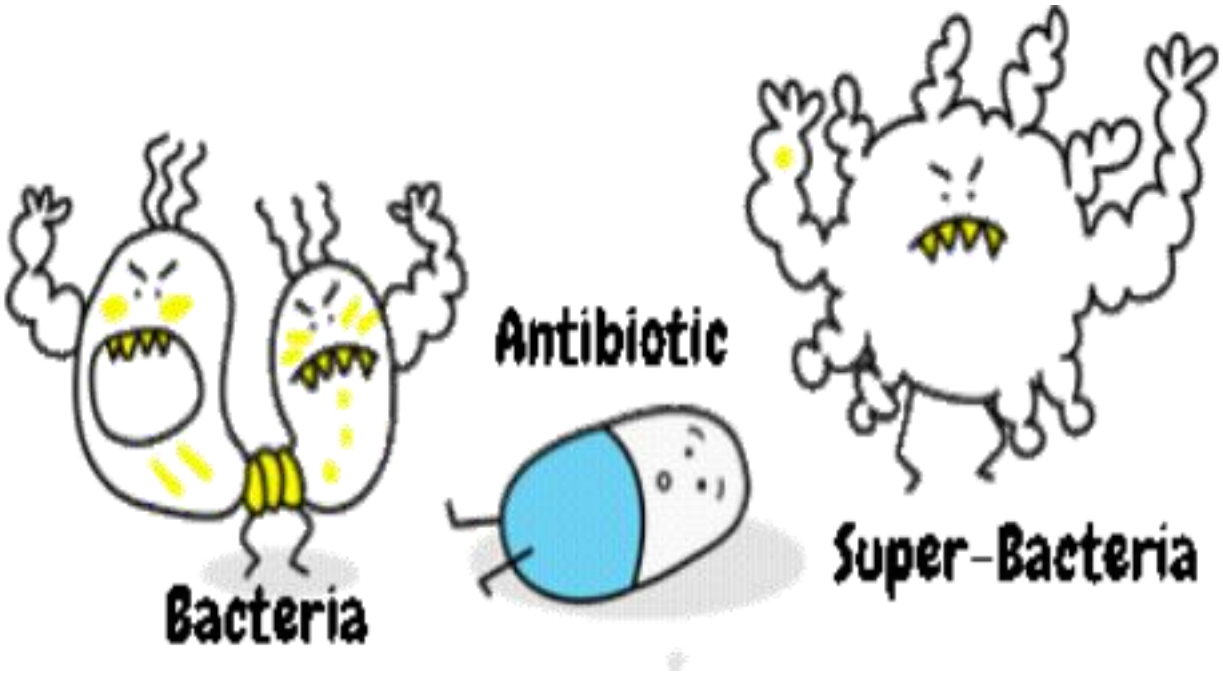


Using Machine-Learning to Discern the Antimicrobial Resistance Profile of Microbes

Date: 14-09-2023

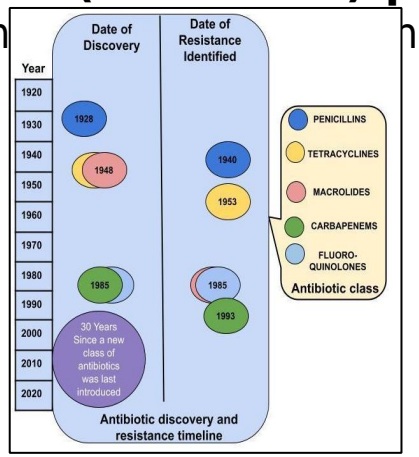
Dr. Manish Kumar (**Associate Professor & Head**)
Department of Biophysics, University of Delhi South Campus, New Delhi, India

Background

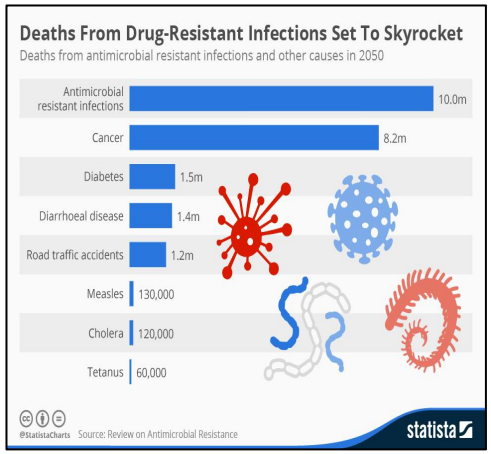
Antimicrobial resistance (AMR) is one of the most **serious public health threats of the twenty-first century**. A systematic review published recently in the ***The Lancet*** reveals, its global impact is **far greater than many infectious diseases such as malaria and AIDS**.

Resistance to antimicrobial agents has become a major source of **morbidity and mortality worldwide**.

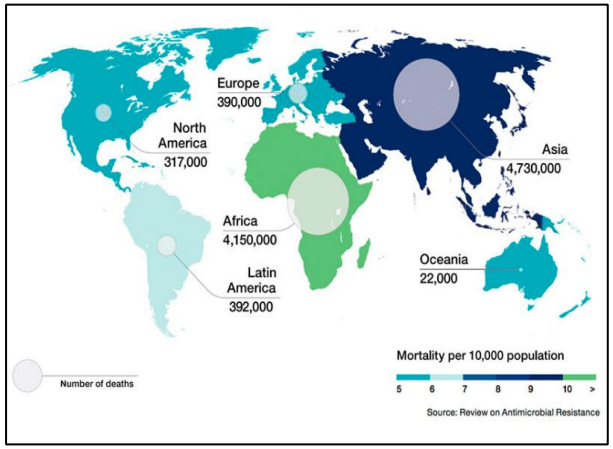
Antibiotic resistance remains a public health threat during the **Coronavirus disease 2019 (COVID-19) pandemic**. The ongoing (COVID-19) pandemic has further com h.



Timeline of discovery of antibiotics and resistance



Drug-resistance infections mortality

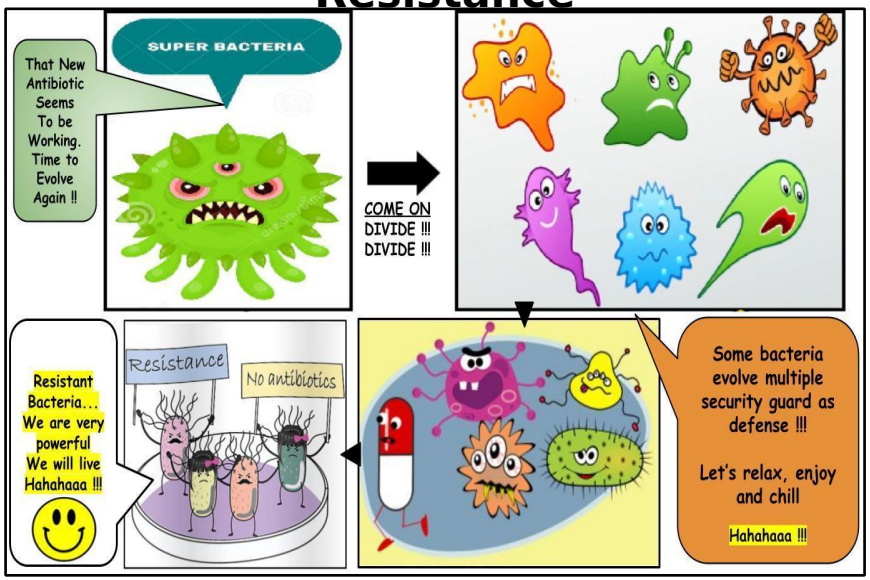


Global distribution of 10 million deaths expected by 2050 due to antimicrobial resistance

Emergence of AMR in Bacteria

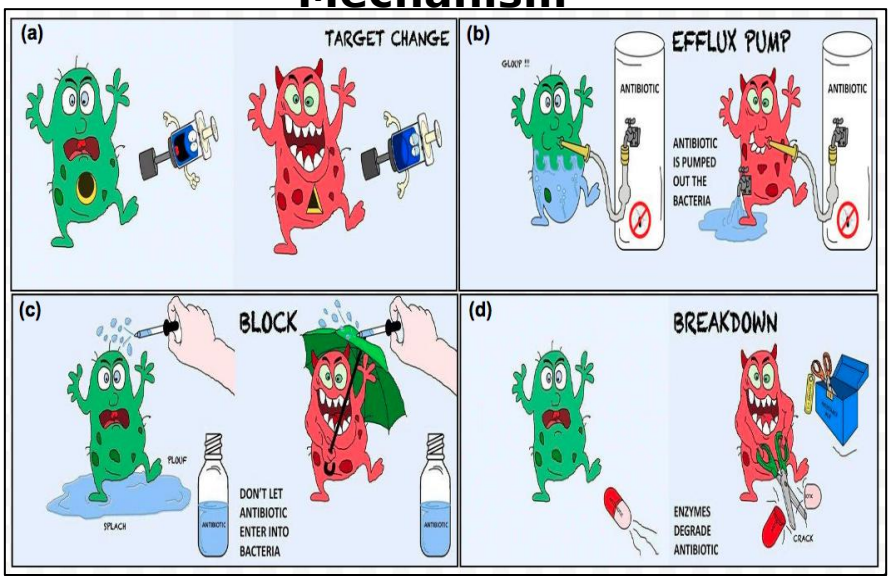
- **Bacterial resistance is considered a major concern in healthcare organizations.** Specially gram-negative bacteria are a leading cause of life-threatening infections and include **nosocomial infections (NI), urinary tract infections (UTIs), nosocomial pneumonia (NP)**, and other inflammatory diseases.

About Antibiotic Resistance



Evolution of AMR community

Resistance Mechanism



Weapons of super-bacteria

To design an *in-silico* resource to discern diversity of antibiotic resistance genes in various -omics datasets

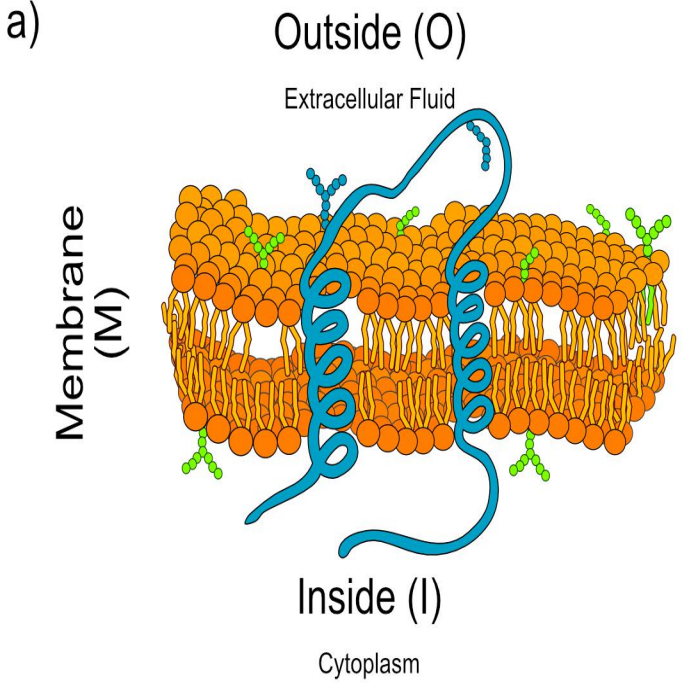
Pandey D. et. al. ***Biology Methods &***

Problem Statement & Genesis of Present Resource

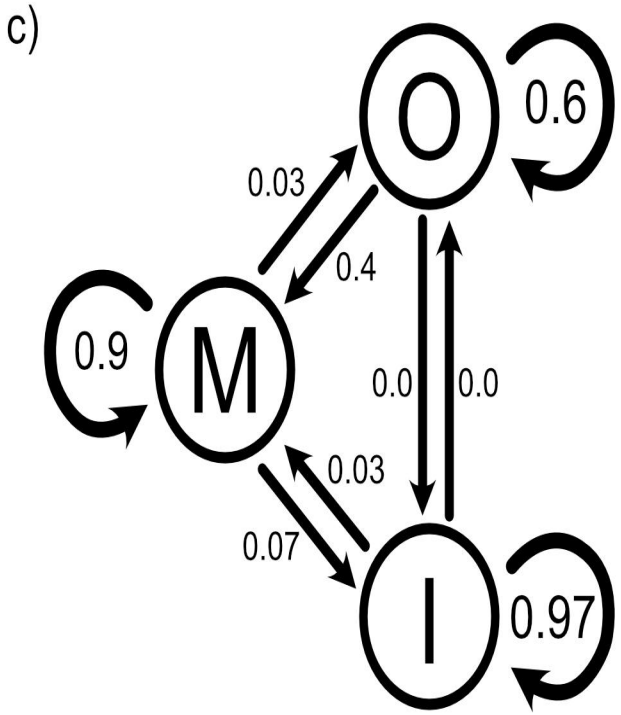
- The prominent ones being Antibiotic Resistance Genes Database (ARDB), Comprehensive β-lactamase Molecular Annotation Resource (CBMAR), ResFinder, Comprehensive Antibiotic Resistance Database (CARD), Resfams, Metagenomic Markov models for Antimicrobial Resistance Characterization (Meta-MARC), Antimicrobial Resistance Gene Finder (AMRFinderPlus) etc.

Limitations associated with previous methods such as:

1. ARDB is no longer updated (**Last Update: 2009**), and its data is incorporated in the CARD database.
2. Resfams is a database of **hidden Markov models (HMMs) developed using the 166 protein families associated with antibiotic resistance** (**Last Update: 2014**).
3. Meta-MARC is based on hierarchical HMMs, which can predict AMR in metagenomic data (either a short read or a longer assembled contig) into resistance class, group, and mechanism. But Meta-MARC result indicated **high false positive prediction and no user-friendly interface is available**.
4. AMRFinderPlus identifies acquired AMR genes and resistance-associated point mutations in protein or assembled nucleotide sequences. **But this tool is difficult for non-programmers**.
5. CARD identifies and annotates ARGs using **BLAST**. **But sequence alignment methods like BLAST work well in comparing sequences with a high degree of similarity (60% or higher) but do not identify a distant homolog.**
6. Also, we found that several resources can **identify/characterize resistance**


a)
Outside (O)
Extracellular Fluid
Membrane (M)
Inside (I)
Cytoplasm


b)
...M E I ...G I V...V D G...
...O O O...M M M...I I I ...


c)
O
M
I
0.6
0.03
0.4
0.9
0.0
0.0
0.03
0.07
0.97

Classification of whole sequences

Given:

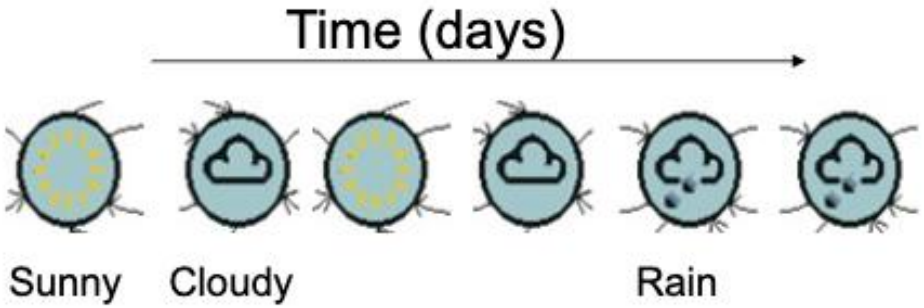
- a set of classes C and
- a number of example sequences in each class,

train a model so that for an unseen sequence we can say to which class it belongs

Example:

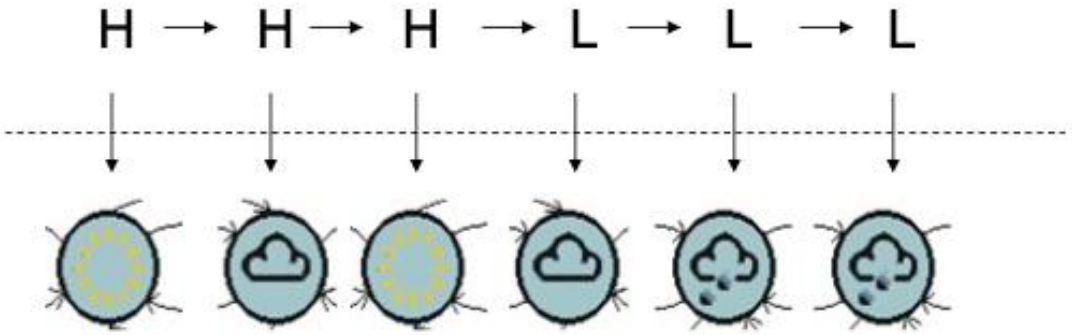
- Given a set of protein families, find family of a new protein
- Given a sequence of packets, label session as intrusion or normal
- Given several utterances of a set of words, classify a new utterance to the right word

Markov Chains



States : Three states - sunny, cloudy, rainy.

Hidden Markov Models



Hidden states : the (TRUE) states of a system that may be described by a Markov process (e.g., High of low pressure systems).

Observable states : the states of the process that are `visible' (e.g., weather).

Components Of HMM

Output matrix : containing the probability of observing a particular observable state given that the hidden model is in a particular hidden state.

Initial Distribution : contains the probability of the (hidden) model being in a particular hidden state at time $t = 1$.

State transition matrix : holding the probability of a hidden state given the previous hidden state.

Multiple alignment

```
A C A - - - A T G
T C A A C T A T C
A C A C - - A G C
A G A - - - A T C
A C C G - - A T C
```

Consensus:

How to distinguish:

```
T G C T - - A G G
A C A C - - A T C
```

Protein Profile HMMs

- Motivation
  - We want an efficient representation of motives.
- What is a Profile?
  - Patterns of conservation, some positions are more conserved than the others

Components Of HMM

Output matrix : containing the probability of observing a particular observable state given that the hidden model is in a particular hidden state.

Initial Distribution : contains the probability of the (hidden) model being in a particular hidden state at time $t = 1$.

State transition matrix : holding the probability of a hidden state given the previous hidden state.

Multiple alignment

```
A C A - - - A T G
T C A A C T A T C
A C A C - - A G C
A G A - - - A T C
A C C G - - A T C
```

Consensus:

How to distinguish:

```
T G C T - - A G G
A C A C - - A T C
```

Protein Profile HMMs

- Motivation
  - We want an efficient representation of motives.
- What is a Profile?
  - Patterns of conservation, some positions are more conserved than the others

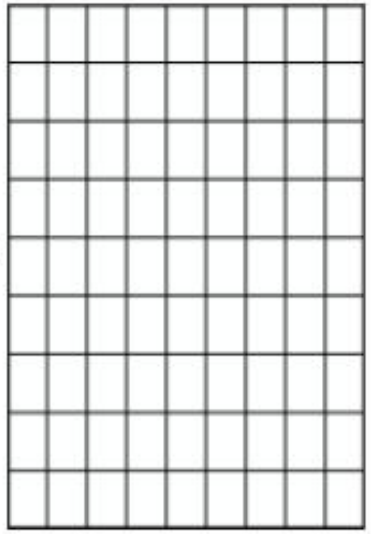
Protein profile representation

Alignment → Profile

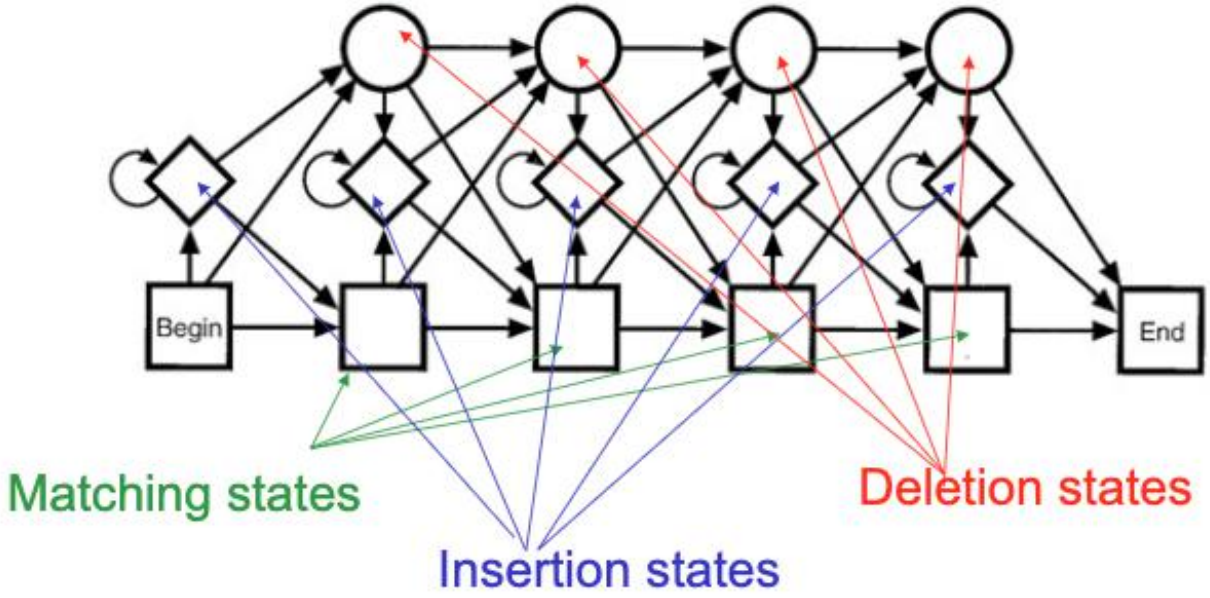
```
HCK_HUMAN/80-135        IIVVALYDYEAIH..HEDLSFQKGDQMVVLEES.G...EWWKARSLATR...KEGYIPSNYVARV
LYN_HUMAN/65-120        DIVVALYFYDGIN..PDDLSFKKGERMKVLEEH.G...EWWKAKSLLTK...KEGFIPSNYVAKL
BLK_MOUSE/54-109        RFVVALFDYAAVN..DRDLQVLKGEKLQVLRST.G...DWWLARSLVTG...REGYVPSNFVAPV
SR64_DROME/98-154       RVVVALYDYKSRD..ESDLSFMKGDRMEVIDDTES...DWWRVVNLTTR...QEGLIPLNFVAEE
CRKL_HUMAN/126-181      EYVRTLVDFPGND..AEDLPFKKGEILVIIEKPEE...QWWSARNKDG....RVGMIPVPYVEKL
NCK1_HUMAN/193-250      HVVQALYPFSSSN..DEELNFEKGDVMDVIEKPEND.PEWWKCRKING....MVGLVPKNYVTVM
MYS3_YEAST/1124-1181    PKFEAAYDFPGSG.SSSELPLKKGDIVFISRDEPS...GWSLAKLLDGS...KEGWVPTAYNTPY
ABL_DROME/207-263       QLFVALYDFQAGG..ENQLSLKKGEQVRILSYNKS..GEWCEAHSDSGN....VGWVPSNYVTPL
ABL1_HUMAN/64-119       NLFVALYDFVASG..DNTLSITKGEKLRVLGYNHN..GEWCEAQTKNG.....QGWVPSNYITPV
SPK1_DUGTI/36-92        YNVKAKYKYAASG..DTDISFEEKEIMYVLEQFDE...FWLKVVKQKDN...KEGLVPSNYVSKQ
SSB1_YEAST/303-359      YKAKALYPYDADDDDAYEISFEQNEILQVSDIEGR....WWKARRANG....ETGIIPSNYVQLI
DRK_DROME/1-56          MEAIAKHDFSATA..DDELSFRKTQILKILNMEDD..SNWYRAELDGK.....EGLIPSNYIEMK
SEM5_CAEEL/1-56         MEAVAEHDFQAGS..PDELSFKRGNTLKVLNKDED..PNWYKAELDGN.....EGFIPSNYIRMT
CSK_CHICK/12-68         TECIAKYNFRGTA..EQDLPFSKGDVLTIVAVTKD..PNWYKAKNKV..G..REGIIPANYVQKR
NCK1_HUMAN/5-59         VVVVAKFDYVAQQ..EQELDIKKGERLWLLDDSKS....WWRVRNSNH....KTGFVPSNYVERK
SPCA_DROME/973-1027     ECVVALYDYTEKS..PREVSMKKGDVLTLLNSNNK...DWWKVEVN...D..RQGFVPAAYIKKI
SPCA_HUMAN/980-1034     QRVNALYDFQARS..PREVTMKKGDVLTLLSSINK...DWWKVEAA...D..HQGIVPAVYVRRL
SCD2_SCHPO/27-84        KVIRALYDYTARK..ATEVSFAKGDFFHVIGREND..KAWYEVCNPAAG...TRGFVPVSHFEEI
BEM1_YEAST/75-130       KVIKAKYSYQAQT..SKELSFKEGEFFYVSGDEKD....WYKASNPSTG...KEGVVPKTYFEVF
SLA1_YEAST/72-130       KKVRAIYDYEQVQNADELTFHENDVFDVFDDKDA...DWLLVKSTVSN...EFGFIPGNYVEPE
RSG1_BOVIN/279-336      RSVRAILPYTKVP.DTDEISFLKGDMFIVHSELEDG.WGWVTNLRTD.....EQGLIVEDLVEEV
YKA7_CAEEL/194-250      FYGIAKFDYAPTQ..SDEMGLRIGDTVLISKKVDA...EWFYGENQRQR...TFGIVPSSYLDIK
STE6_SCHPO/3-58         FQSTAISDYEHSS.NPSFLKFSAGDTIIVIEVLED...GWCDGICSE.....KRGWFPTSCIDSS
NCF1_HUMAN/159-213      QTYRAIADYEKTS..GSEMALSTGDVVEVVEKSES...GWWFCQMKA.....KRGWIPASFLEPL
NCF1_HUMAN/229-283      EPYVAIKAYTAVE..GDEVSLLEGEAVEVIHKLLD...GWWVIRKDD.....VTGYFPSNYLQKS
SCD2_SCHPO/126-183      LFGIVQFDFAAER..PDELEAKAGEAIIIIARSNM...EWLVAKPIGRLG..GPGLIPLSFIQLR
BEM1_YEAST/158-215      LYAIVLYDFKAEK..ADELTIYVGEMLFICAHNNC...EWFIAKPIGRLG..GPGLVPVGFVSII
SLA1_YEAST/356-413      KRGIVQYDFMAES..QDELTIKSGDKVYILDDKKS..KDWWMCQLVDSG...KSGLVPAQFIEPV
NCF2_HUMAN/243-297      EAKRVLFGFVPET..KEELQVMPGNIVFVLKKGND...NWATVMFNG.....QKGLVPCNYLEPV
BOI2_YEAST/46-105       PNYIAINEYFKRN..EDELDMKPGDKIKVITDDEEYKDGWYFGRNLRTH...EEGLYPVVFTQKI
BZI1_YEAST/496-553      GKNKVLYAYVQKD..DDEITITPGDKISLVARDTG..SGWTKINNDTTG...ETGLVPTTYIRIS
```

Amino Acids

Positions



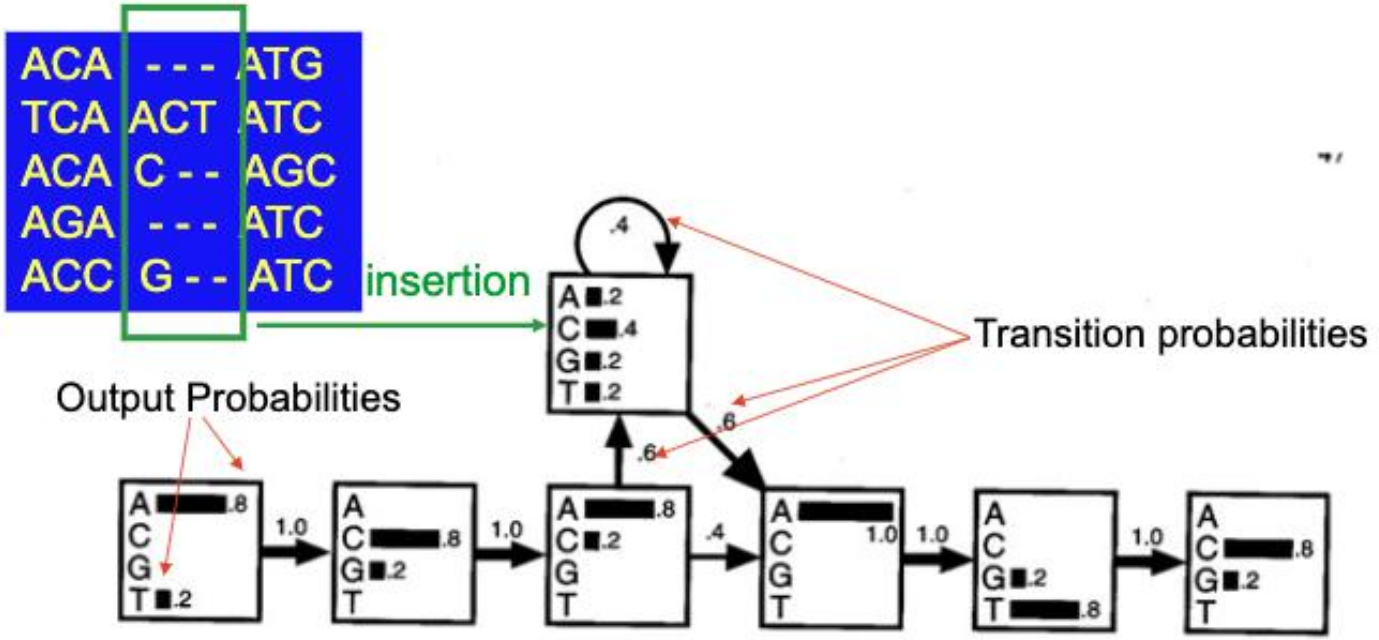
HMM topology (Hidden states)



No of matching states = average sequence length in the family

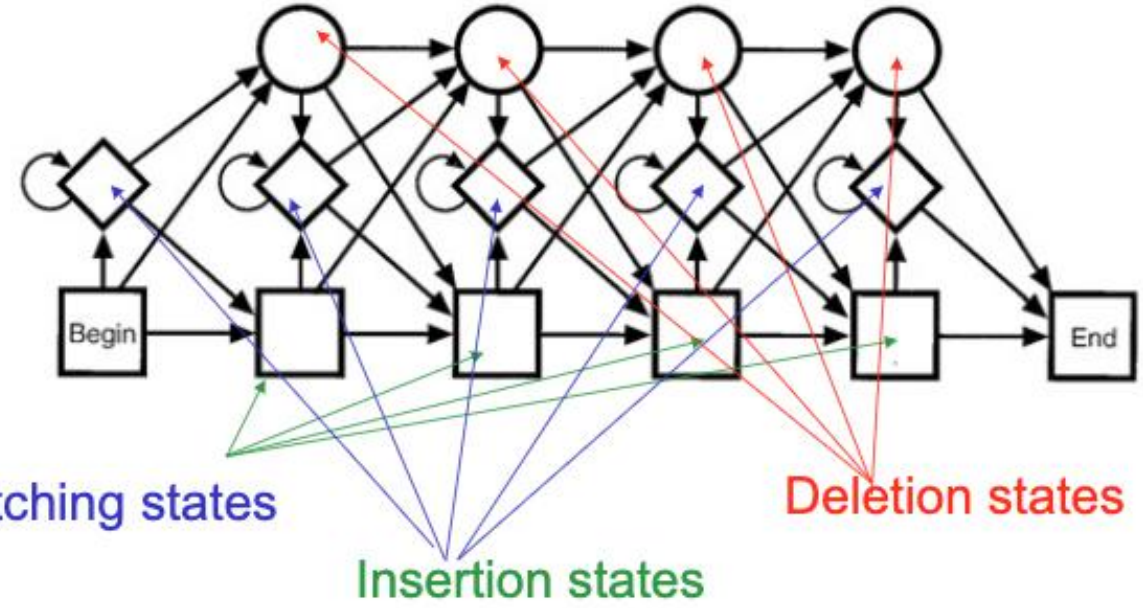
PFAM Database - of Protein families
(pfam.wustl.edu)

Building – *from an existing alignment*



A HMM model for a DNA motif alignments, The transitions are shown with arrows whose thickness indicate their probability. In each state, the histogram shows the probabilities of the four bases.

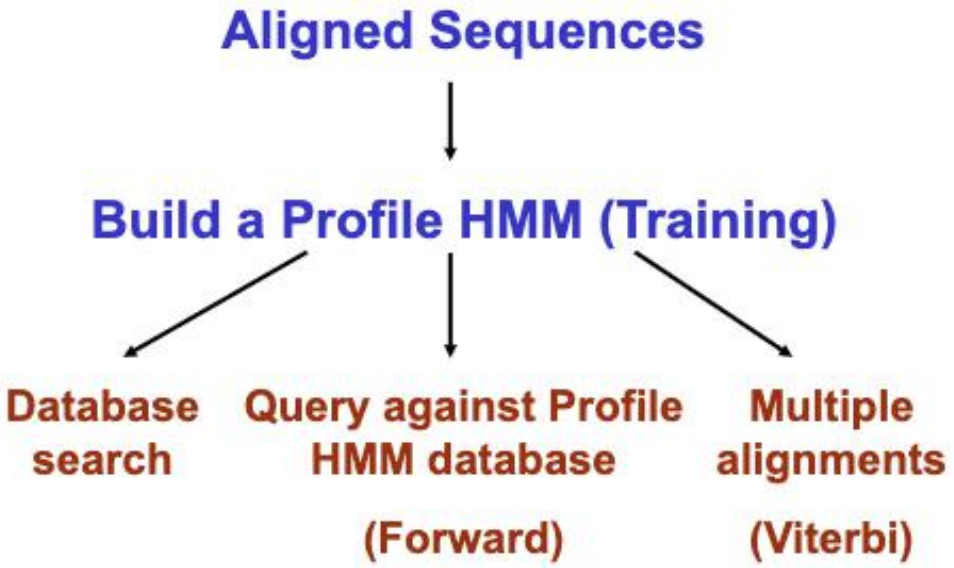
Building – *Final Topology*



No of matching states = average sequence length in the family

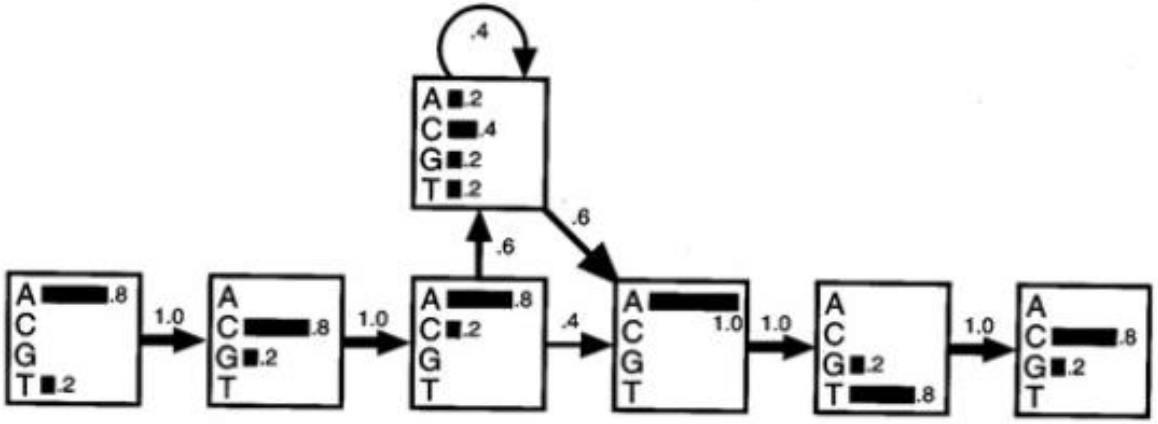
PFAM Database - of Protein profiles
(http://pfam.wustl.edu)

An Overview


Aligned Sequences
Build a Profile HMM (Training)
Database search
Query against Profile HMM database
(Forward)
Multiple alignments
(Viterbi)

Query a new sequence

Suppose I have a query sequence, and I am interested in which family it belongs to?



Consensus sequence: ACAC - - ATC

P (ACACATC) = 0.8x1 x 0.8x1 x 0.8x0.6 x 0.4x0.6 x 1x1 x 0.8x1 x 0.8 = 4.7 x 10^{-2}

Scoring

P (ACACATC) = 0.8x1 x 0.8x1 x 0.8x0.6 x 0.4x0.6 x 1x1 x 0.8x1 x 0.8 = 4.7×10^{-2}

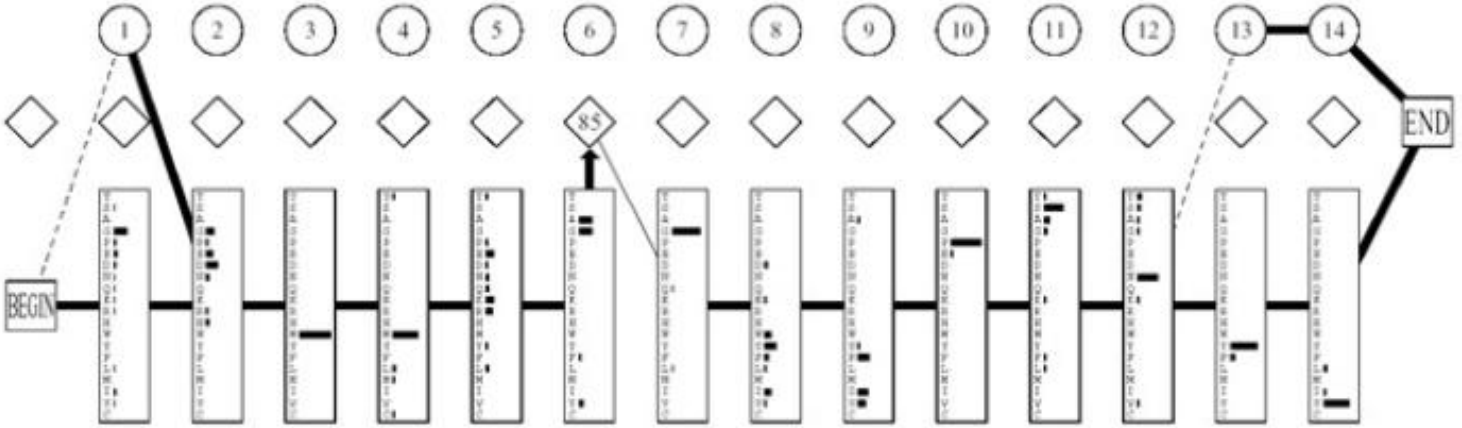
$$\text{log-odds for sequence } S = \log \frac{P(S)}{0.25^L} = \log P(S) - L \log 0.25$$

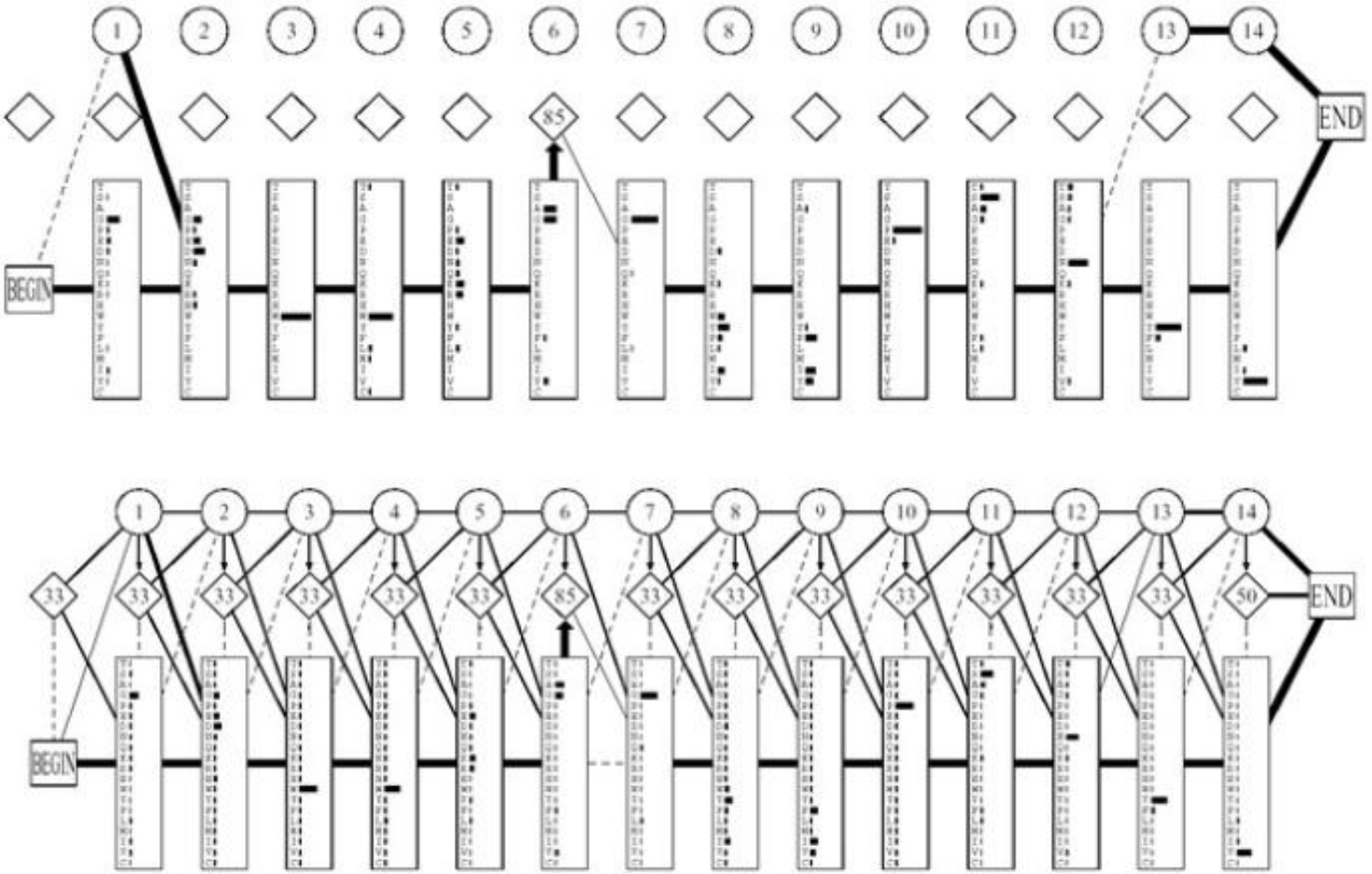
$$\begin{aligned} \text{log-odds}(\texttt{ACACATC}) &= 1.16 + 0 + 1.16 + 0 + 1.16 - 0.51 + \\ &\quad 0.47 - 0.51 + 1.39 + 01.16 + 0 + 1.16 \\ &= 6.64. \end{aligned}$$

PHMM Example

An alignment of 30 short amino acid sequences chopped out of a alignment of the SH-3 domain. The shaded area are the most conserved and were represented by the main states in the HMM. The unshaded area was represented by an insert state.

```
GGWWRGdy.ggkkqLWFPSNYV
IGWLNGynetteergDFPGTYV
PNWWEGql..nnrrGIFPSNYV
DEWWQArr..deqiGIVPSK--
GEWWKAqs..tgqeGFIPFNFV
GDWWLArs..sgqtGYIPSNYV
GDWWDAel..kgrrGKVPSNYL
-DWWEArslssghrGYVPSNYV
GDWWYArslitnseGYIPSTYV
GEWWKArslatrkeGYIPSNYV
GDWWLArslvtgreGYVPSNFV
GEWWKAkslsskreGFIPSNYV
GEWCEAqt.kngq.GWVPSNYI
SDWWRVvnlttrqeGLIPLNFV
LPWWRArd.kngqeGYIPSNYI
RDWWEFrsktvytpGYYESGYV
EHWWKVkd.alqnvGYIPSNYV
IHWWRVqd.rnqheGYVPSSYL
KDWWKVev..ndrqGFVPAAYV
VGWMPGlnertrqrGDFPGTYV
```



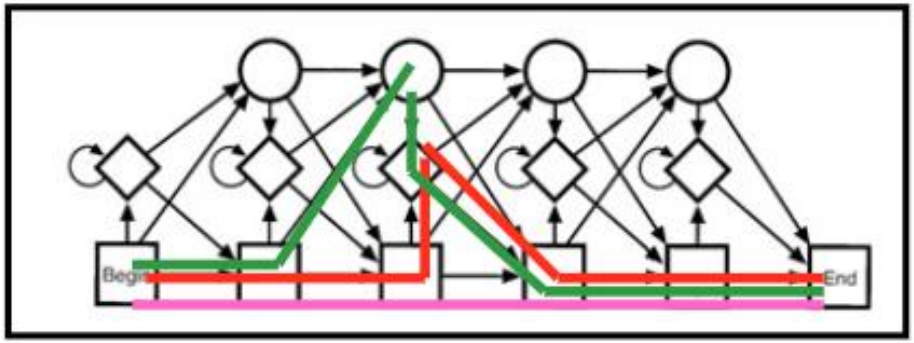

1
2
3
4
5
6
7
8
9
10
11
12
13
14
85
BEGIN
END
1
2
3
4
5
6
7
8
9
10
11
12
13
14
33
33
33
33
33
33
85
33
33
33
33
33
33
33
50
BEGIN
END

Database Searching

- Given HMM *M*, for a sequence family, find all members of the family in data base.

Multiple Alignments

- Try every possible path through the model that would produce the target sequences
  - Keep the best one and its probability.
  - Output : Sequence of match, insert and delete states
- Viterbi alg. Dynamic Programming



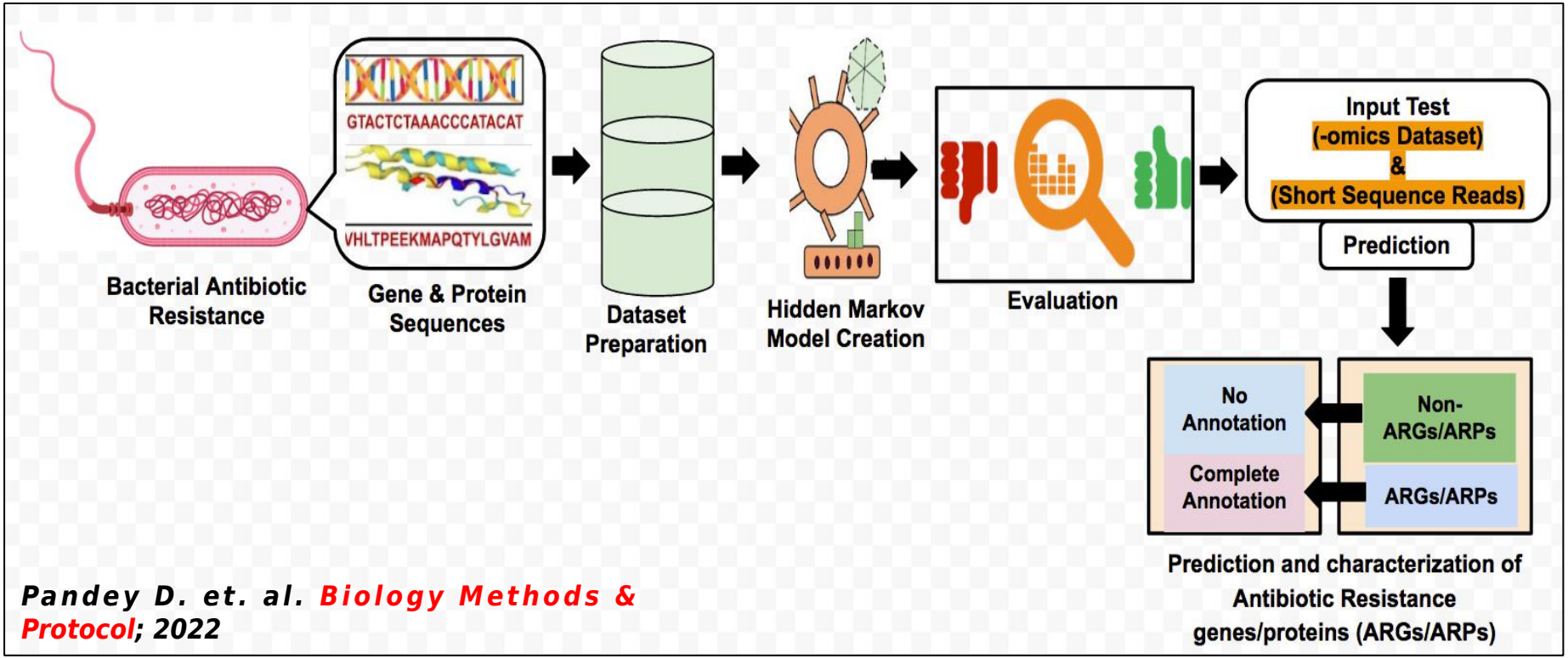
Advantages

- Characterize an entire family of sequences.
- Position-dependent character distributions and position-dependent insertion and deletion gap penalties.
- Built on a formal probabilistic basis
- Can make libraries of hundreds of profile HMMs and apply them on a large scale (**whole genome**)

Motivation

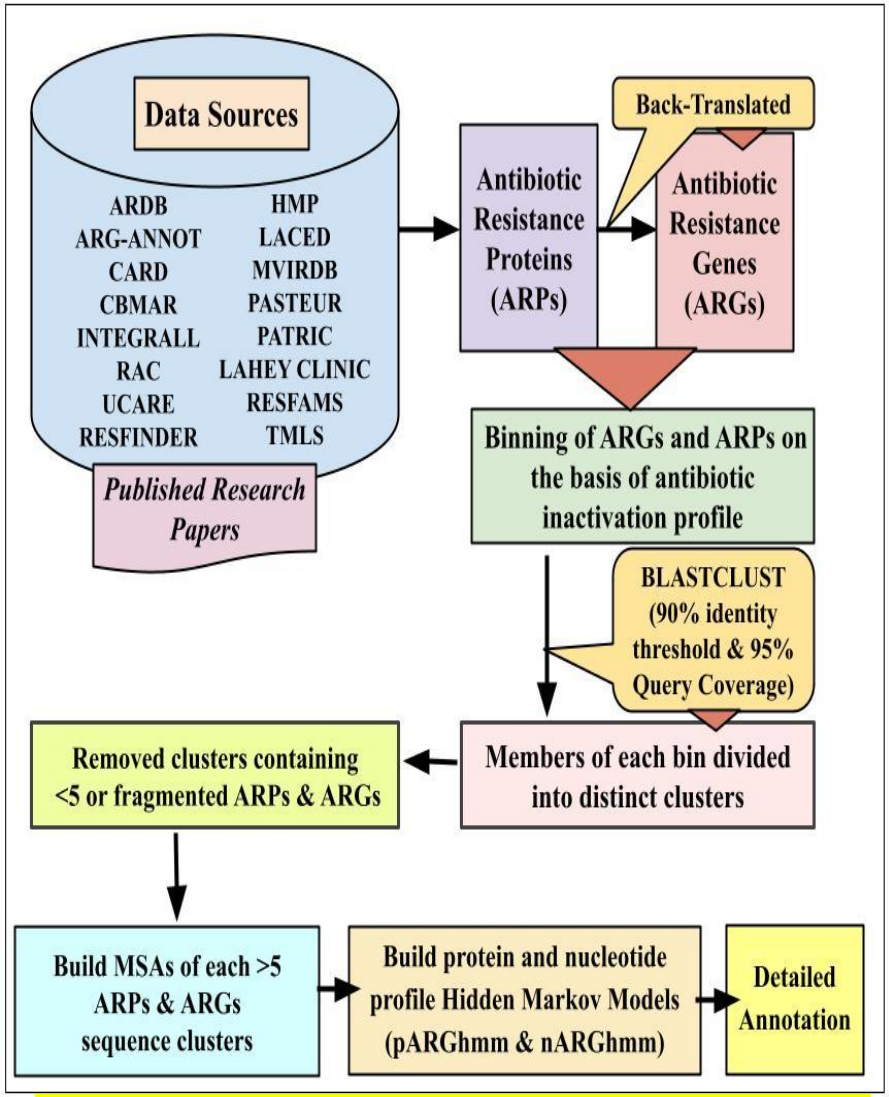
- Thus, we have described a new *in-silico* tool for **rapid monitoring, characterization, and surveillance** of **all bacterial antibiotic resistance genes (ARGs)** which named as **Bac**terial **A**ntibiotic **R**esistance **scan** (BacARscan).
- **This tool has the edge over its predecessors as it can also discern ARGs in short sequencing reads and fragmented contigs.**
- BacARscan can be easily integrated into a **user-defined ARG annotation pipeline for the detection of ARG variants in the microbial genomes.**

Schema of the tool



Pandey D. et. al. Biology Methods & Protocol; 2022

Workflow & Data Statistics



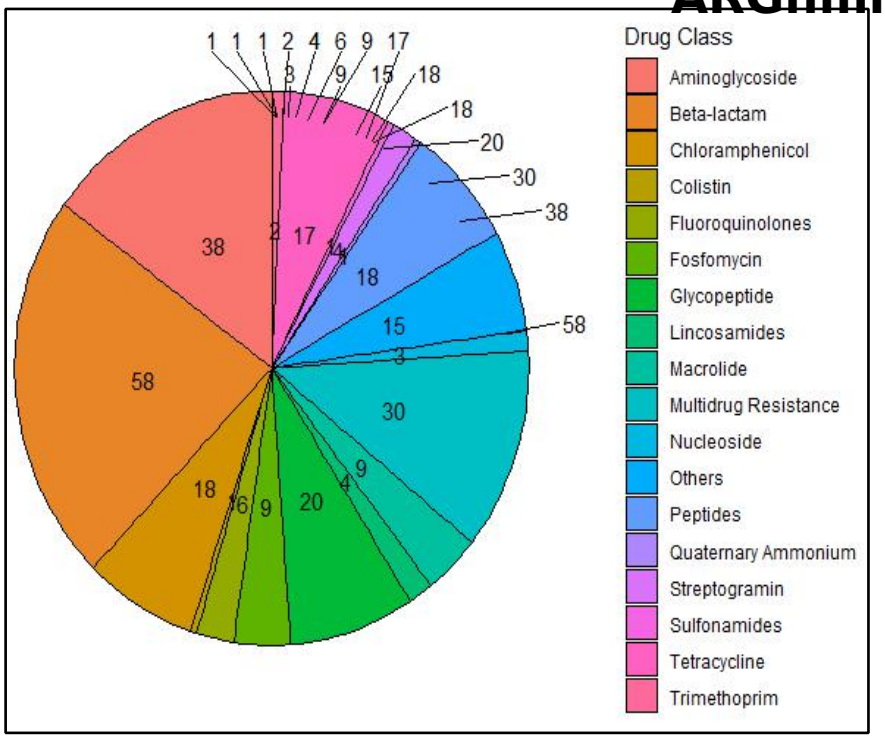
Methodology used for the development of BacARscan

S. No.	ARG Database	Number of Sequences	
		Before redundancy reduction	After redundancy reduction (Duplicates removed)
1.	ARDB	7828	7825
2.	ARG	1689	1601
3.	CARD	2158	2155
4.	CBMAR	3273	3273
5.	INTEGRALL	11132	11132
6.	RAC	6911	6911
7.	TMLS	1983	1983
8.	UCARE-DB	99	99
9.	LAHEY CLINIC	3562	3562
10.	RESFAM	3169	1745
11.	RESFINDER	2156	2008
12.	HMP	7828	7825
13.	LACED	483	448
14.	MVIR-DB	64711	61469
15.	PASTEUR	1123	1123
Total AR gene/protein sequences		118105	113159

Data Statistics of Antibiotics resistance gene/protein sequences retrieved from various ARG databases; before and after redundancy reduction

Pandey D. et. al. Biology Methods & Protocol; 2022

Functional annotation of protein (p) and nucleotide (n) ARGhmm



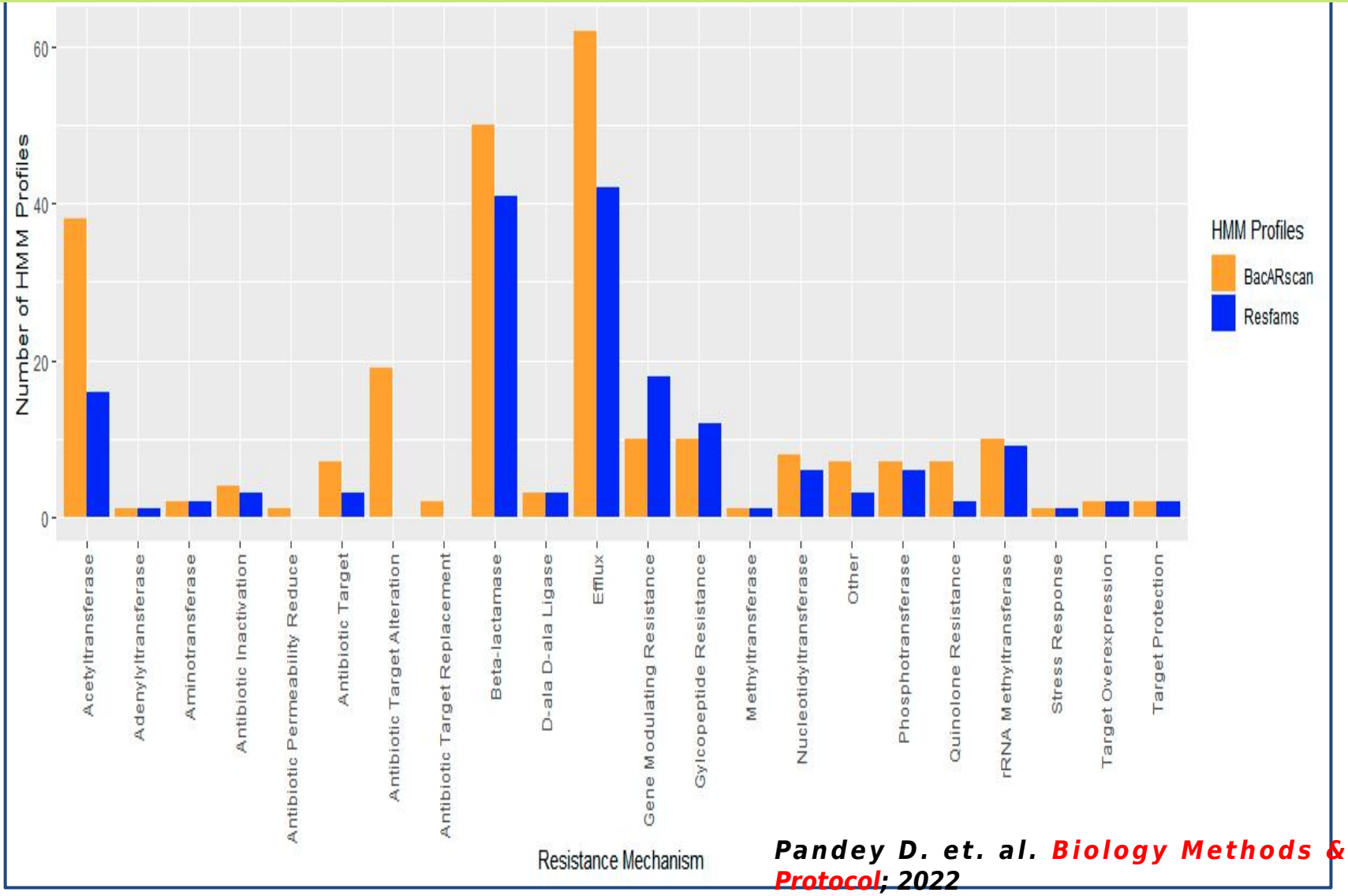
Distribution of ARG HMM profiles into various antibiotic classes. The numerical value indicates the number of HMM that inactivates the antibiotic

ARGhmm profiles includes:

- Class and subclass of antibiotics against which the query proteins/genes impart resistance
- Resistance mechanism
- Antimicrobial resistance spectrum
- AMR protein name & families
- Function of AMR genes
- UniProt ID against each ARGhmm

Pandey D. et. al. Biology Methods & Protocol; 2022

Comparison of Resfams and BacARscan profiles – HMM models on the basis of their resistance mechanism



Pandey D. et. al. Biology Methods & Protocol; 2022

Benchmark Datasets

Dataset-I: Evaluation dataset

Positive dataset: Protein clusters ≥ 5 sequences

Negative dataset: Protein clusters < 5 sequences

Dataset-II: Short sequence reads

From back-translation of positive and negative data

Length: 100 nt

Each sequence: 20 short

Dataset-III:Independent dataset

1) 60 Penicillin-binding proteins (PBPs) or DD-peptidase

2) 369 Non-antibiotic resistant bacterial efflux Pandey et al., Sci. Rep. 2020)

Dataset-IV: Annotation of ARGs in different strains of ESKAPE pathogens

Five proteomes of each organism of ESKAPE

Dataset VI: Clinical

metagenomic data

16 metagenomic samples from human patients of cholecystectomy, six from human bile and five from gut and saliva each. (Kujiraoka et al., 2017; Frontiers in Microbiology).

(DDBJ Accession: DRA005134).

Dataset-V: Validation Dataset

To benchmark BacARscan vis-a-vis other ARG prediction and annotation methods.

Source: CARD database (date: 29-07-2022) **4422 ARG sequences**.

Short reads of 151nt length at 20x coverage were simulated. 100,000 short reads randomly selected for benchmarking.

Simulated Non-ARG short-read Data:

Source: 2 million short-reads from complete genome of a probiotic strain of *Enterococcus faecium* Strain T-110 (NCBI Genome Accession Number: CP006030) Natarajan et.al. (2015)

The comparative evaluation was carried out among BacARscan, Meta-MARC, and ResFinder

Pandey D. et. al. Biology Methods & Protocol; 2022

Performance of BacARscan (pARGhmm & nARGhmm)

Dataset-I: Evaluation dataset

Modules	pARGhmm				nARGhmm			
Parameters / No. of top hits	True Positive	False Positive	Precision(%)	F-measure (%)	True Positive	False Positive	Precision(%)	F-measure (%)
1	228	26	89.76%	94.60%	231	23	90.94%	95.25%
3	229	25	90.15%	94.82%	235	19	92.51%	96.11%
5	**234**	**20**	**92.12 %**	**95.90%**	237	17	93.30%	96.53%
7	233	21	91.73%	95.68%	236	18	92.91%	96.32%
9	232	22	91.33%	95.47%	240	14	94.48%	97.16%
11	209	45	82.28%	90.28%	**241**	**13**	**94.88%**	**97.37%**
13	182	72	71.65%	83.48%	240	14	94.48%	97.16%
15	158	96	62.20%	76.69%	238	16	93.70%	96.74%

Pandey D et. al. Biology Methods & Protocol; 2022

Comparison of proposed method BacARscan with existing methods using homologous sequences

Dataset-III:Independent dataset

Method	Type of dataset used	True Negative	False Positive	True Negative Rate (%)	False Positive Rate (%)
BacARscan	**Penicillin-binding proteins (PBPs)**	**54**	**06**	**90%**	**10%**
AMRFinderPlus		48	12	80%	20%
Meta-MARC		51	09	85%	15%
RGI-CARD		45	15	75%	25%
Resfams		56	04	93.33%	6.67%
BacARscan	**Non-antibiotic efflux proteins (non-ARE)**	**366**	**23**	**94.08%**	**5.91%**
AMRFinderPlus		352	37	90.48%	9.51%
Meta-MARC		363	26	93.31%	6.68%
RGI-CARD		298	91	76.60%	23.39%
Resfams		365	24	93.83%	6.16%

E-value Threshold	# of simulated reads	# of reads predicted (Hits Found)		
		BacARScan	Meta-MARC	ResFinder
1e-6	100000 AR Short Reads	58703	69294	88831
1e-3		66802	77667	89580
Default (10)		78680	89778	99875

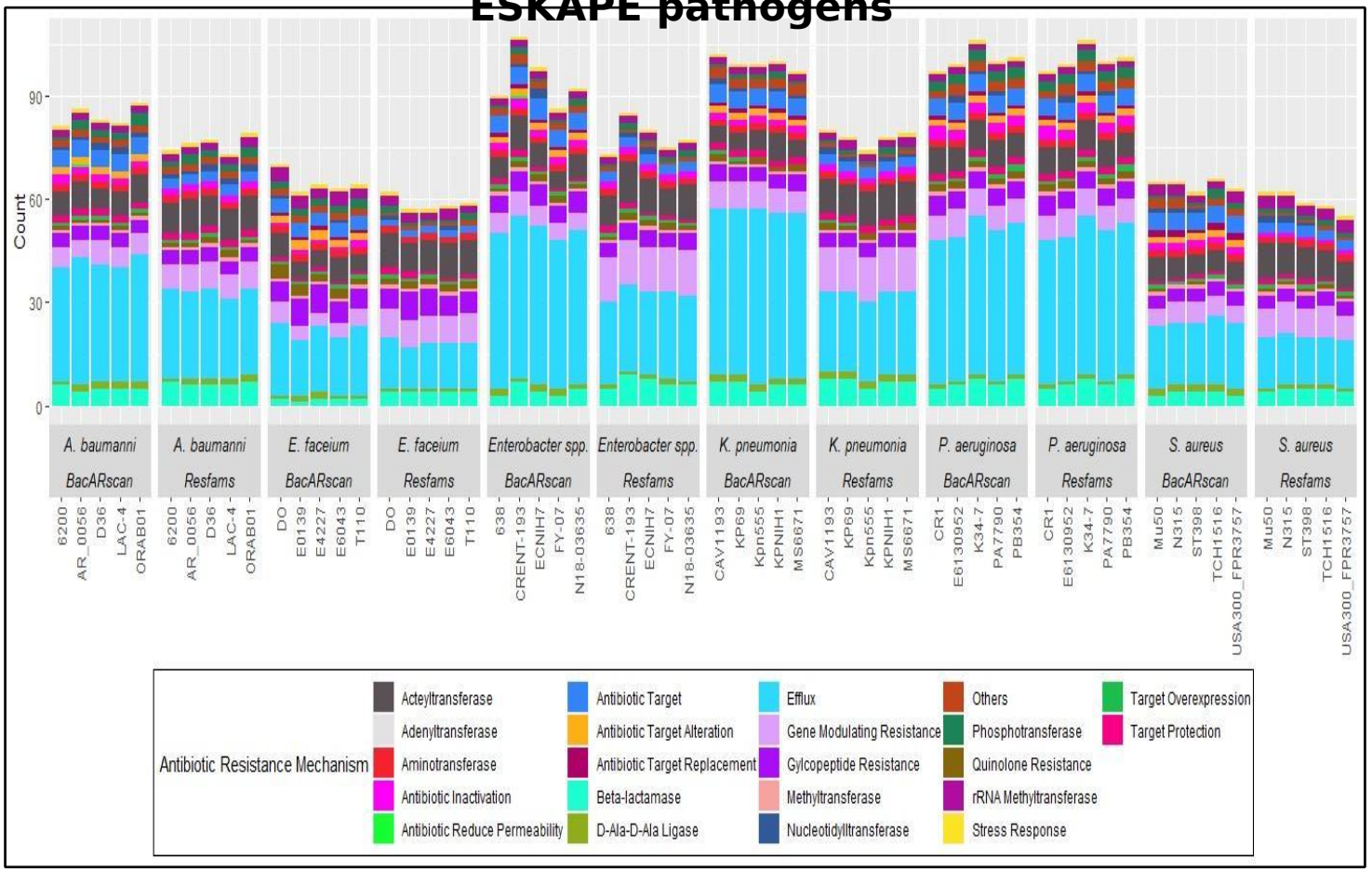
Performance of BacARscan and other off-the-shelf tools in predicting antibiotic resistance, an external test set of ARG short-read data

E-value Threshold	Tools	# of Reads Predicted (Hits Found) & Unique ARGs			
		# of Reads Predicted	# of Unique ARGs	False Positive Rate (%)	True Negative Rate (%)
1e-6	BacARscan	**3979**	**19**	**0.20%**	**99.80%**
	Meta-MARC	22331	56	1.12%	98.88%
	ResFinder	1912	5	0.10%	99.90%
1e-20	BacARscan	**238**	**3**	**0.02%**	**99.98%**
	Meta-MARC	9034	18	0.46%	99.54%
	ResFinder	1648	3	0.09%	99.91%
1e-50	BacARscan	**0**	**0**	**0**	**0**
	Meta-MARC	0	0	0	0
	ResFinder	1500	3	0.08%	99.92%

Performance of BacARscan and other off-the-shelf tools in predicting antibiotic resistance in an external test set of Non-ARG short-read data

Pandey D. et. al. Biology Methods & Protocol; 2022

Dataset-IV: Annotation of ARGs of 30 proteomes of ESKAPE pathogens

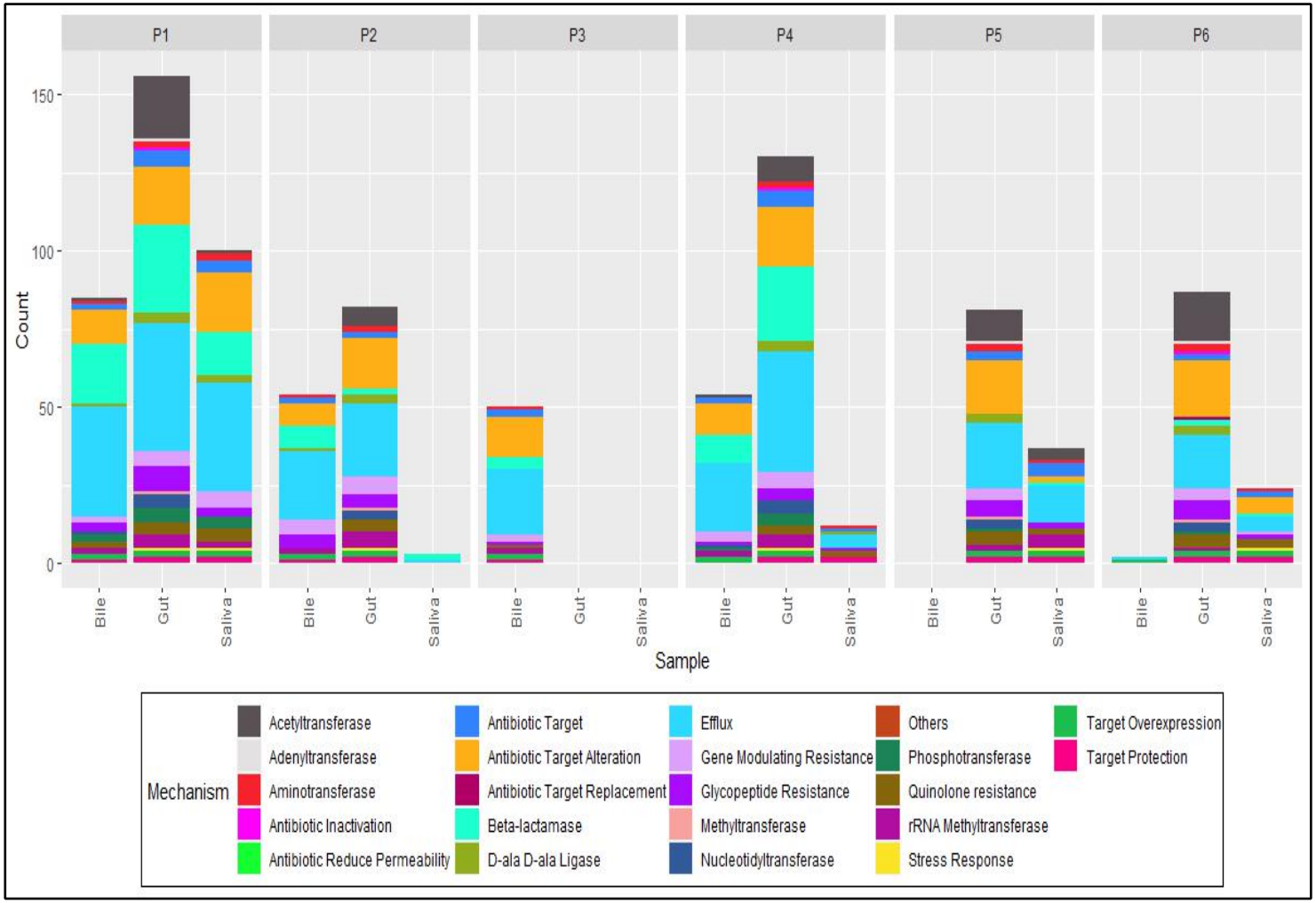


Comparison of prediction of ARGs and their resistance mechanism pattern between Resfams and BacARscan on ESKAPE pathogens

Pandey D. et. al. Biology Methods & Protocol; 2022

Dataset VI: Clinical metagenomic data

Comparative evaluation of prediction efficiency of BacARscan on metagenomic data



Kujiraoka, M. *et al. Front. Microbiol.* 8, 685 (2017)

P: Patient

Pandey D. et. al. Biology Methods & Protocol; 2022

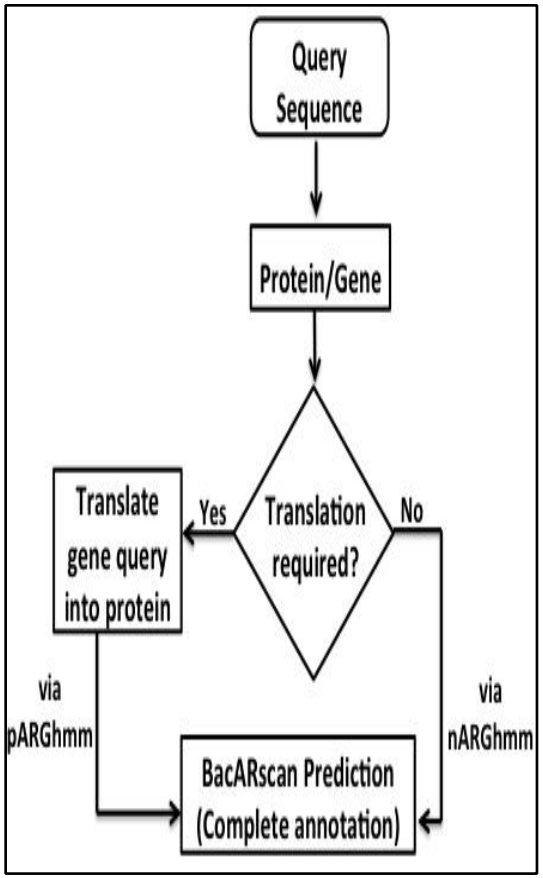
Web interface and standalone tool

- **In the BacARscan web tool, a user has the option to choose between query sequence type and nature of HMM-profiles (either 'Protein/pARGhmm' or 'Gene/nARGhmm').**
- **The web platform of BacARscan can process 100 sequences at a time**
- For [illegible] ersion

Speed Assessment

- **30 complete proteomes (6 ESKAPE organisms × 5 different strains) containing 1,28,305 nearly 31 minutes** to complete the annotation of all **30 proteomes**, ~ **one minute per proteome**.

- **Intel(R) Xeon(R) 4 Core E5507 2.27 GHz processor with 6GB DDR4 RAM, 64-bit Red Hat Enterprise Linux operating system (Release 6.2)**.



Prediction schema of BacARscan

Pandey D. et. al, Biology Methods & Protocol; 2022

Web Interface

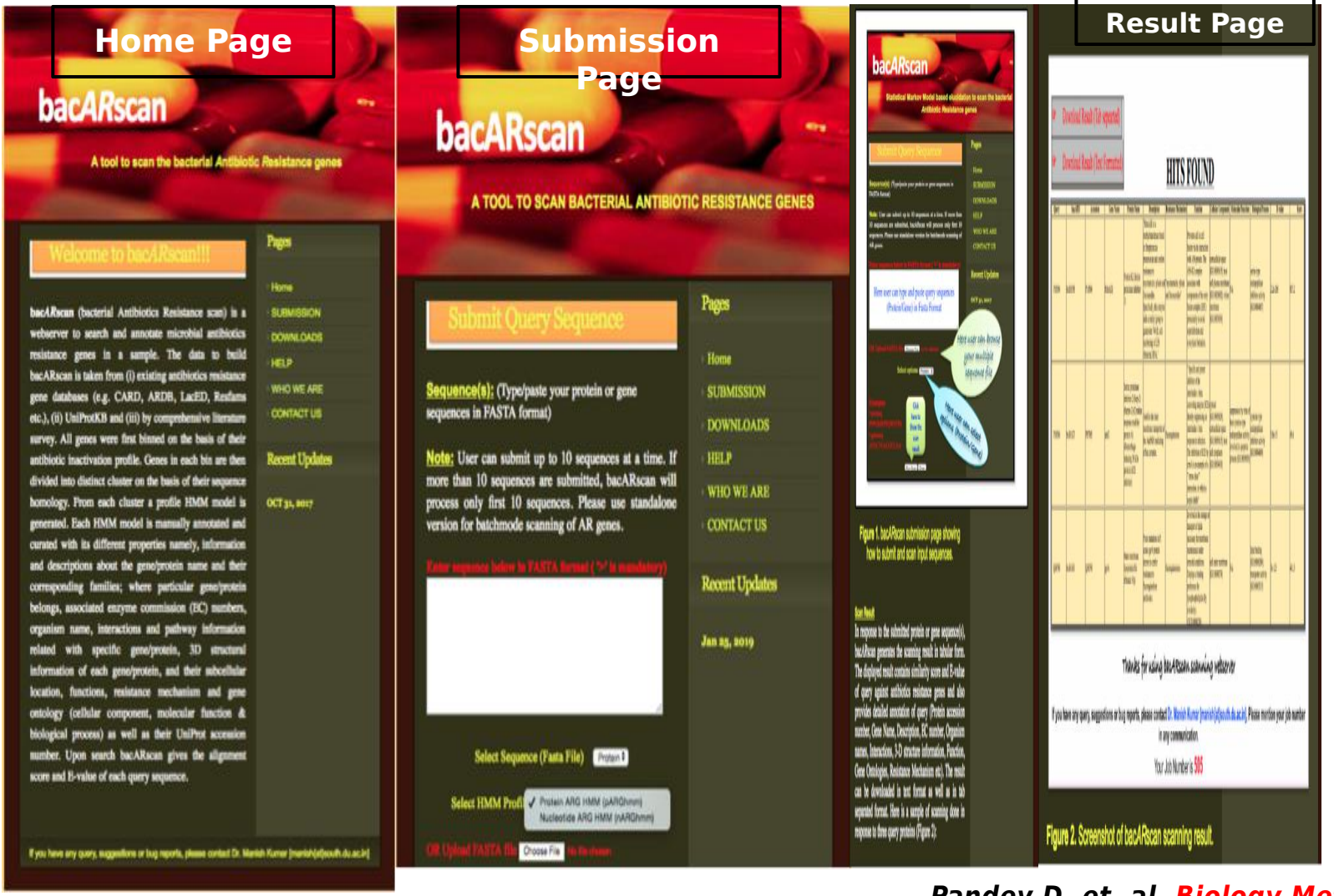


Figure 1 bacARscan submission page showing how to submit and scan input sequences.

Figure 2. Screenshot of bacARscan scanning result.

Web Link:
http://www.proteininformatics.org/mkumar/bacarscan/
Github Link:

Pandey D. et. al. Biology Methods & Protocol; 2022

Conclusions

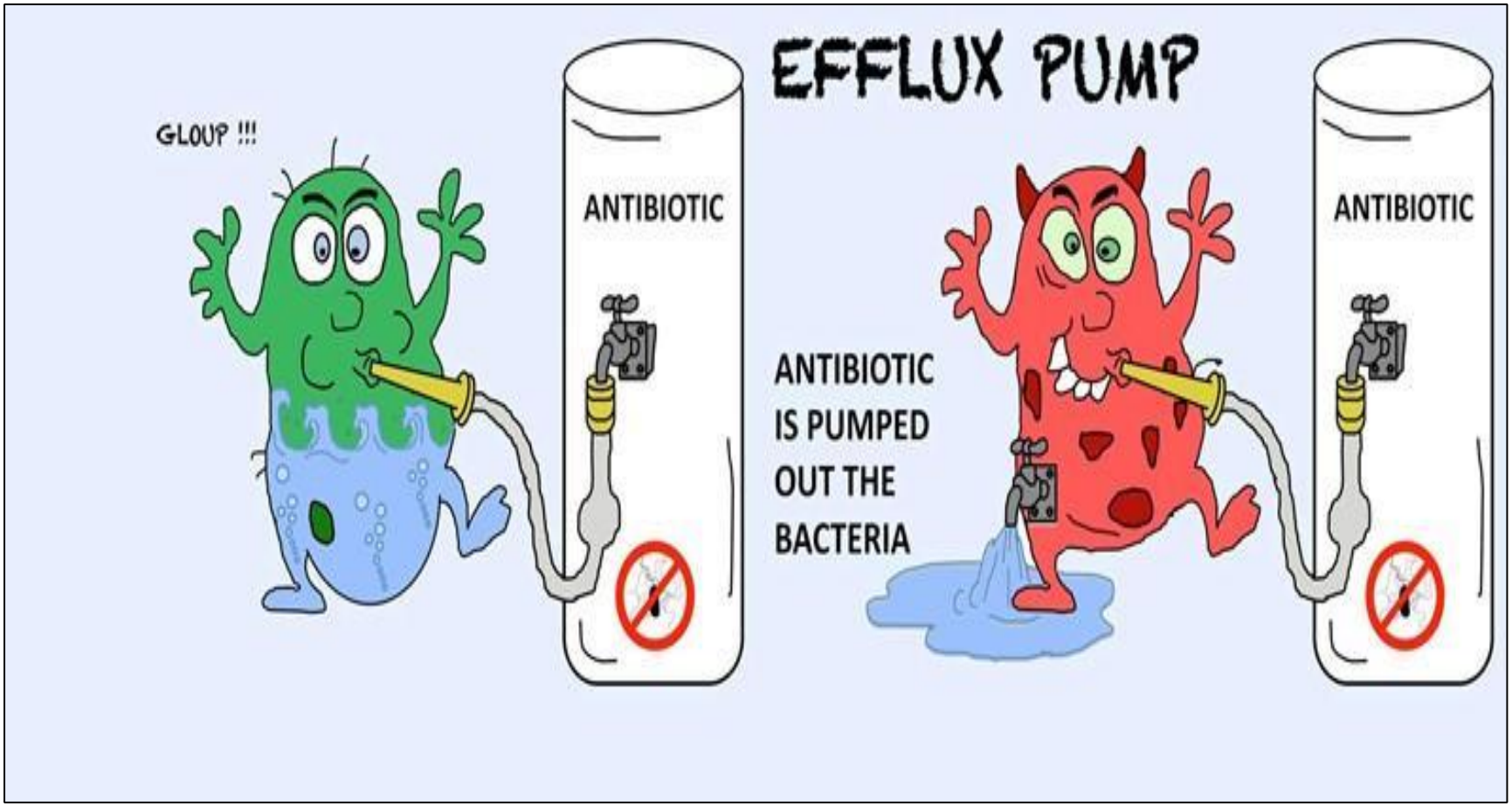
- **BacARscan: *in-silico* ARG annotation resource that can be used for rapid monitoring, surveillance, and characterization of antibiotic resistance determinants in both genomics and proteomic datasets**.
- **Current version of BacARscan supports prediction using 254 ARG families**.
- Comparison with other *in-silico* resources like **AMRFinderPlus, Meta-MARC, Resfams, and CARD revealed that BacARscan's ability to discern ARGs in -omics datasets was much more significant than its predecessors.** Also it indicated **less false positive prediction of ARG by BacARscan vis-a-vis other methods**.
- One of the most notable improvements of BacARscan over other ARG annotation methods is its ability to work on both genomes and short reads sequence libraries with equal efficiency and without any requirement for assembly of short reads.

Potential use of BacARscan

- Can identify ARGs in an -omics (proteomics/genomics and metagenomic) datasets.
- BacARscan can also be combined with traditional surveillance and thus can complement the traditional methods of ARG annotation.

Pandey D. et. al. Biology Methods & Protocol: 2022

To develop a two-tier system to predict and categorize bacterial efflux-mediated antibiotic resistance proteins and their families



Pandey D. et. al. Scientific Reports; 2020
Pandey D. et. al. Nature Protocol Exchange; 2021

Motivation

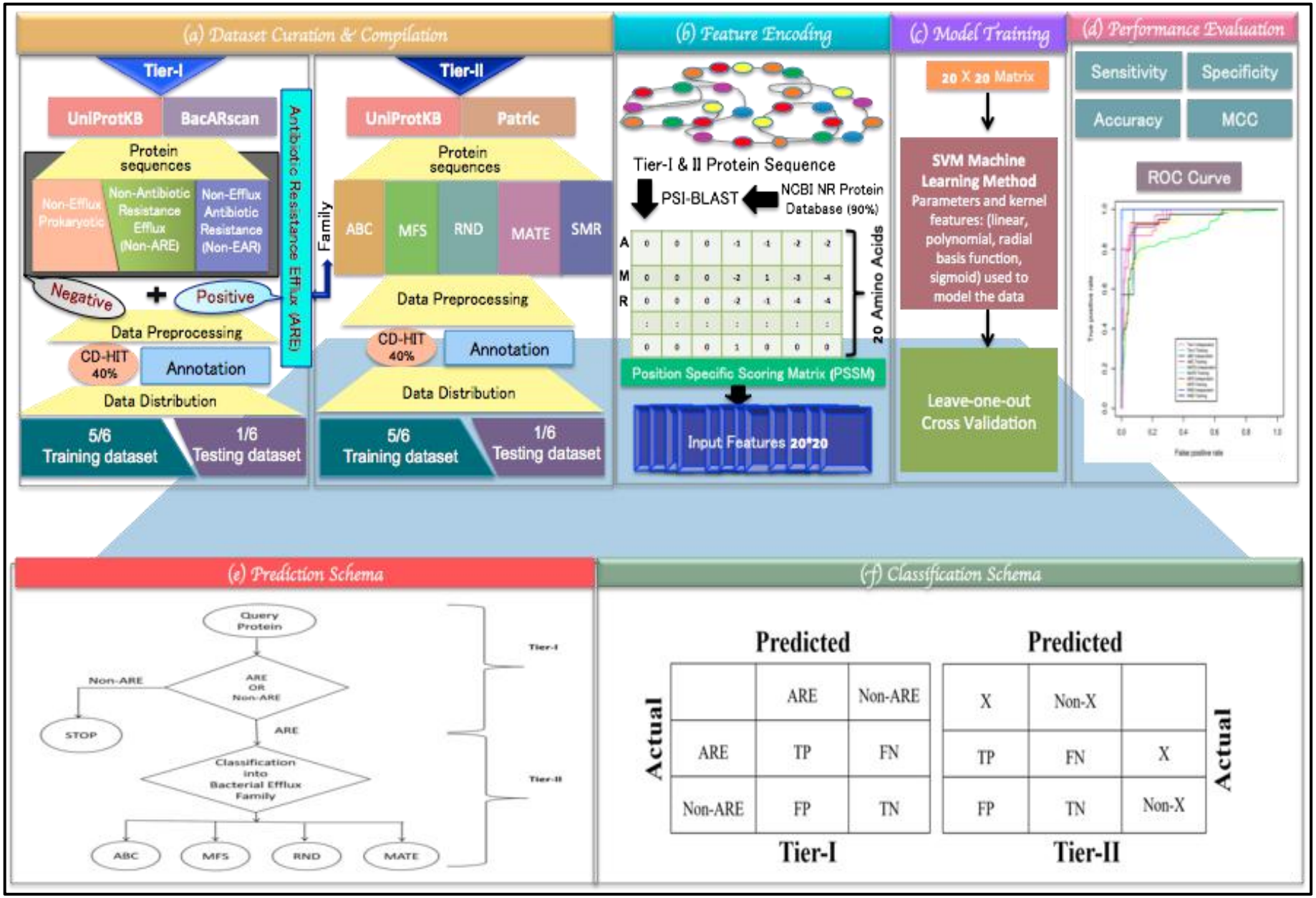
- **Efflux proteins are present in both Gram-positive and Gram-negative bacteria.**
- **Prokaryotic Efflux pumps are divided into five classes: Major facilitator superfamily (MFS), ATP -binding cassette (ABC) superfamily, Small multidrug resistance (SMR) family, Resistance- nodulation cell division (RND) superfamily, Multi- antimicrobial extension (MATE).**
- Efflux protein pumps constitute between 6-18% of all the transporters present in any bacterial species. Efflux pumps might be specific for one substrate or may transport a range of structurally dissimilar compounds **(including antibiotics of multiple classes). Efflux pumps were associated with multiple drug resistance (MDR) in bacteria.**
- **We could not find any *in-silico* tool that can discriminate bacterial antibiotic resistance efflux (ARE) proteins from efflux proteins which do not efflux out antibiotics (non-ARE), and/or can predict the family to**

- **BacEffluxPred:** a machine-learning based two-tier *in-silico* tool that **discriminates bacterial ARE proteins from non-ARE and also predicts its respective family**.
- BacEffluxPred completes a prediction cycle in **two different tiers**.
- **Tier-I: discrimination between ARE and non-ARE proteins**
- **Tier-II: prediction of ARE protein(s) family**.

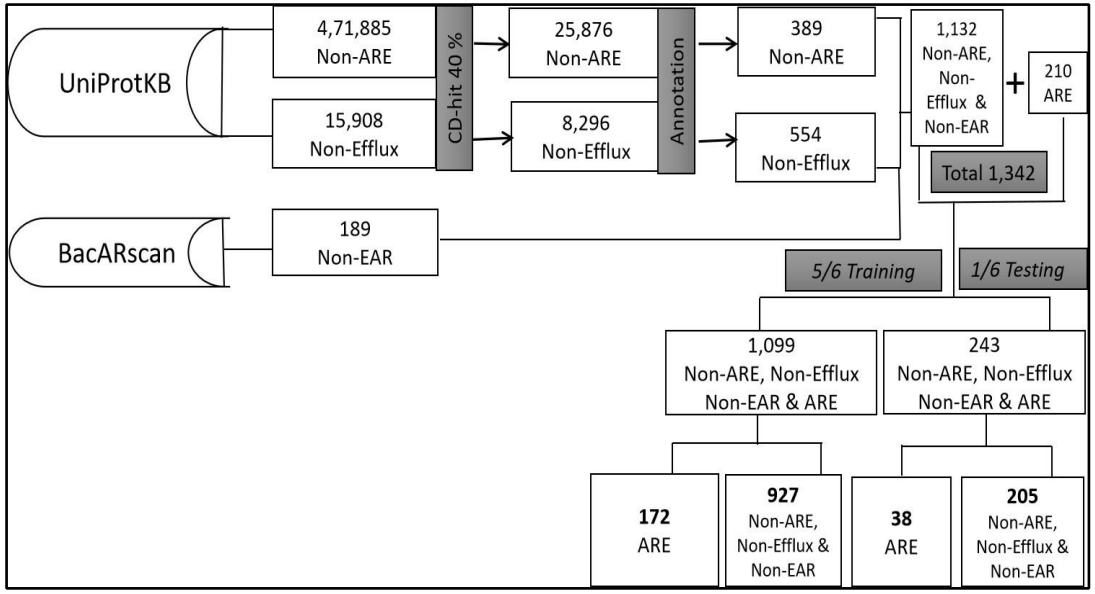
Pandey D. et. al. Scientific Reports; 2020
Pandey D. et. al. Nature Protocol Exchange; 2021

Workflow

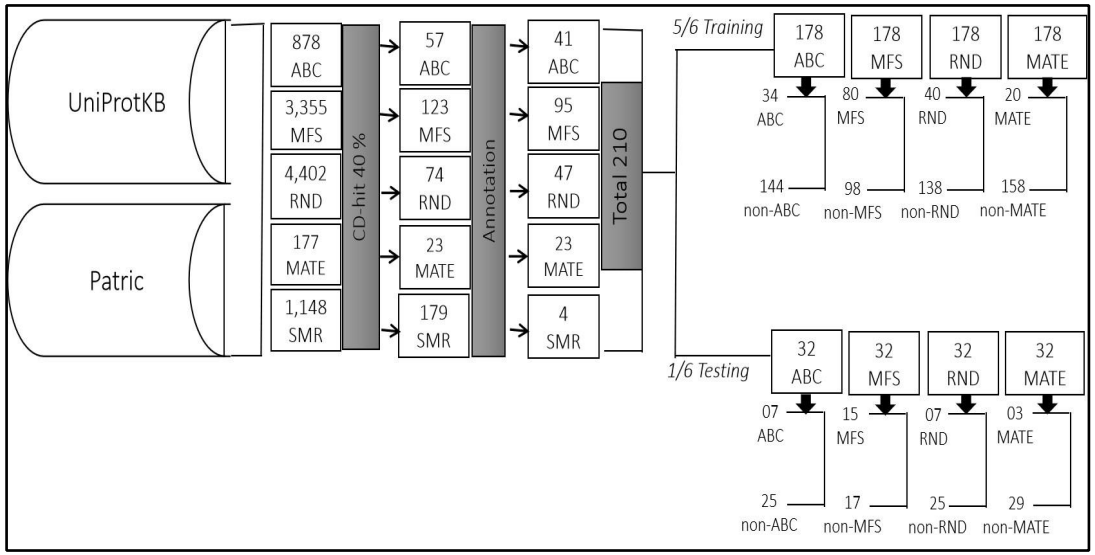
BacEffluxPred



Data Sources and Compilation



Tier-I dataset compilation: Numerical values indicate the number of proteins. ARE: antibiotic resistance efflux proteins, non-ARE: non-antibiotic resistance efflux proteins, non-efflux: non-efflux prokaryotic proteins, and non-EAR: non-efflux antibiotic resistance proteins.



Tier-II dataset compilation: Numerical values indicate the number of proteins. ABC, MFS, RND, MATE and SMR are efflux protein families.

Pandey D. et. al. Scientific Reports; 2020
Pandey D. et. al. Nature Protocol Exchange; 2021

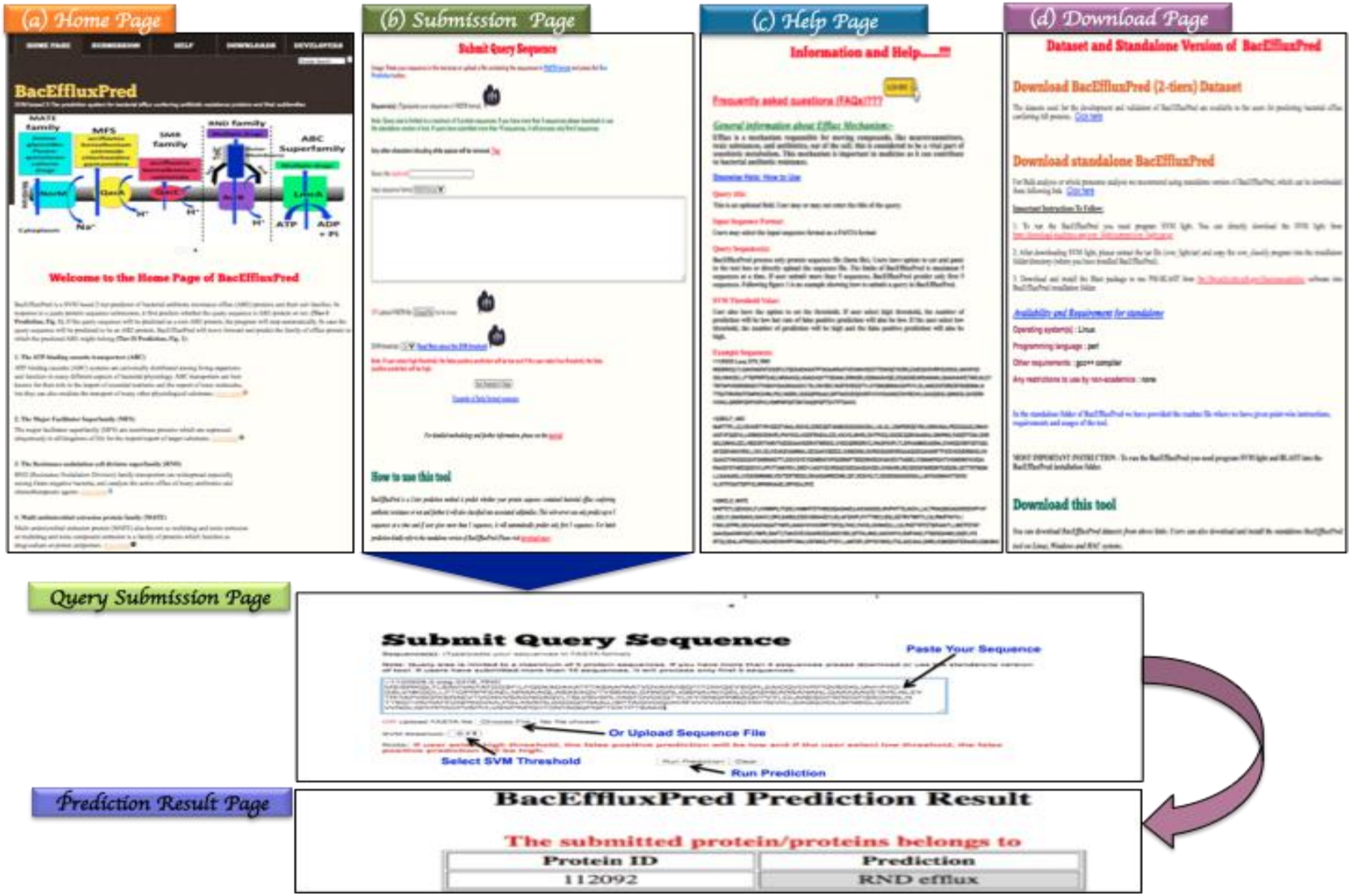
Performance of SVM models at training and independent testing dataset during LOOCV at tier–I and II

Threshold	Tier		Training Dataset					Independent Testing Dataset				
			AC (%)	SEN (%)	SPE (%)	MCC	AUC	AC (%)	SEN (%)	SPE (%)	MCC	AUC
-0.4	Tier-I		85.81	80.23	86.84	0.57	0.87	94.24	86.84	95.61	0.79	0.95
-0.4	Tier-II	ABC	92.13	88.24	93.06	0.77	0.96	93.75	100.00	92.00	0.85	0.96
-0.3		MFS	85.39	87.50	83.67	0.71	0.92	93.75	93.33	94.12	0.87	0.97
-0.4		RND	91.01	90.00	91.30	0.76	0.94	93.75	100.00	92.00	0.85	1.00
0.3		MATE	99.44	95.00	100.00	0.97	0.99	100.00	100.00	100.00	1.00	1.00

The overall performance of SVM models during LOOCV at tier-I and tier-II. AC, SEN, SPE, MCC and AUC represent accuracy, sensitivity, specificity, Matthew's correlation coefficient (MCC) and area under ROC curve (AUC)

The highly successful predictor will have MCC value near to 1, while opposite and random predictions have MCC value -1 and 0 respectively

Tool Page

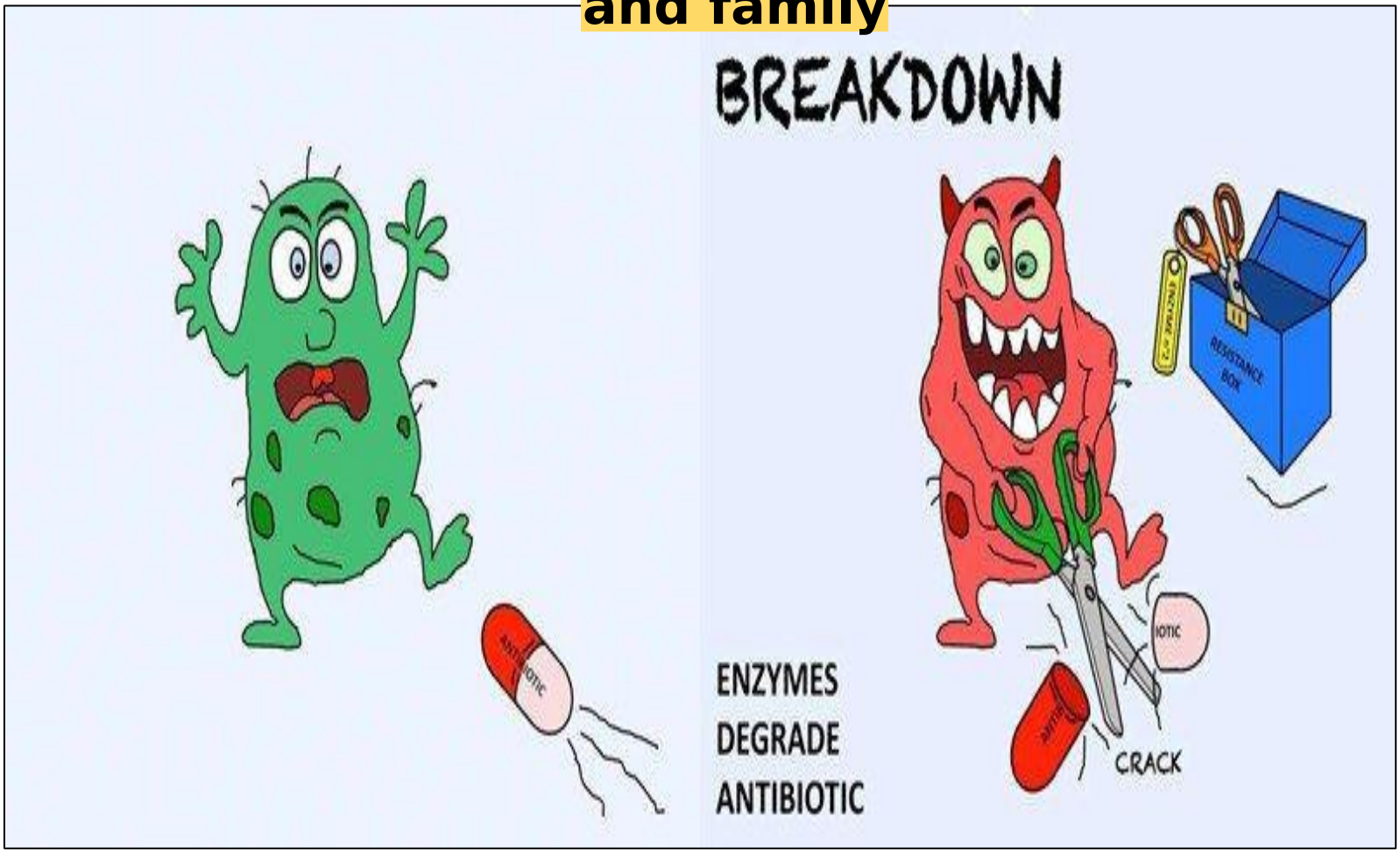


Pandey D. et. al. Scientific Reports; 2020

Pandey D. et. al. Nature Protocol Exchange; 2021

Web Link: http://proteininformatics.org/mkumar/baceffluxpred

To design an online tool for the prediction and classification of β-Lactamase in class, subclass, and family



Pandey D. et al. Frontiers in Microbiology, 2022

Motivation

β-lactams are the **most commonly prescribed drug for treatment of Gram-negative bacterial infection**. Despite **70 years of clinical use, β-lactam antibiotics** still remain at the **forefront** of antimicrobial chemotherapy.
The resistance against β-lactam antibiotics is due to development of a **highly diverse group of enzymes**, **collectively called β-lactamases** (BLs), **that hydrolyze the amide bond of a β-Lactam ring to make it ineffective**.

Over the years, several classification systems have been developed to classify BLs. However, the most popular schemes are:

(i) Ambler's classification scheme, which was based on the amino acid sequence similarity

(ii) Bush, Jacoby, and Medeiros classification scheme, which was based on substrate and inhibitor profiles

Ambler's classification scheme categorized BLs into **four classes: A, B, C, and D**. Class A, C, and D are also known as **serine BLs because they have an active-site serine to catalyze the hydrolysis**.

Class B BLs are known as Metallo β-Lactamases (MBLs) since they use zinc ions (Zn2+) for their activity.

MBLs are distinct from the serine BL in sequence, structure fold, and catalytic mechanism and they are further divided into three **subclasses, B1, B2, and B3**, based on **their active site geometry and overall homology**.

Motivation

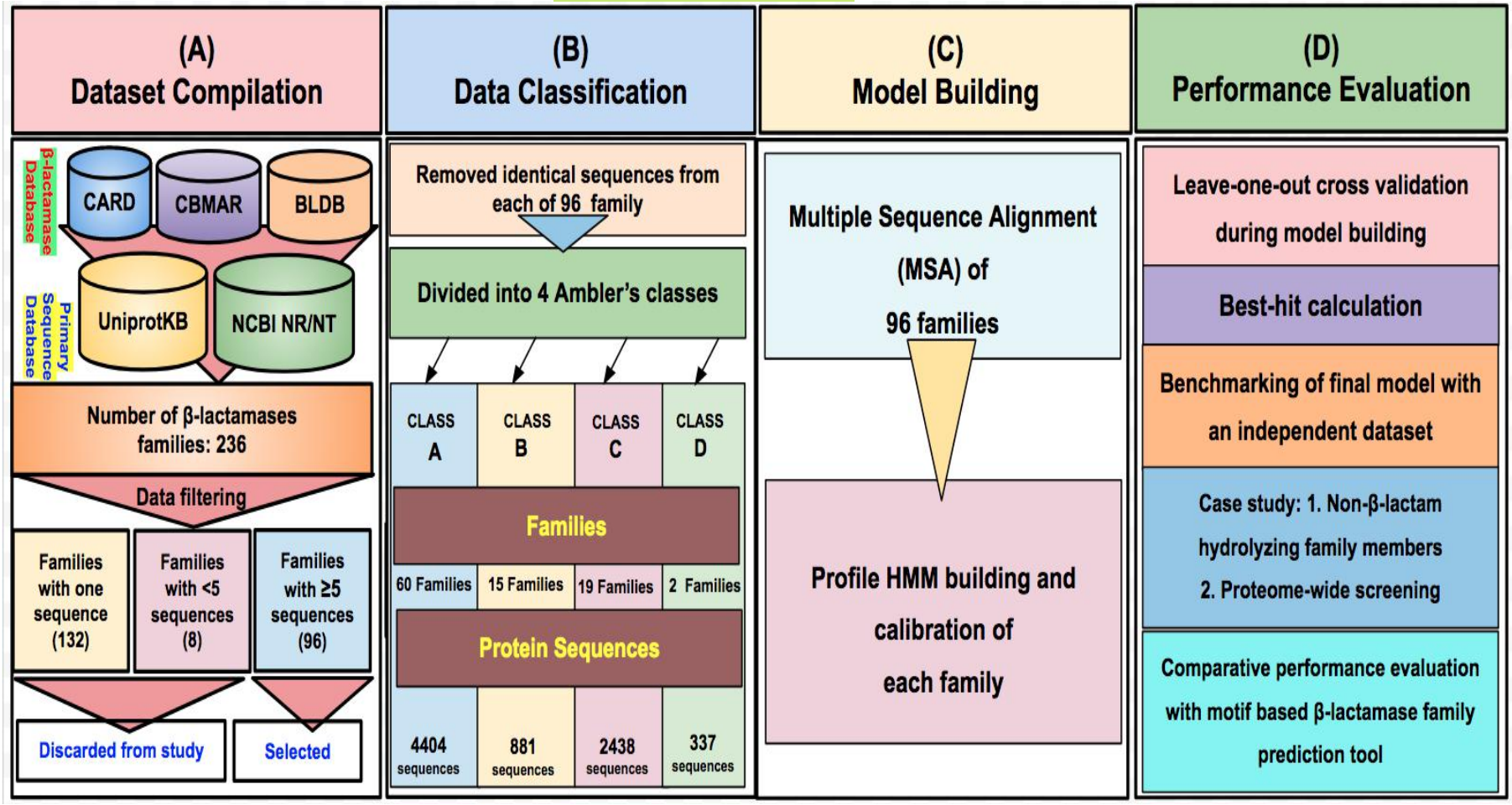
- Several screening tests have been developed to identify the family of BLs at both gene and whole genome levels. However, **these methods are resource and time-consuming**.
- An alternative approach for rapid annotation of BLs family is to use **computational methods**, which can quickly identify BLs genes/proteins and classify them into the family.
- The most popular computational approach is using BLAST search against either general-purpose **molecular biology databases such as NCBI NR/NT** or **UniProtKB/Swiss Prot or BL-specific databases such as BLDB, BLAD, LacED, ARDB, CARD, and our laboratory has also developed a database of β–lactamases named CBMAR**.

- **However, there are a few limitations of LactFP. The most critical limitation of LactFP was that it was developed using a dataset compiled in 2014.** Over time information about **new family members and many mutations in different families has been accumulated in the databases**. Hence LactFP might not be capable of predicting all BL families correctly. This indicates that a tool capable of predicting more BL families is the need of the hour.

- However, most prediction methods except LactFP were restricted only to the prediction upto class level (e.g. *β*Lact-Pred, CNN-BLPred, PredLactamase, or subclass (e.g. BlaPred). **LactFP predicts the class, sub-class, and family of a BL protein on the basis of presence of a family-specific motif called fingerprint in the primary amino acid sequence.**

β-LacFamPred: a machine learning based classifier that can annotate BLs up to the family level. β-LacFamPred can be used on both genomic and proteomic data.

Workflow



workflow depicting the methodology used for developing β-LacFamPred

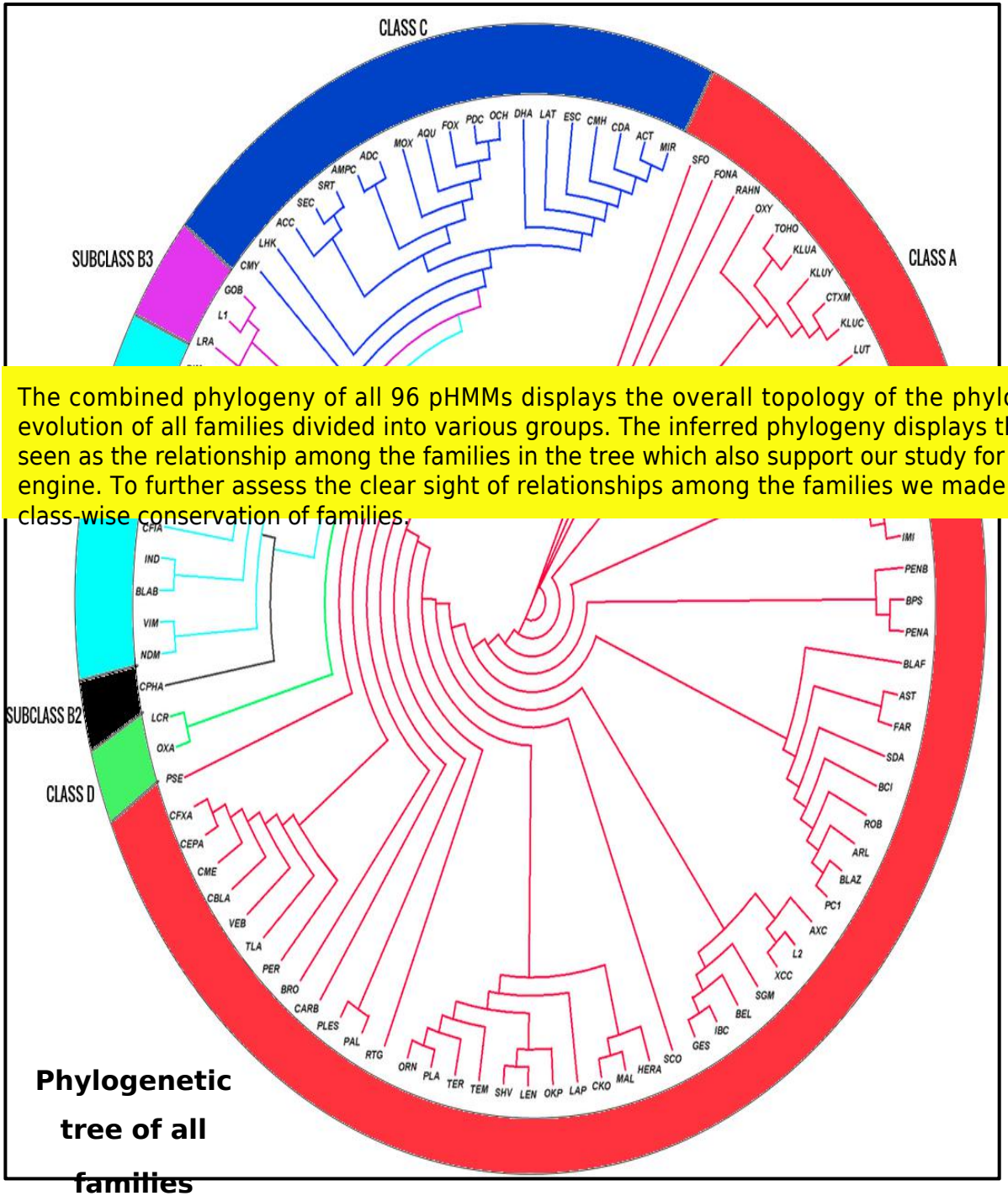
Training Dataset

The family-wise sequences of BLs were obtained from **BLDB & CBMAR databases**. The number of protein sequences in **each family was also augmented from CARD, UniProtKB, and NCBI NR databases**. Sequences of each family were manually curated **using literature and UniProtKB annotations**. **We also removed the fragmented sequences from each family**.

Class	Sub-class	Total Families	Families with one sequence	Families with <5 sequences	Families with ≥5 sequences
A	--	64# + 13* = 77	17#	0	47# + 13* = 60
B	B1	20# + 35* = 55	11# + 31* = 42	3*	9# + 1* = 10
	B2	3# + 3* = 6	2# + 2* = 4	1*	1#
	B3	13# + 42* = 55	9# +38* = 47	1# + 3* = 4	3# +1* = 4
C	--	14# + 9* = 23	4#	0	10# + 9* = 19
D	--	2# +18* =20	18*	0	2#

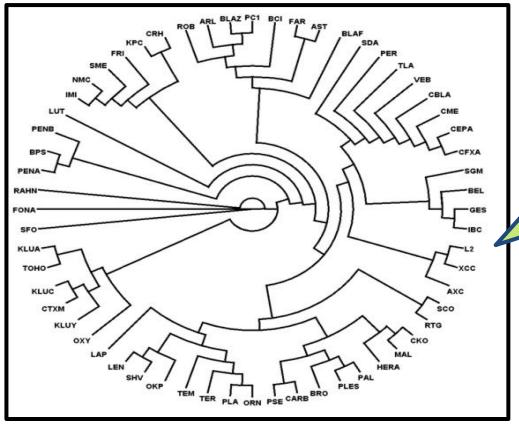
Statistics of BL families retrieved from CBMAR and BLDB databases.

BL families with **less than five sequences or single sequences were also removed from further studies**. If multiple copies of identical sequences were present in a family, then all except one sequence were removed.

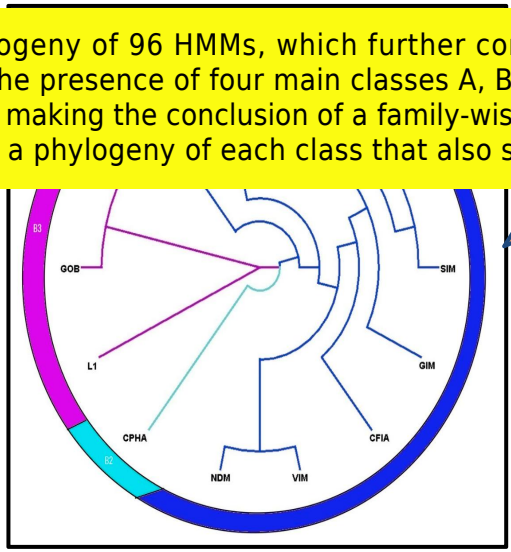


Phylogenetic tree of all families

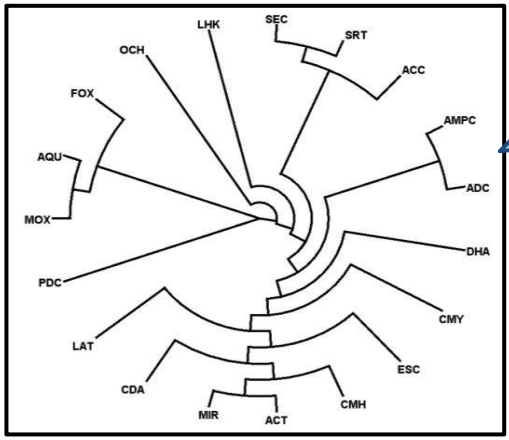
The combined phylogeny of all 96 pHMMs displays the overall topology of the phylogeny of 96 HMMs, which further confirmed the evolution of all families divided into various groups. The inferred phylogeny displays the presence of four main classes A, B, C & D also seen as the relationship among the families in the tree which also support our study for making the conclusion of a family-wise prediction engine. To further assess the clear sight of relationships among the families we made a phylogeny of each class that also supports the class-wise conservation of families.



CLASS A



CLASS B



CLASS C

Benchmarking Independent Dataset

- We used the reference *bla* gene sequences obtained from Lee et al. **(Lee et al., 2015 *Antimicrobial Agents and Chemotherapy)*** for benchmarking.
- These BL sequences were used to develop **molecular probes for PCR-based methods** to detect ***bla* genes in various pathogenic isolates**.
- The total number of *bla* gene sequences were 1342, belonging to all four Ambler's classes, A-D, and 29 families of BLs.

Methodology

Construction of the β-LacFamPred HMMs

- Sequences of each BL family were multiply aligned using the Muscle 3.8 program at default parameters.
- Using the hmmbuild function of the HMMER tool (version 3.1), we build HMM models of each family of BLs.

Comparative Evaluation

- We compared the performance of β-LacFamPred with well-known ARG annotation methods: AMRFinderPlus, RGI-CARD, ResFinder, and Meta-MARC.
- We have also included LactFP as it assigns the family of a BL sequence based on the presence of a

Functional Annotation

- All 96 BL HMMs were annotated using (a) ARG databases, namely **DeepARG - ARGminer, CARD, ARDB, (b) UniProtKB, and (c) published research papers**.
- The annotation details mentioned with each HMM are **resistance mechanisms, class, and name of antibiotic against which the family confers the resistance, family, class, subclass, and phenotypic information as per Jacoby and Bush classification scheme.**
- Each HMM was also tagged with the information of their action in terms of **their spectrum, namely broad spectrum, extended spectrum, and narrow spectrum.**

Contd..

Cross-validation and Performance Metrics

- To test the efficiency of each HMM in discriminating between the family and non-family members, we used the **leave-one-out approach of cross-validation (LOOCV)**. The performance of methods was assessed using the standard evaluation metrics namely, **precision, recall, accuracy, and F1 score**.

Case Study #1

Performance evaluation on homologous dataset

- To further assess the capability of β-LacFamPred for identifying BL class, subclass, and families, we performed an additional independent evaluation using a Penicillin-Binding Proteins (PBPs) dataset. PBPs are membrane-associated proteins involved in the biosynthesis of peptidoglycan components of bacterial cell walls. PBP and BLs belong to the superfamily of serine penicillin-recognizing enzymes and have similar conserved protein folds.
- PBP and BLs are homologous proteins, but PBP does not provide antibiotic resistance against BLs. Also, BLs are considered to have evolved from penicillin-binding proteins. PBPs were not part of the dataset on which β-LacFamPred prediction models were developed.

Out of 60 PBP sequences, only four were wrongly predicted as BLs.

Case Study #2

To confirm the discriminatory capability of β-lactamase, and non-β-lactamase, we created a second independent dataset consisting of glyoxalase II, which belongs to the metallo-beta-lactamase (MBL) superfamily of proteins. The sequences of the glyoxalase II were retrieved from the UniProtKB database.

We found a total of 57 full-length sequences of glyoxalase II. At e-value 1e-15 none of the glyoxalase II sequences were predicted as BL. When e-value was increased to 1e-10, 1e-6 and 0.1 the number gradually increased to 17, 43 and 43 respectively. The result was consistent with previous work that had shown the requirement of more stringent e-value cutoff to reduce the number of false positive predictions (Gibson et al., 2015; McArthur et al., 2013; Zankari et al., 2012).

Performance Comparison with Existing Methods

Method	Type of data	TP	FP	TN	FN	Precision (%)	Recall (%)	F-measure (%)	Accuracy
β-LacFamPred	**Protein sequences**	**1320**	**22**	**37554**	**22**	**98.36 %**	**98.36 %**	**98.36 %**	**0.99**
RGI-CARD		1115	227	37349	227	83.08%	83.08%	83.08%	0.98
AMRFinderPlus		1026	316	37260	316	76.45%	76.45%	76.45%	0.99
LactFP		742	600	36976	600	55.29%	55.29%	55.29%	0.96
β-LacFamPred	**Gene sequences**	**1337**	**5**	**37571**	**5**	**99.62 %**	**99.62 %**	**99.62 %**	**0.99**
Meta-MARC		1199	143	37433	143	89.34%	89.34%	89.34%	0.99
ResFinder		1242	100	37476	100	92.54%	92.54%	92.54%	0.99

Advantages and Limitations of Present and Previously Developed BL Family Prediction Method

Feature	LactFP	β-LacFamPred
Training data source	UniProtKB/TrEMBL	CBMAR, BLDB, CARD, UniProtKB, NCBI NR/NT
Total dataset	**605**	**8060**
Less than 5 sequence family used	Yes	No
One sequence family used	No	No
Similarity tool and threshold used	Blast (1e-4)	Blast (1e-6)
Total families	**71**	**96**
Benchmark data source	**None**	**Lee et al. (2015)**
Data redundancy threshold	Not mentioned	CD HIT (100%)
Tool used to develop prediction Model	Meme/Mast	HMM
Cross-validation method	**No**	**Leave-one-out cross validation (LOOCV)**
Web Server	Yes	Yes
Input data	**Only Protein sequences**	**Protein/Gene sequences**

Proteome–wide screening of β–Lactamases

Recently (Y. Wang et al., 2021) developed a method deep learning-based method, DeepBL, for predicting and classifying BLs on the basis of their protein sequences.

To characterize the complete repertoire of BLs, they annotated all reviewed bacterial protein sequences (334542 in total) from the UniProtKB database.

Ambler's Class	Number of proteins predicted as BL by DeepBL/ Annotated as BL by UniProt	Number of proteins predicted as BL by β-LacFamPred/ Annotated as BL by UniProt	Number of class B predicted as BL and their sub-class prediction	Number of families in which predicted BLs were distributed as per β-LacFamPred
A	2876/80	86/77	-	26
B	665/91	246/145	21 (B1)	5
			2 (B2)	1
			223 (B3)	3
C	335/13	67/15	-	10
D	231/15	29/15	-	2
Total	**4107/199**	**428/252**	**246**	**47**

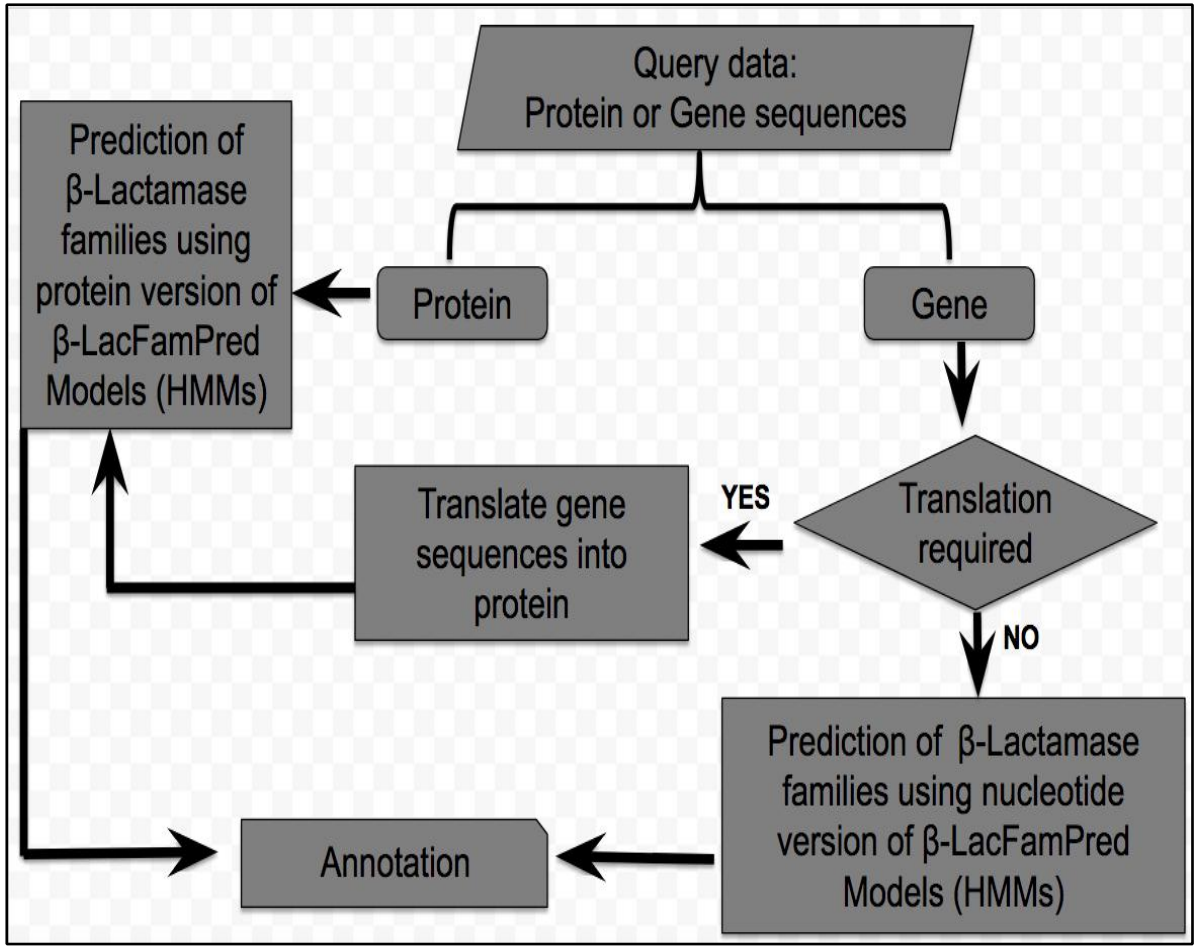
Number of proteins predicted as BL by DeepBL and β-LacFamPred and annotation statistics of UniProt therein

Comparative prediction outputs of DeepBL, UniProtKB and β–LacFamPred

S. No.	ID	Prediction Tools					
		DeepBL	UniProtKB		β-LacFamPred		
1.	Q9EZQ7	Class A	Class-A Beta Lactamase	Beta-lactamase AST-1	Class A	-	AST
2.	Q9S424	Class A	Class-A Beta-Lactamase	Beta-lactamase SHV-13	Class A	-	SHV
3.	P28585	Class A	Class-A Beta-	Beta-lactamase	Class A	-	CTXM
6.	P26918	Class B	Class-B Beta-Lactamase	Metallo-Beta-Lactamase type 2 cphA	Class B	Sub-class B2	CPHA
7.	A0A096ZEC9	Class A	Class-B Beta-Lactamase	Metallo-Beta-Lactamase type 2 cphA	Class B	Sub-class B2	CPHA
8.	O05465	Class C	Class-C Beta-Lactamase	Beta-lactamase ampc	Class C	-	AmpC
9.	B3U538	Class D	Class-D Beta-Lactamase	Beta-lactamase OXA-133	Class D	-	OXA
10.	Q00983	Class D	Class-D Beta-Lactamase	Beta-lactamase LCR-1	Class D	-	LCR
11.	A6V707	Not Beta-Lactamase	Class-B Beta-Lactamase	Metallo-Beta-Lactamase	Class B	Sub-class B3	L1
12.	O31760	Not Beta-Lactamase	Class-B Beta-Lactamase	Metallo-Beta-Lactamase	Class B	Sub-class B1	IMP
13.	A0A0H2UR93	Class A	Glucosyl transferase 3	Gtf3 glucosyl-transferase family	Non-Beta-Lactamase		
14.	B6I4P3	Not Beta-Lactamase	L-rhamnose mutarotase	Rhamnose mutarotase family	Non-Beta-Lactamase		
15.	V6F4W4	Not Beta-Lactamase	Magnetosome protein MamZ	Major facilitator superfamily	Non-Beta-Lactamase		

The results showed that the number of false positive predictions in β–LacFamPred was significantly lower than DeepBL and β–LacFamPred can be used to predict and annotate new BLs that are not known yet.

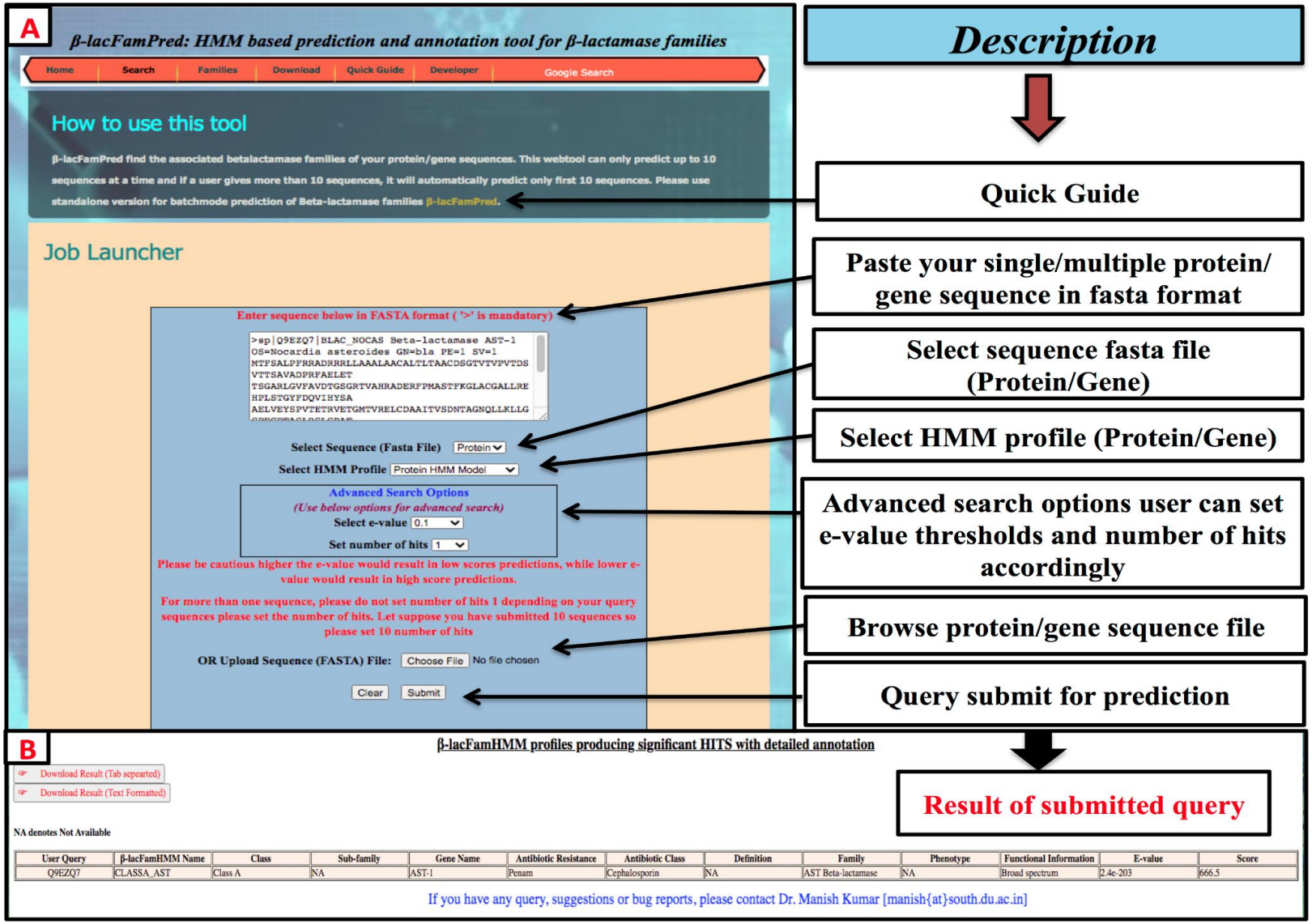
β-LacFamPred Web-server and Standalone Tool



The overall schema of the prediction methodology of the tool

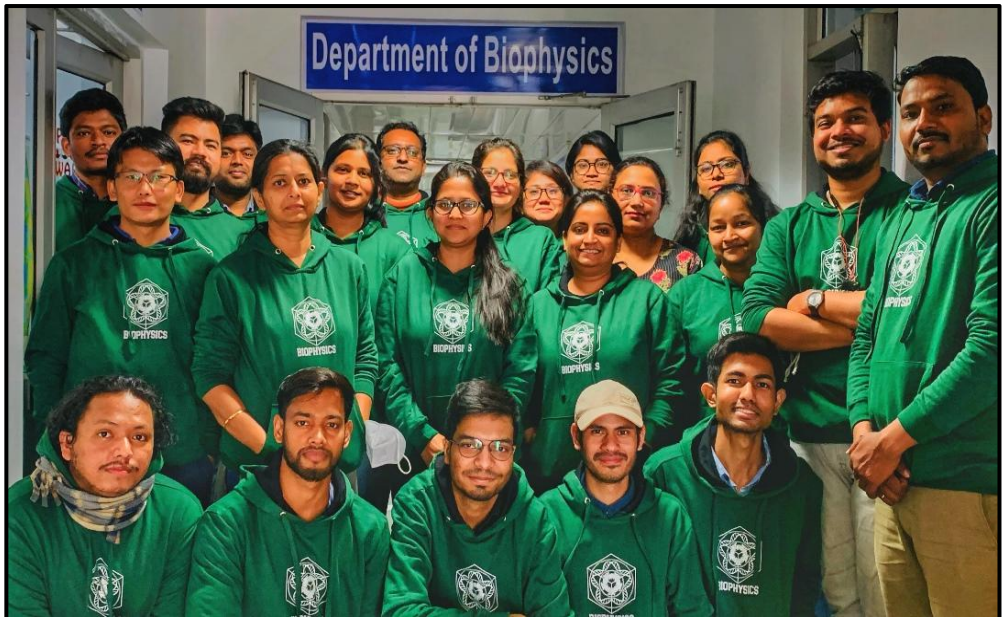
Screenshot of web-tool

Web Link:
http://proteininformatics.org/mkumar/blacfampred



A snapshot of the search and prediction page of the 'β-LacFamPred' web server

Acknowledgement



A CENTURY OF ACADEMIC EXCELLENCE
UNIVERSITY OF DELHI
YEARS
1922-2022
सत्यमेव जयते
Department of Science & Technology
Govt. of India
inspire
INDIAN COUNCIL OF MEDICAL RESEARCH
Serving the nation since 1911