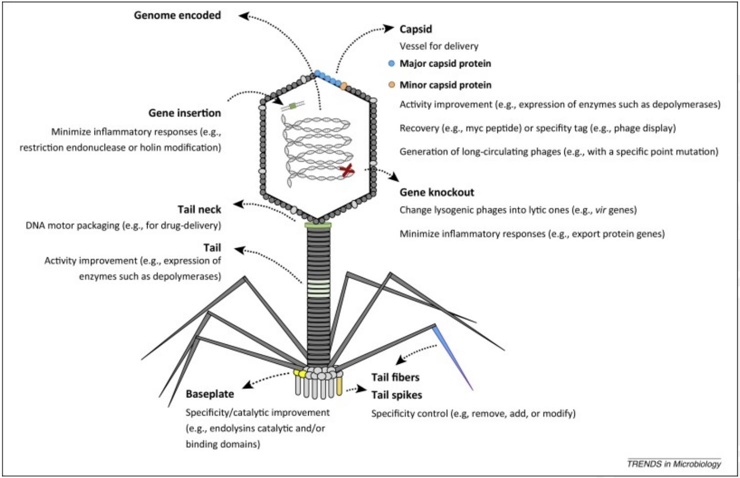
Artificial Neural networks for the prediction of phage protein function

Adrian Cantu

San Diego State University
Computational Science Research Center

May 21th 2019

BacterioPhage



Genome encoded

Capsid

Vessel for delivery

● **Major capsid protein**

● **Minor capsid protein**

Activity improvement (e.g., expression of enzymes such as depolymerases)

Recovery (e.g., myc peptide) or specifity tag (e.g., phage display)

Generation of long-circulating phages (e.g., with a specific point mutation)

Gene insertion

Minimize inflammatory responses (e.g., restriction endonuclease or holin modification)

Gene knockout

Change lysogenic phages into lytic ones (e.g., *vir* genes)

Minimize inflammatory responses (e.g., export protein genes)

Tail neck

DNA motor packaging (e.g., for drug-delivery)

Tail

Activity improvement (e.g., expression of enzymes such as depolymerases)

Baseplate

Specificity/catalytic improvement (e.g., endolysins catalytic and/or binding domains)

Tail fibers

Tail spikes

Specificity control (e.g, remove, add, or modify)

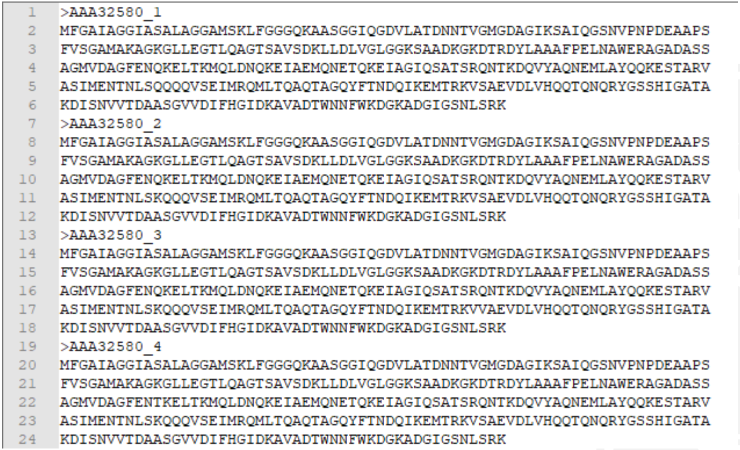
TRENDS in Microbiology

Databases

Class	Raw sequences	After manual curation	After 90% dereplication
Major capsid	112,987	105,653	13,172
Minor capsid	2,901	1,903	656
Baseplate	75,599	19,293	2,090
Major tail	66,513	35,030	3,249
Minor tail	94,628	80,467	3,886
Portal	210,064	189,143	18,622
Tail fiber	29,132	18,514	3,191
Tail shaft	37,885	35,570	4,933
Collar	4,224	3,709	1,262
Head-Tail joining	60,270	58,658	6,713
Other	733,006	-	162,709

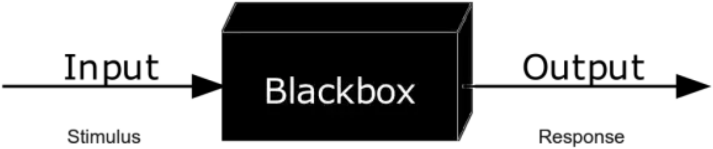
Table: The classes database by the numbers

Protein Sequences

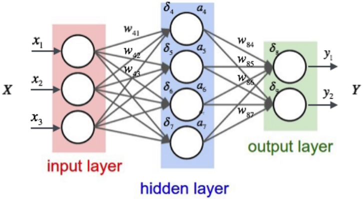


```
>AAA32580_1
MFGAIAGGIASALAGGAMSKLFGGGQKAASGGIQGDVLATDNNTVGMGDAGIKSAIQGSNVPNPDEAAPS
FVSGAMAKAGKGLLEGTLQAGTSAVSDKLLDLVGLGGKSAADKGKDTRDYLAAAFPELNAWERAGADASS
AGMVDAGFENQKELTKMQLDNQKEIAEMQNETQKEIAGIQSATSRQNTKDQVYAQNEMLAYQQKESTARV
ASIMENTNLSQQQQVSEIMRQMLTQAQTAGQYFTNDQIKEMTRKVSAEVDLVHQQTQNQRYGSSHIGATA
KDISNVVTDAASGVVDIFHGIDKAVADTWNNFWKDGKADGIGSNLSRK
>AAA32580_2
MFGAIAGGIASALAGGAMSKLFGGGQKAASGGIQGDVLATDNNTVGMGDAGIKSAIQGSNVPNPDEAAPS
FVSGAMAKAGKGLLEGTLQAGTSAVSDKLLDLVGLGGKSAADKGKDTRDYLAAAFPELNAWERAGADASS
AGMVDAGFENQKELTKMQLDNQKEIAEMQNETQKEIAGIQSATSRQNTKDQVYAQNEMLAYQQKESTARV
ASIMENTNLSKQQQVSEIMRQMLTQAQTAGQYFTNDQIKEMTRKVSAEVDLVHQQTQNQRYGSSHIGATA
KDISNVVTDAASGVVDIFHGIDKAVADTWNNFWKDGKADGIGSNLSRK
>AAA32580_3
MFGAIAGGIASALAGGAMSKLFGGGQKAASGGIQGDVLATDNNTVGMGDAGIKSAIQGSNVPNPDEAAPS
FVSGAMAKAGKGLLEGTLQAGTSAVSDKLLDLVGLGGKSAADKGKDTRDYLAAAFPELNAWERAGADASS
AGMVDAGFENQKELTKMQLDNQKEIAEMQNETQKEIAGIQSATSRQNTKDQVYAQNEMLAYQQKESTARV
ASIMENTNLSKQQQVSEIMRQMLTQAQTAGQYFTNDQIKEMTRKVVAEVDLVHQQTQNQRYGSSHIGATA
KDISNVVTDAASGVVDIFHGIDKAVADTWNNFWKDGKADGIGSNLSRK
>AAA32580_4
MFGAIAGGIASALAGGAMSKLFGGGQKAASGGIQGDVLATDNNTVGMGDAGIKSAIQGSNVPNPDEAAPS
FVSGAMAKAGKGLLEGTLQAGTSAVSDKLLDLVGLGGKSAADKGKDTRDYLAAAFPELNAWERAGADASS
AGMVDAGFENTKELTKMQLDNQKEIAEMQNETQKEIAGIQSATSRQNTKDQVYAQNEMLAYQQKESTARV
ASIMENTNLSKQQQVSEIMRQMLTQAQTAGQYFTNDQIKEMTRKVSAEVDLVHQQTQNQRYGSSHIGATA
KDISNVVTDAASGVVDIFHGIDKAVADTWNNFWKDGKADGIGSNLSRK
```

F:Sequence –> Function



Artificial Neural Networks



ANN have been shown to be universal approximators of continuous functions in $\mathbb{R}^n$

$$d = \left(\int_0^{2\pi} |f_1(t) - f_2(t)|^p dt \right)^{\frac{1}{p}}$$

where $1 < p < \infty$

$$\begin{pmatrix} Z_1 \\ Z_2 \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ Z_{410} \end{pmatrix} = X$$

$$\begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \\ Y_5 \\ Y_6 \\ Y_7 \\ Y_8 \\ Y_9 \\ Y_{10} \end{pmatrix} = Y$$

where $\sum_{n=1}^{10} Y_n = 1$

The 'black box' function

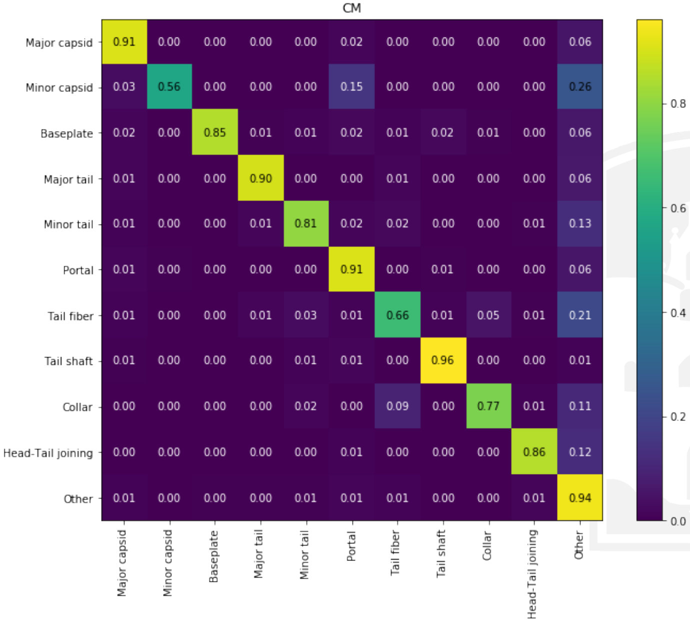
$$F(X) = \underbrace{[10 \ast 200]}_{W_3} \left(\underbrace{[200 \ast 200]}_{W_2} \left(\underbrace{[200 \ast 407]}_{W_1} \underbrace{[407 \ast 1]}_{X} + \underbrace{[200 \ast 1]}_{\delta_1} \right) + \underbrace{[200 \ast 10]}_{\delta_2} \right) + \underbrace{[10 \ast 1]}_{\delta_1}$$

289,866 Trainable parameters

Accuracy

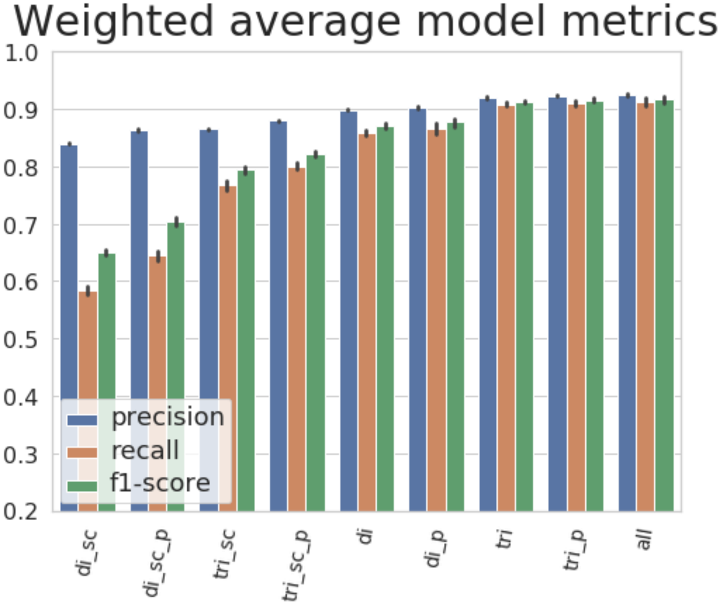
	Precision	*Recall*	*f1 − score*	*Support*
Major capsid	0.88	0.92	0.90	1232
Minor capsid	0.27	0.57	0.36	51
Baseplate	0.54	0.87	0.67	180
Major tail	0.82	0.88	0.85	289
Minor Tail	0.65	0.77	0.70	345
Portal	0.87	0.90	0.88	1640
Tail Fiber	0.54	0.67	0.60	272
Tail shaft	0.91	0.94	0.93	444
Collar	0.75	0.80	0.77	129
Head − Tail Joining	0.74	0.84	0.79	647
Other	0.97	0.93	0.95	15254
weighted avg	0.82	0.79	0.79	675

Results Confusion matrix

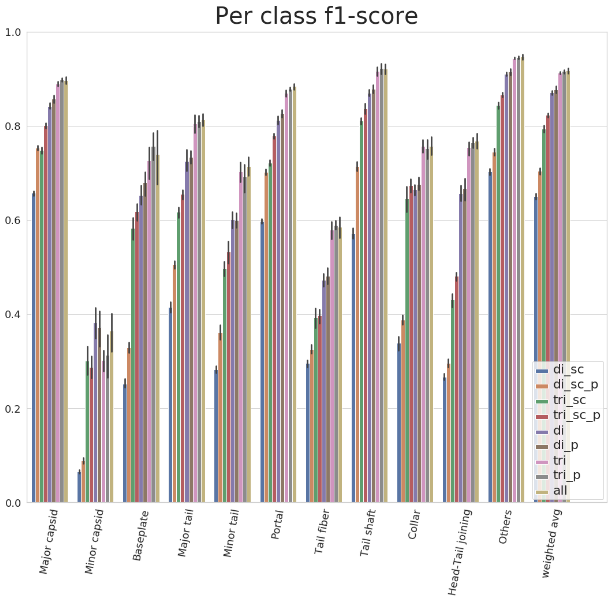


	Major capsid	Minor capsid	Baseplate	Major tail	Minor tail	Portal	Tail fiber	Tail shaft	Collar	Head-Tail joining	Other
Major capsid	0.91	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.06
Minor capsid	0.03	0.56	0.00	0.00	0.00	0.15	0.00	0.00	0.00	0.00	0.26
Baseplate	0.02	0.00	0.85	0.01	0.01	0.02	0.01	0.02	0.01	0.00	0.06
Major tail	0.01	0.00	0.00	0.90	0.00	0.00	0.01	0.00	0.00	0.00	0.06
Minor tail	0.01	0.00	0.00	0.01	0.81	0.02	0.02	0.00	0.00	0.01	0.13
Portal	0.01	0.00	0.00	0.00	0.00	0.91	0.00	0.01	0.00	0.00	0.06
Tail fiber	0.01	0.00	0.00	0.01	0.03	0.01	0.66	0.01	0.05	0.01	0.21
Tail shaft	0.01	0.00	0.00	0.00	0.01	0.01	0.00	0.96	0.00	0.00	0.01
Collar	0.00	0.00	0.00	0.00	0.02	0.00	0.09	0.00	0.77	0.01	0.11
Head-Tail joining	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.86	0.12
Other	0.01	0.00	0.00	0.00	0.01	0.01	0.01	0.00	0.00	0.01	0.94

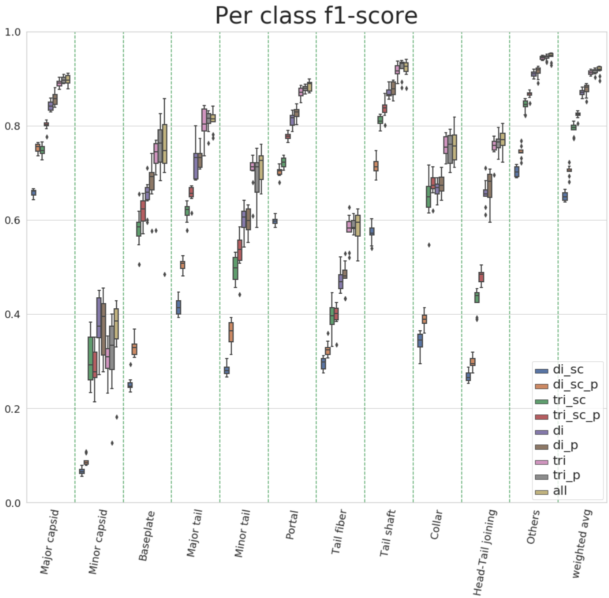
Weighted average



Per class f1-score



Per class f1-score



website

http://edwards.sdsu.edu/PhANNies/

Conclusions

- ANN is slow to train but fast to run.
- Robots will rule the world