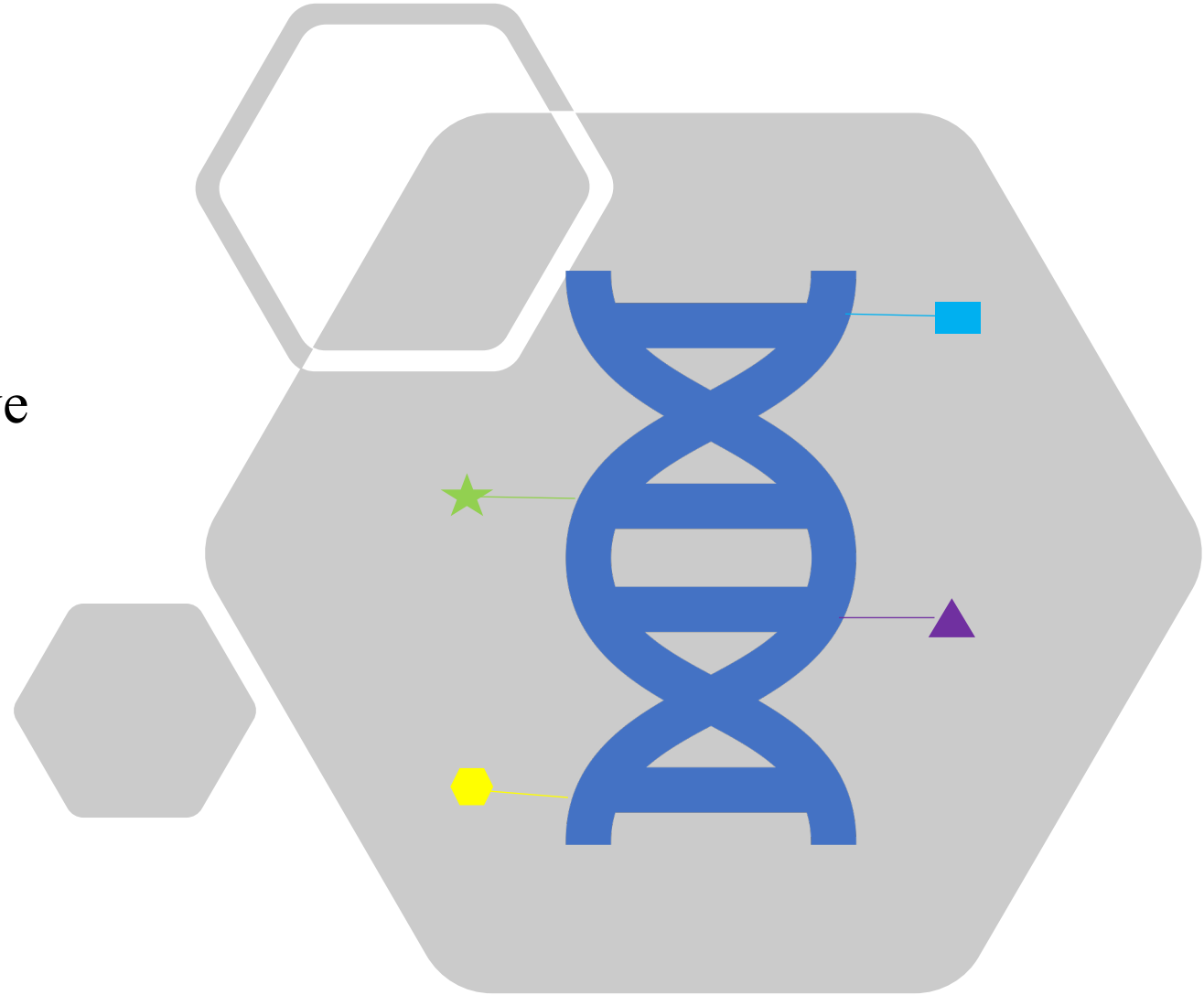
Functional Annotation

Team 2

Danielle Temples, Courtney Astore, Ujani Hazra, Sooyoun Oh, Rhiya Sharma

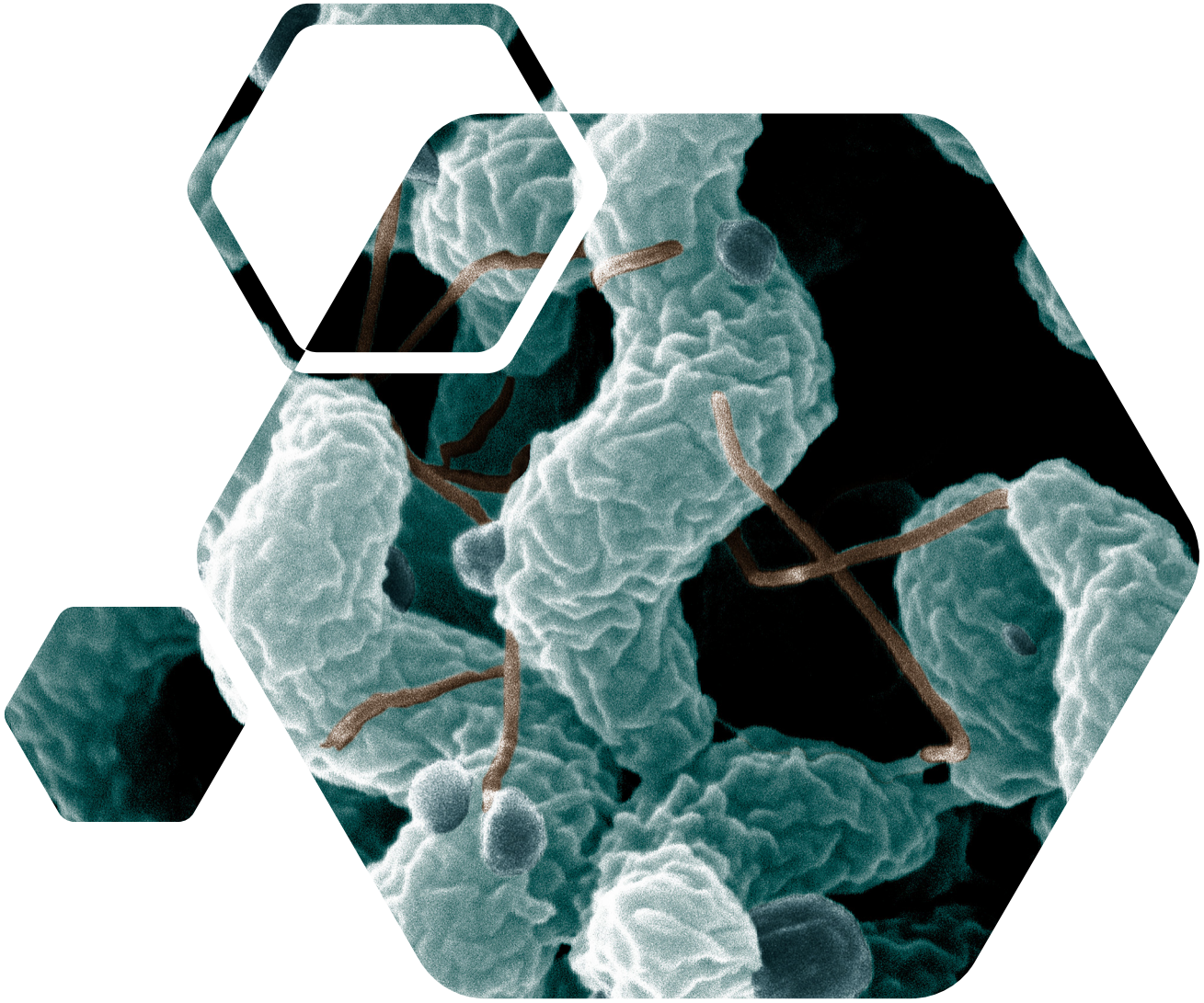
Outline

- Pathogenic Organism
- Functional Annotation & Objective
- Functional Annotation Steps with Software Selection
  - Clustering
  - Homology
  - *Ab initio*
- Future Steps & Deliverables



Campylobacter jejuni

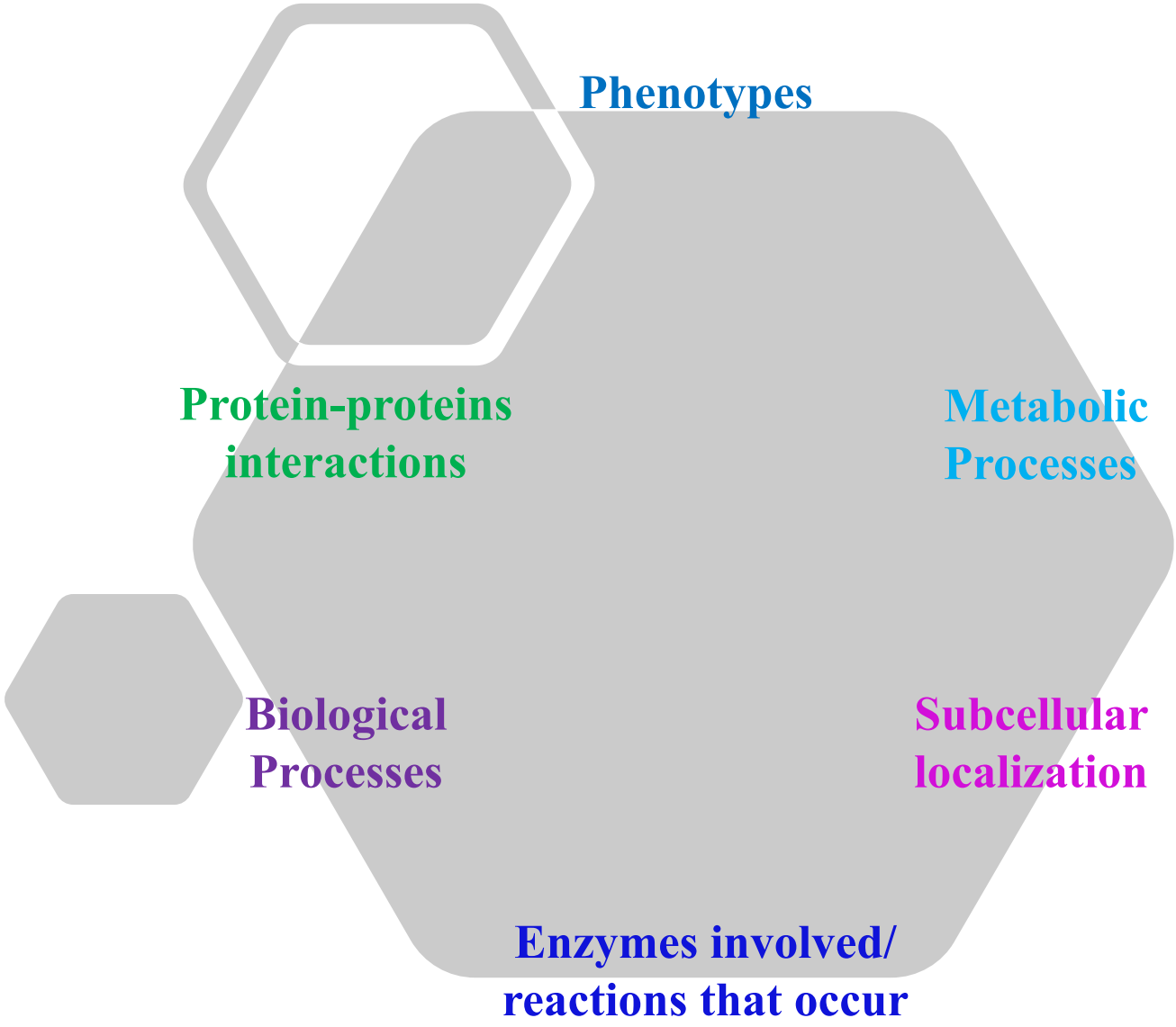
- Campylobacter species are the major cause of human bacterial gastroenteritis
- *C. jejuni* and *C. coli* together account for over 95% of Campylobacter infections in humans
- Certain strains are linked with the development of the neurological disorder Guillain-Barre syndrome (GBS)
- Gram negative bacteria
- Outer membrane is highly related to antibiotic resistance



What is Functional Annotation?

The process of determining the biological function(s) of genes and proteins

Objective: Perform a full functional annotation on the genes and proteins determined by the Gene Prediction group that is relevant to *C. jejuni*



Homology vs. a*b initio*-Based Techniques

Homology

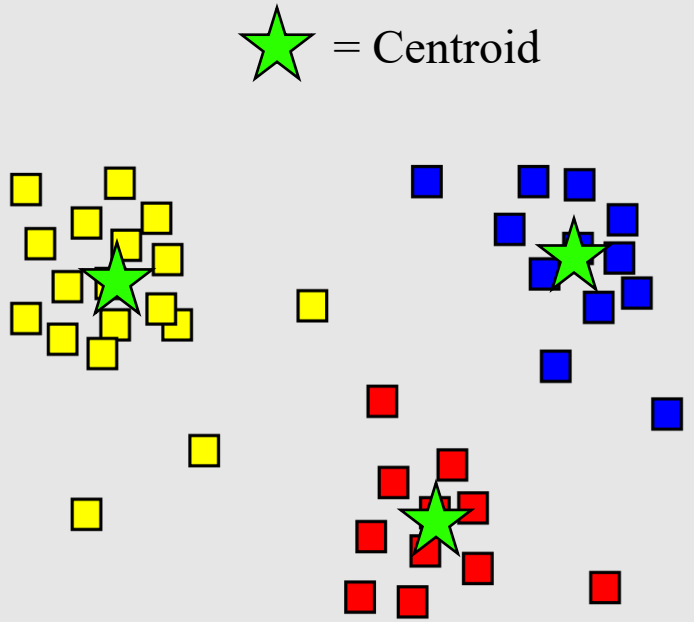
- Determine function via sequence similarity to already functionally annotated sequences
- This is limited by what we already know

ab-initio

- Determine function via predictive model without comparing to existing sequences
- This is based on laws of nature
- Difficult to verify without experiments

Why Clustering?

- Significant sequence similarity implies shared ancestry that often leads to shared function
- Clustering such sequences can reduce repeat queries in homology-based annotations
- **Reducing repeats improves speed and storage costs**



How Can Clustering Reduce Redundancy?

"After clustering, the size of a NR [non-redundant] query data set ... can often **be 50% to many times smaller** than the original data set. So overall, the annotation using NR data sets can **be easily accelerated by 10-fold**."

Data set[a]	Number sequences	Total	Cutoff [b] (%)	Clusters	Reduced to (%)	Time (minutes)[c]
NCBI NR	12 054 819	4.3 GB	90	7 036 029	58	181
16S (Silva + Greengene)	555 530	799 MB	98	154 170	28	90
NCBI microbial genomes	3 355	6.4 GB	90	1 279	38	389
NCBI virus sequences	1 042 347	1.3 GB	95	288 701	28	480

Li, Weizhong & Fu, Limin & Wu, Sitao & Wooley, John. (2012). Ultrafast clustering algorithms for metagenomic sequence analysis. *Briefings in bioinformatics*. 13. 10.1093/bib/bbs035.

Clustering Comparison

Tool	Release Date	Algorithm	DNA/Protein	Reproducibility	Runtime	Citations
CD-HIT	2001	Greedy clustering using short word filtering and word counting	Both	Output depends on input order	*O(NK)*	**780***
UCLUST	2010	Greedy clustering using threshold BLAST sequence identity comparison	Both	Output depends on input order	*O(NK)*	**11457**
Linclust	2018	Greedily clusters sequences to *m* centroid sequences that share the most *k*-mers	Both	Output depends on input order	***O(N)***	42
MeShClust	2018	Mean-shift algorithm that finds locally optimal cluster centroids	DNA	**Creates locally optimal clusters**	$O(N^2)$	15

Homology Categories

Prophage:

- Play an important role in the evolution of bacterial genomes and their pathogenicity
- Can change or knock out gene functions; alter gene expression

Virulence:

- A pathogen's ability to infect or damage a host
- Ex: toxins, surface coats that inhibit phagocytosis, surface receptors that bind to host cells

Fully Automated Functional Annotation Tools:

- Tools that annotate a spectrum of features related to the function

Antibiotic Resistance:

- When bacteria develop the ability to defeat the drugs designed to kill them
- Leads to higher medical costs, prolonged hospital stays, and increased mortality

Operons:

- A functional unit of transcription and genetic regulation
- Identifying these may enhance our knowledge of gene regulation & function which is a key addition to genome annotation

Prophage

ProphET

- ***PROPH****age* ***E****stimation* ***T****ool*
- Identifies prophages in bacterial genomes with high precision and offers a fast, highly scalable alternative
- Uses three steps: similarity search, calculation of the density of prophage genes, and edge refinement

PHASTER

- ***PHA****ge* ***S****earch* ***T****ool* ***E****nhanced* ***R****elease*
- Rapid identification and annotation of prophage sequences
- Performs database comparisons as well as phage "cornerstone" feature identification steps to locate, annotate and display prophage sequences and prophage features

Prophage Comparison

Tool	Release Date	Algorithm	DNA/Protein	Citations
ProphET	2019	Similarity search, calculate density of prophage genes, and edge refinement	DNA/GFF	37
PHASTER	2015	Database comparisons & phage 'cornerstone' feature identification	DNA/GFF	799

	Sensitivity	PPV
Prophinder	69.8	73.5
PHAST	77.0	69.2
PHASTER	78.1	74.6
ProphET	73.3	84.2
PhySpy	79.0	47.8
PhySpy*	77.7	55.8

High and low values are represented on a scale from red to blue. PhySpy: Predictions using the generic training set. PhySpy*: Predictions using a taxonomically-optimized training set (Materials and Methods; S3 Table).

https://doi.org/10.1371/journal.pone.0223364.t001

Reis-Cunha JL, Bartholomeu DC, Manson AL, Earl AM, Cerqueira GC (2019) ProphET, prophage estimation tool: A stand-alone prophage sequence prediction tool with self-updating reference database. PLOS ONE 14(10): e0223364. https://doi.org/10.1371/journal.pone.0223364

Virulence

VFDB

- ***V**irulence **F**actor **D**ata**B**ase*
- Provide virulence structure features, functions, and mechanisms used to allow pathogens to conquer new niches and circumvent host defense mechanisms
- BLAST based identification of virulence genes

T3DB

- ***T**oxin and **T**oxin **T**arget **D**ata**B**ase*
- Combines detailed toxin data with comprehensive toxin target information
- Toxin metabolism prediction, toxin/drug interaction prediction, and general toxin hazard awareness by the public
- Offers local BLAST search

Virulence Comparison

Tool	Release Date	Algorithm	DNA/Protein	Citations
VFDB	2005	BLAST-like algorithm	Both	513
T3DB	2010	BLAST-like algorithm	Both	105

http://www.t3db.ca/

http://www.mgc.ac.cn/cgi-bin/VFs/v5/main.cgi

Fully Automated Functional Annotation

PANNZER2

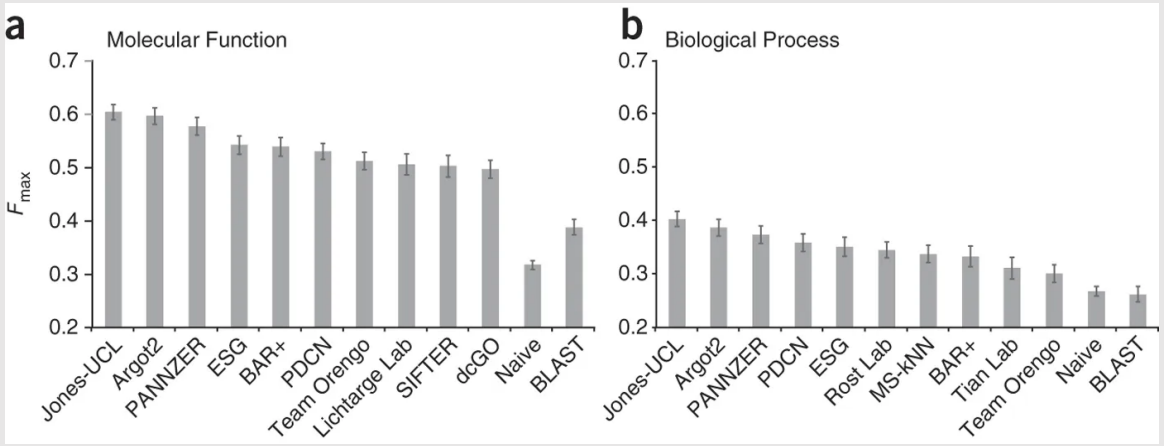
- ***P****rotein* ***ANN****otation with* ***Z****-sco****RE***
- Provides both Gene Ontology (GO) annotations and free text description predictions
- Uses SANSparallel to perform high-performance homology searches
- Updated on a monthly schedule

BLAST

- ***B****asic* ***L****ocal* ***A****lignment* ***S****earch* ***T****ool*
- A database is searched for high-scoring local alignments with a query
- The annotations on the sequence that score the highest alignment are assigned to the query sequence, provided the alignment score passes a threshold

Fully Automated Comparison

Tool	Release Date	Algorithm	DNA/Protein	Citations
PANNZER2	2018	SANSparallel: interactive homology search against Uniprot	Both	25
BLAST	2010	Heuristic algorithm	Both	11484



Radivojac, P., Clark, W., Oron, T. et al. A large-scale evaluation of computational protein function prediction. Nat Methods 10, 221–227 (2013). https://doi.org/10.1038/nmeth.2340

Server	GO prediction	DE prediction	>1000 query sequences	Prob. estimate	Open source	Last database update/update schedule
PANNZER2	Yes	Yes	Yes	Yes	Yes	Monthly (synchronised with UniProt)
ARGOT	Yes	No	Yes	No	No	11/2016
PFP	Yes	No	No	yes	No	Unknown
FunFam	Yes	No	No	No	Data can be downloaded	Daily
INGA	Yes	No	No	Yes	No	04/2015
eggNOG	Yes	Keyword	Yes	No	Yes	11/2017
dcGO	Yes	No	Error	Yes	No	06/2016*

Törönen, Petri et al. "PANNZER2: a rapid functional annotation web server." Nucleic acids research vol. 46,W1 (2018): W84-W88. doi:10.1093/nar/gky350

Antibiotic Resistance

CARD

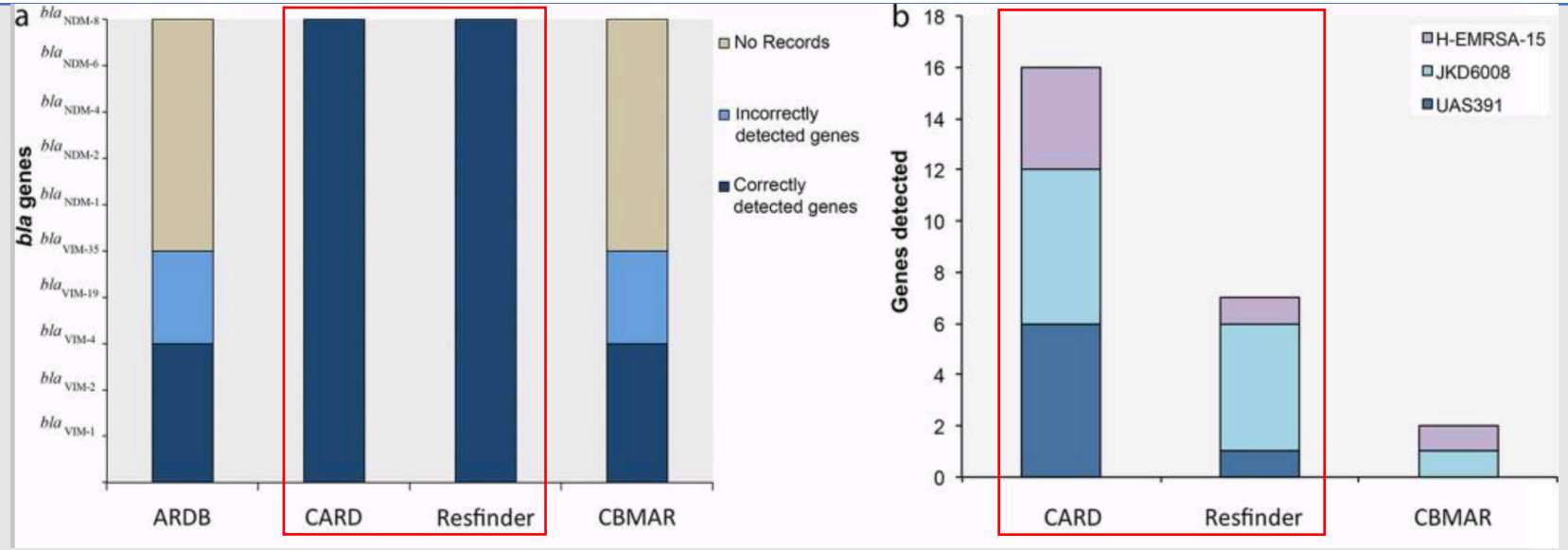
- ***C****omprehensive* ***A****ntibiotic* ***R****esistance* ***D****atabase*
- Provides data, models, and algorithms relating to the molecular basis of antimicrobial resistance
- Can be used for analysis of genome sequences using the **Resistance Gene Identifier**

ResFinder

- Identification of acquired antimicrobial resistance genes in total or partial sequenced isolates of bacteria.
- Based on a database of more than 2,000 resistance genes covering 12 types of antimicrobial resistance agents which is searched using BLAST

Antibiotic Resistance Comparison

Tool	Release Date	Algorithm	DNA/Protein	Citations
Resfinder	2012	Uses BLAST for identification of acquired antimicrobial resistance genes in whole-genome data	Both	1981
CARD	2015	ARO data associated with detection models that can be used for prediction of resistome and for analysis of genome sequences using RGI	Both	799



Xavier, Basil Britto et al. "Consolidating and Exploring Antibiotic Resistance Gene Data Resources." *Journal of clinical microbiology* vol. 54,4 (2016): 851-9. doi:10.1128/JCM.02717-15

Operons

OperonDB

- Detects and analyzes conserved gene pairs
- For each conserved gene pair, calculate an estimate of probability that the genes belong to the same operon
- Considers other alternative possibilities:
  - Functionally unrelated genes may have the same order due simply because they were adjacent in a common ancestor
  - Genes may be adjacent in two genomes by chance alone, or due to horizontal transfer of the gene pair

ODB

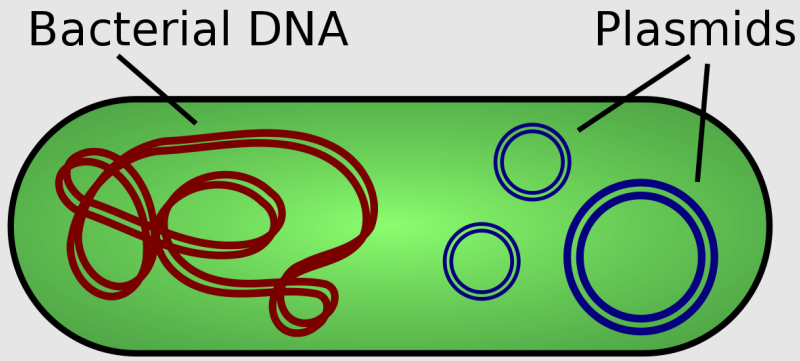
- ***O**peron **D**ata**B**ase*
- Collects data of all known and conserved operons in completely sequenced genomes
- ODB proposes the idea of reference operons as a new operon prediction tool. A reference operon, a set of possible orthologous genes that organize operons, is defined by clustering all known operons

Operon Comparison

Tool	Release Date	Algorithm	Citations
OperonDB	2010	Calculate a probability that genes belong to the same operon for each conserved gene pair	112
ODB v4	2011	Predict by mapping orthologous genes on a reference operon	26

Plasmids

- Plasmids are genetic structures that can replicate independently of chromosomal DNA
- In bacteria, they are usually circular and double-stranded
- In *Campylobacter* species, "...plasmids contained various replication-associated and conjugation-associated genes that showed homology with a plasmid from *Actinobacillus actinomycetemcomitans*"



https://en.wikipedia.org/wiki/Plasmid

Plasmid Software

Tool	Release Date	Algorithm	Citations
PlasmidSeeker	2018	Uses *k*-mer abundance to distinguish between bacterial and plasmid sequences	14
PlasmidFinder	2014	Uses nucleotide-nucleotide BLAST to distinguish plasmid sequences, web-only	1023
PlasFlow	2018	Neural network approach for identification of bacterial plasmid sequences in environmental samples	62

ab-initio Categories

Transmembrane Proteins (Cell Membrane and Outer Membrane):

- Bacteria have the ability to export effector proteins in membranes of eukaryotic host
- Integral membrane protein that function as gates or docking sites that allow or prevent the entry or exit of materials across the cell membrane

Signal Peptides:

- Guide secretory proteins to find their correct locations outside the cell membrane for signal transduction

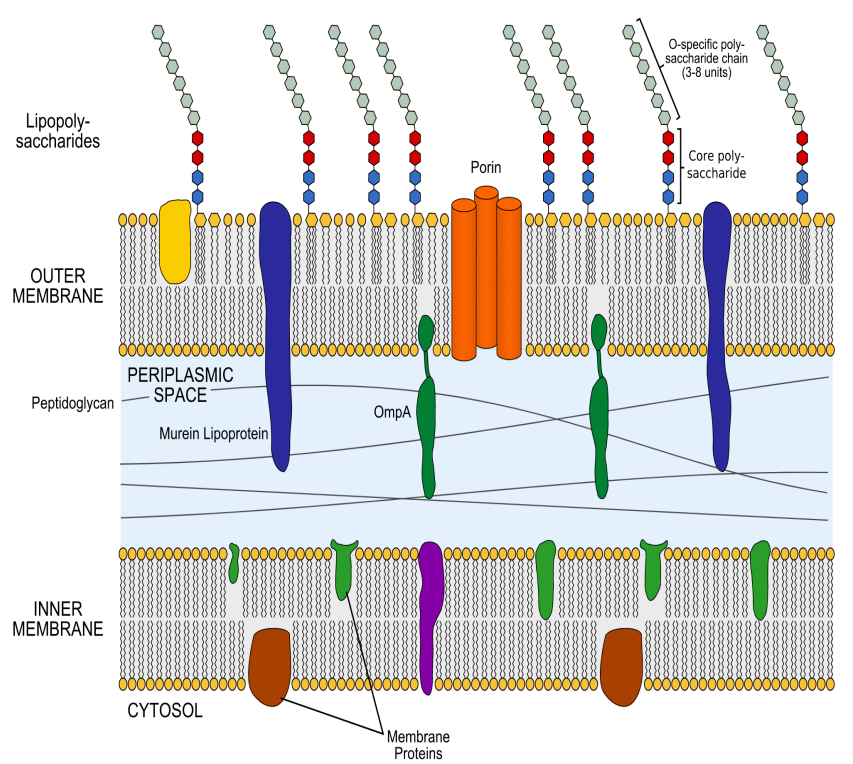
CRISPR:

- Provides immunity to the bacteria against Bacteriophages
- Contributes to the Virulence and Pathogenicity of the bacteria

Gram-negative Cell Envelope

- The Gram-negative Bacteria cell wall is composed of a thin layer of peptidoglycan surrounded by a membranous structure called the outer membrane
- Outer membrane makes it more difficult to treat the bacteria, by regulating the passage of antibiotics inside the cell
- The lipopolysaccharide component of the outer membrane acts as a virulence factor and causes disease in animals
- The cell membrane is composed of a phospholipid bilayer
- The transmembrane proteins act as gates or docking sites, regulating the exchange of material across the cell boundary

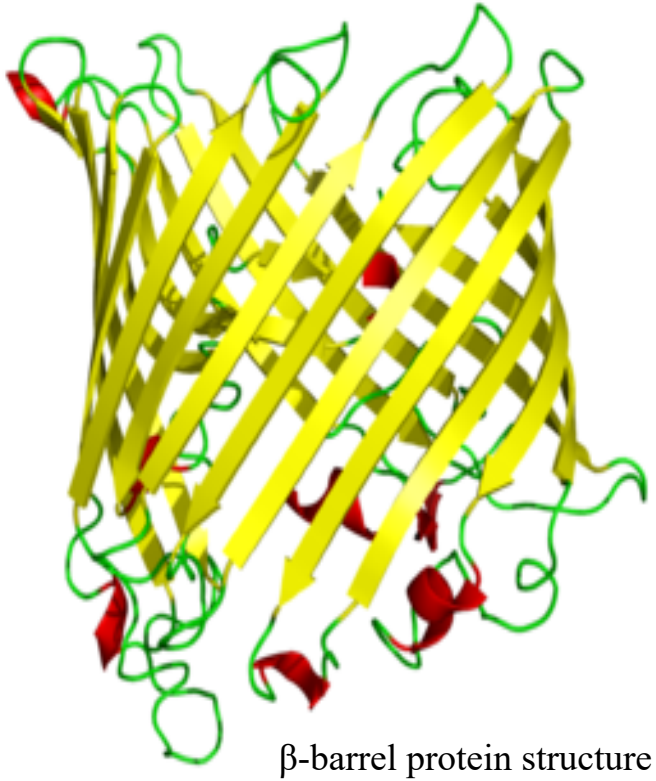
Gram-negative Cell Wall and Plasma Membrane



https://en.wikipedia.org/wiki/Gram-negative_bacteria#/media/File:Gram_negative_cell_wall.svg

Outer Membrane Proteins

- The outer membrane is the first line of defense for Gram-negative bacteria against toxic compounds
- We are interested in two types of proteins of the OM: lipoproteins and β-barrel proteins
- β-barrel proteins, such as the porins, play a fundamental role in pathogenicity and represent useful targets for therapeutic development
- Porin channels prevent the entry of harmful chemicals and antibiotics like penicillin. These channels can also expel out antibiotics making it much more difficult to treat in comparison to gram-positive bacteria
- We will be looking into ab-initio tools that predict β-barrel proteins



β-barrel protein structure

https://en.wikipedia.org/wiki/Beta_barrel#/media/File:Sucrose_porin_1a0s.png

Comparison of β-barrel Protein Prediction Tools

Method	% Globular	% β-Barrels	% Non-β-barrels
PSORT	–	a	–
TBB-pred	92.3	88.8	–
MCM-BB	92.5	89.2	–
ProfTMB	–	45	100
PRED-TMBB	89	89	
HMM-B2TMR	90	84	
BBF	–	–	–
HUNTER	96.9	82.4	96
Wimley	–	–	–
Liu AaSecStru	92.5	85.5	
BOMP	–	88	98.8

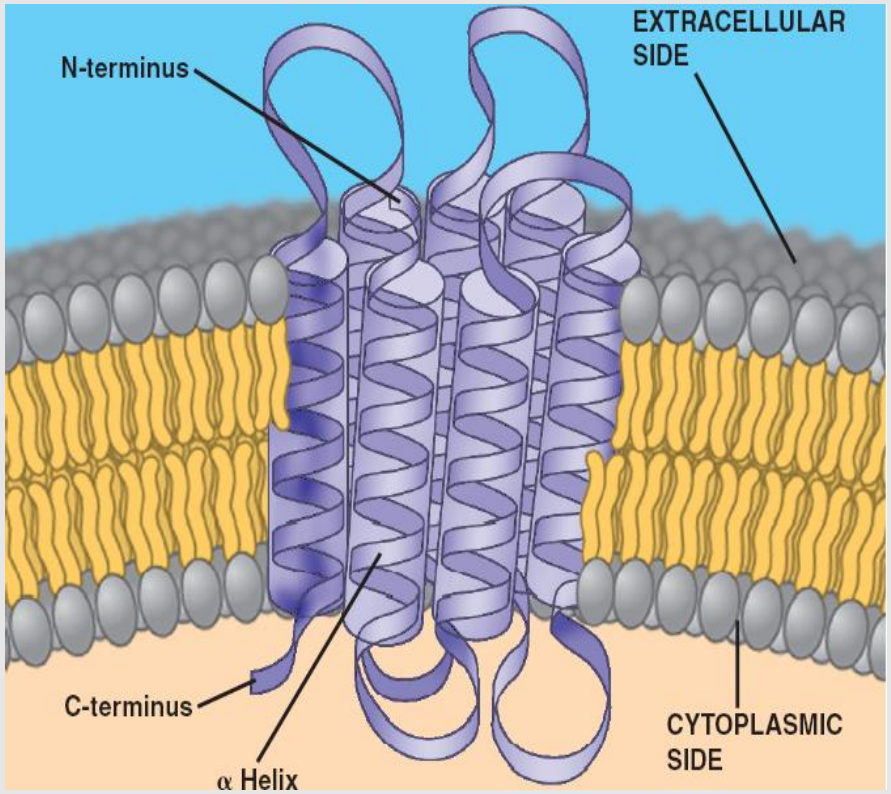
Berven, F.S., Karlsen, O.A., Straume, A.H. *et al.* Analysing the outer membrane subproteome of *Methylococcus capsulatus* (Bath) using proteomics and novel biocomputing tools. *Arch Microbiol* **184,** 362–377 (2006). https://doi.org/10.1007/s00203-005-0055-7

β-barrel Protein Prediction Comparison

Tool	Release Date	Citations	Basis
TBB-pred	2004	123	Uses Neural Networks/SVM, command-line tool available
MCM-BB	2004	18	Uses HMM, only web version available, very low citations.
PRED-TMBB	2004	359	Uses HMM, only web version available.
HMM-B2TMR	2002	223	Uses HMM, not sure about availability of tool
Liu AaSecStru	2003	48	Not available
BOMP	2004	187	Not available

Cell Membrane Proteins

- The fundamental structure of the membrane is the phospholipid bilayer, which forms a stable barrier
- Proteins embedded within the phospholipid bilayer carry out the specific functions of the plasma membrane, including selective transport of molecules and cell-cell recognition
- Transmembrane proteins span the lipid bilayer with portions exposed on both sides of the membrane. They are amphipathic, having both hydrophobic and hydrophilic regions
- We will be looking into ab-initio tools that predict α-helix transmembrane proteins



A-helix Transmembrane Protein

Comparison of Transmembrane Protein Prediction Tools

Table III. Prediction of the number and location of TM helices: non-redundant dataset[a]

Prediction method	N_P	N_C	Q_P (%)	Q_{3TM} (%)	Q_{3NTM} (%)	Q_3 (%)	N_{TM}
ALOM2	178	175	80.1	41.8	97.8	73.8	35
DAS	297	260	92.2	61.5	95.2	80.7	43
HMMTOP2	270	254	94.4	70.4	91.1	82.2	58
MEMSAT 1.5	256	246	93.9	69.2	92.8	82.6	55
MEMSAT 2	250	237	91.6	67.2	93.2	82.1	41
MPEX	261	246	93.0	69.4	87.2	79.5	50
PHD	249	238	92.1	62.1	94.1	80.3	41
SPLIT4	254	250	95.8	78.3	90.3	85.2	61
TMAP	249	240	92.9	76.1	86.2	81.9	48
TM-FINDER	235	226	90.1	67.5	91.7	81.3	44
TMHMM2	246	239	93.1	71.1	92.5	83.3	49
TMPRED	259	247	93.8	68.5	92.1	82.0	53
TOPPRED2	273	253	93.5	70.7	89.5	81.4	52

Cuthbertson, J. M., Doyle, D. A., & Sansom, M. S. (2005). Transmembrane helix prediction: a comparative evaluation and analysis. *Protein Engineering Design and Selection*, *18*(6), 295-308.

α-helix Transmembrane Protein Prediction Comparison

Tool	Year of publication	Citations	Algorithm	Basis
TMHMM2	2001	9675	HMM	High number of citations, uses Hidden Markov Model, stand-alone version available, most widely used, excellent documentation, easy to use
HMMTOP2	2001	1849	HMM	High number of citations, uses Hidden Markov Model, stand-alone version available, proper documentation, software is maintained and up-to date
MEMSAT 3	2007	451	Neural Network	Comparatively low citations, poor documentation, uses Neural Networks
MEMSAT-SVM	2009	369	SVM	Comparatively low citations, poor documentation, uses Support Vector Machines (SVM)
SPLIT 4	2002	226	Preference Functions	Low citations, only web server available, uses Preference Functions to predict models

Transmembrane Protein

HMMTop2

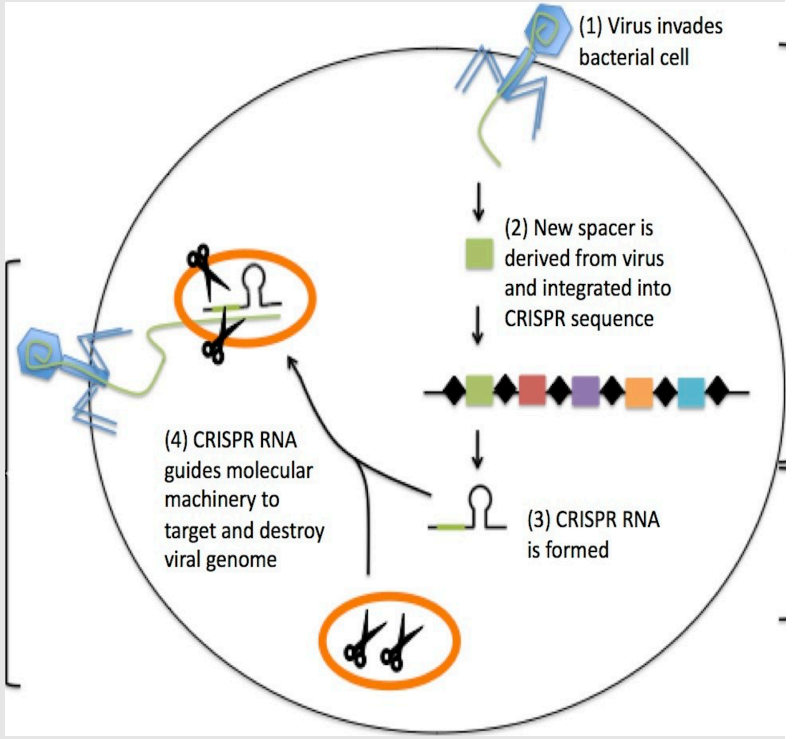
- *Input:* FASTA format
- *Algorithm*: Hidden Markov Model
- *Reliable Mode*: Runs multiple iterations. It searches or makes optimization for the best topology. The results are reliable but more time-consuming
- *Output:* Number of transmembrane helices predicted, position of each residue with respect to the cell membrane (outside/inside the cell, transmembrane segment)

TMHMM2

- *Input: FASTA format*
- *Algorithm*: Hidden Markov Model
- User can input up to 10,000 protein sequences at a time
- *Output*: Number of transmembrane helices predicted, position of each residue with respect to the cell membrane (outside/inside the cell, transmembrane segment), optional graphical output visualizing the predicted helix

CRISPR

- **CRISPR** (**C**lusters of **R**egularly **I**nterspaced **S**hort **P**alindromic **R**epeats) is a family of DNA sequence found in prokaryotic genomes
- The bacteria capture snippets of DNA from invading viruses and use them to create CRISPR arrays
- The CRISPR arrays allow the bacteria to "remember" the viruses. If the virus attacks again, the bacteria produces RNA segments from the CRISPR arrays to target the viruses' DNA
- The bacteria then use Cas9 or a similar enzyme to cut the DNA apart, which disables the virus
- CRISPR-Cas9 is a naturally occurring genome editing system in bacteria, providing immunity. These systems have significant impact in altering the bacterial physiology in term of its virulence and pathogenicity



The steps of CRISPR-mediated immunity

http://sitn.hms.harvard.edu/flash/2014/crispr-a-game-changing-genetic-engineering-technique/

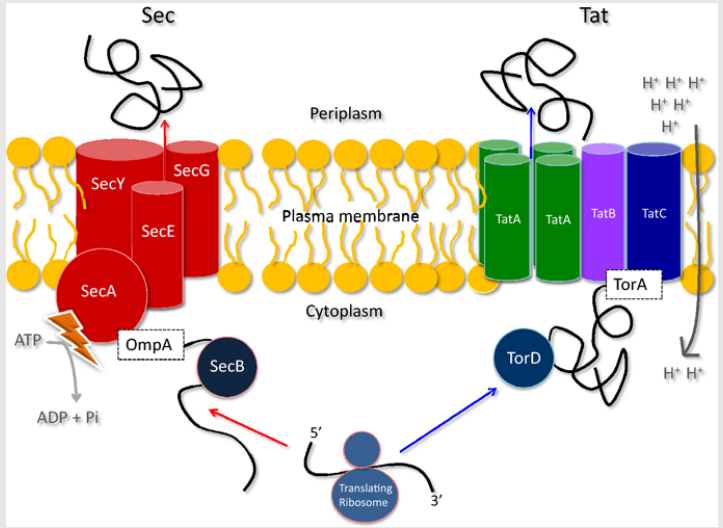
CRISPR Prediction Tools

Tool	Year	Citations	Basis
PilerCR	2007	231	Comparatively low citations, excellent documentation, software is maintained and up-to date
CRT	2007	430	High number of citations, proper documentation, maintained

- *Input:* FASTA format
- *Output:* Spacer and repeat sequences in the genome along with the coordinates

Signal Peptides

- These are short peptides (typically 16-30 AA in sequence length) present at N-terminus of a fraction of recently synthesized secretory proteins destined for secretory pathway
- Generally function to initiate cell translocation by directing secretary proteins across the cytoplasmic membrane to the cytoplasm
- There are 2 major pathways/channels involved translocation of secretory proteins which include different integral membrane protein subunits:
  - Sec : Secretion pathway
    - Transports unfolded proteins
  - Tat: Twin Arginine translocation pathway
    - Transports folded proteins (by way of amino acid interactions)
  - Those cleaved by signal peptidase I (SPI) are standard and those cleaved by signal peptidase II (SPII) are lipoproteins, which is consistent for both pathways



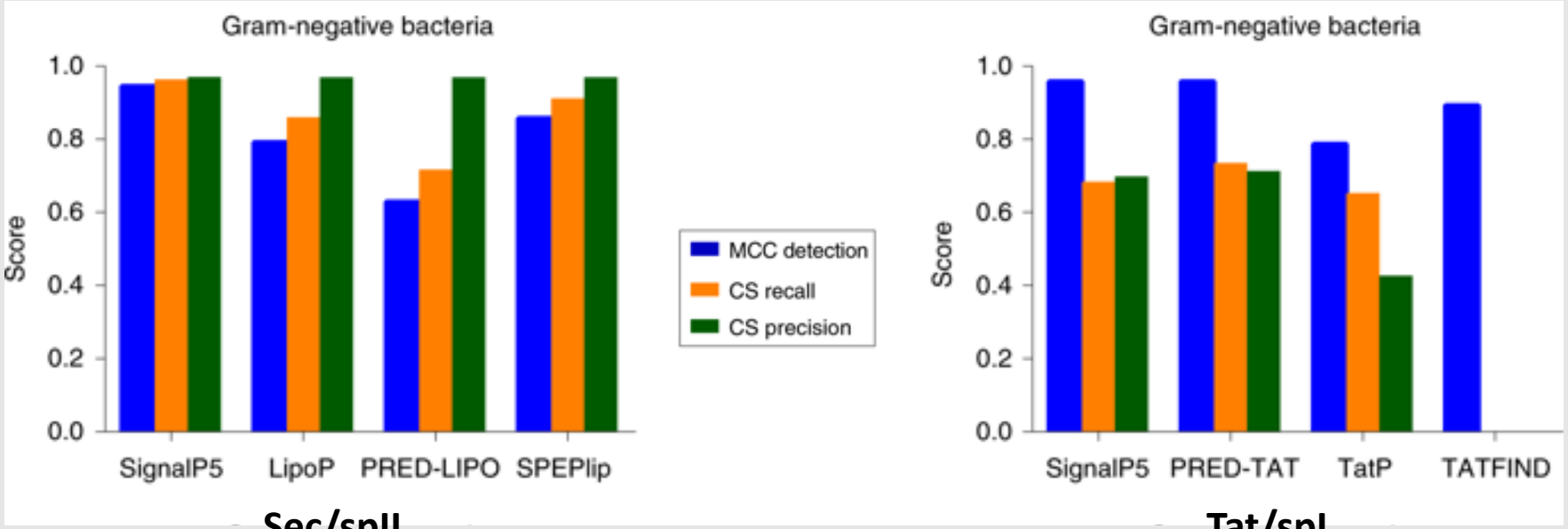
Sec vs. Tat pathways involved in translocation

Signal Peptide Prediction Comparison

Tool	Year	Citations	Algorithm	Basis
SignalP v5	2019	153	Deep neural networks	Predicts presence of signal peptides and location of cleavage sites, created and optimized from multiple versions, doesn't determine lipoproteins
tatP	2005	502	Neural network	Predicts the presence of the Tat pathway, which is required for virulence and other cellular activities
Phobius	2004	1,291	HMM	Stand-alone version, sufficient documentation, predicts both signal peptide and transmembrane proteins
LipoP 1.0	2003	1,055	HMM	Stand-alone version, determines lipoprotein signal peptides

Comparison of Signal Peptides Prediction Tools

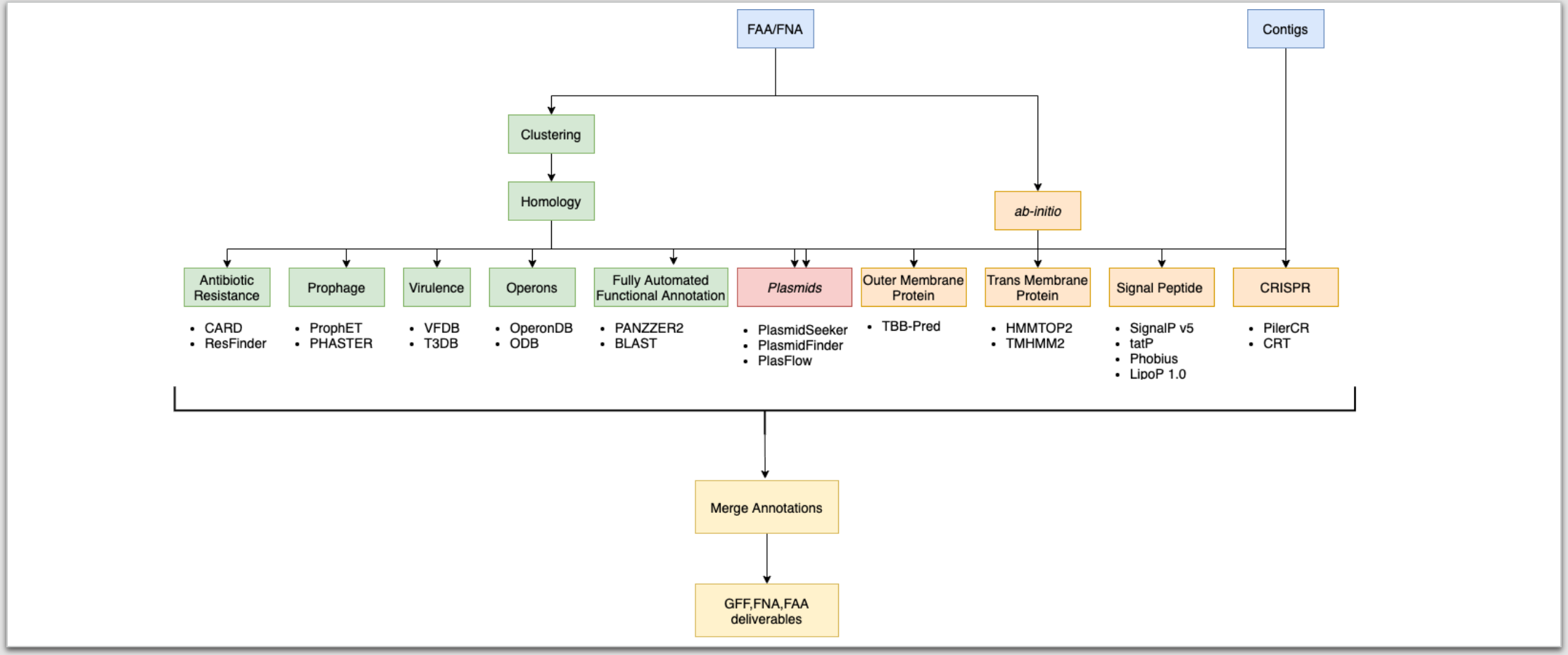
MCC = Matthews Correlation Coefficient (quality of binary classifications)
CS = Cleavage site



Sec/spII **Tat/spI**

Almagro Armenteros, J.J., Tsirigos, K.D., Sønderby, C.K. *et al.*SignalP 5.0 improves signal peptide predictions using deep neural networks. *Nat Biotechnol* 37, 420–423 (2019). https://doi.org/10.1038/s41587-019-0036-z

Bioinformatics Pipeline



References

Re-annotation and re-analysis of the *Campylobacter jejuni* NCTC11168 genome sequence https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1899501/

Genomic Characterization of *Campylobacter jejuni* Strain M1 . https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2928727/

A Genome-Wide Association Study to Identify Diagnostic Markers for Human Pathogenic *Campylobacter jejuni* Strains https://www.frontiersin.org/articles/10.3389/fmicb.2017.01224/full

Campylobacter jejuni transcriptional and genetic adaptation during human infection https://centerforimmunizationresearch.org/wp-content/uploads/2018/08/Campylobacter-jejuni-transcriptional-and-genetic-adaptation-during-human-infection.pdf

A proteome-wide protein interaction map for *Campylobacter jejuni* https://link.springer.com/article/10.1186/gb-2007-8-7-r130

Li, Weizhong & Fu, Limin & Wu, Sitao & Wooley, John. (2012). Ultrafast clustering algorithms for metagenomic sequence analysis. *Briefings in bioinformatics*. 13. 10.1093/bib/bbs035.

Combining multiple functional annotation tools increases coverage of metabolic annotation . https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-018-5221-9

https://vbn.aau.dk/ws/portalfiles/portal/306201236/Hsu_et_al._2019_Comparative_genomics_and_genome_biology_of_Campylobacter_showae.pdf

https://services.healthtech.dtu.dk

https://link.springer.com/protocol/10.1007%2F978-1-4939-1115-8_22

https://www.nature.com/articles/s41467-018-05969-w

Hunt, Martin et al. "ARIBA: Rapid Antimicrobial Resistance Genotyping Directly From Sequencing Reads". *Microbial Genomics*, vol 3, no. 10, 2017. *Microbiology Society*, doi:10.1099/mgen.0.00013

Xavier, Basil Britto et al. “Consolidating and Exploring Antibiotic Resistance Gene Data Resources.” *Journal of clinical microbiology* vol. 54,4 (2016): 851-9. doi:10.1128/JCM.02717-15

Alcock, Brian P et al. "CARD 2020: Antibiotic Resistome Surveillance With The Comprehensive Antibiotic Resistance Database". *Nucleic Acids Research*, 2019. *Oxford University Press (OUP)*, doi:10.1093/nar/gkz935

Pertea, M. et al. "Operondb: A Comprehensive Database Of Predicted Operons In Microbial Genomes". *Nucleic Acids Research*, vol 37, no. Database, 2009, pp. D479-D482. *Oxford University Press (OUP)*, doi:10.1093/nar/gkn784. Accessed 2 Mar 2020.

Okuda, S. "ODB: A Database Of Operons Accumulating Known Operons Across Multiple Genomes". *Nucleic Acids Research*, vol 34, no. 90001, 2006, pp. D358-D362. *Oxford University Press (OUP)*, doi:10.1093/nar/gkj037. Accessed 2 Mar 2020.

Pawel S Krawczyk, Leszek Lipinski, Andrzej Dziembowski, PlasFlow: predicting plasmid sequences in metagenomic data using genome signatures, *Nucleic Acids Research*, Volume 46, Issue 6, 6 April 2018, Page e35, https://doi.org/10.1093/nar/gkx1321

Marasini, Daya et al. “Phylogenetic Relatedness Among Plasmids Harbored by *Campylobacter jejuni* and *Campylobacter coli* Isolated From Retail Meats.” *Frontiers in microbiology* vol. 9 2167. 12 Sep. 2018, doi:10.3389/fmicb.2018.02167

Carattoli, Alessandra et al. “In silico detection and typing of plasmids using PlasmidFinder and plasmid multilocus sequence typing.” *Antimicrobial agents and chemotherapy* vol. 58,7 (2014): 3895-903. doi:10.1128/AAC.02412-14

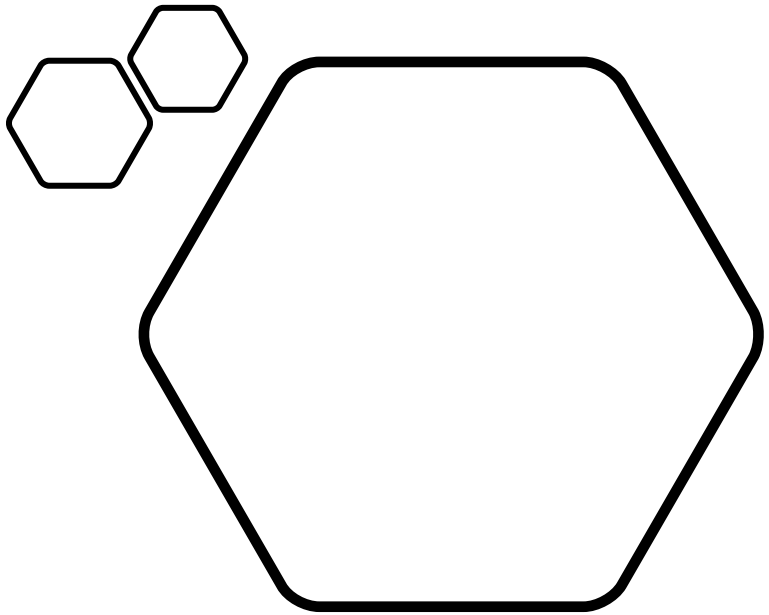
Marasini, Daya et al. “Phylogenetic Relatedness Among Plasmids Harbored by *Campylobacter jejuni* and *Campylobacter coli* Isolated From Retail Meats.” *Frontiers in microbiology* vol. 9 2167. 12 Sep. 2018, doi:10.3389/fmicb.2018.02167

Carattoli, Alessandra et al. “In silico detection and typing of plasmids using PlasmidFinder and plasmid multilocus sequence typing.” *Antimicrobial agents and chemotherapy* vol. 58,7 (2014): 3895-903. doi:10.1128/AAC.02412-14

Roosaare M, Puustusmaa M, Möls M, Vaher M, Remm M. 2018. PlasmidSeeker: identification of known plasmids from bacterial whole genome sequencing reads. *PeerJ*6:e4588 https://doi.org/10.7717/peerj.4588

Weizhong Li, Lukasz Jaroszewski, Adam Godzik, Clustering of highly homologous sequences to reduce the size of large protein databases , *Bioinformatics*, Volume 17, Issue 3, March 2001, Pages 282–283, https://doi.org/10.1093/bioinformatics/17.3.282

Steinegger, M., Söding, J. Clustering huge protein sequence sets in linear time. *Nat Commun* 9, 2542 (2018). https://doi.org/10.1038/s41467-018-04964-5



Questions?

Thank you!