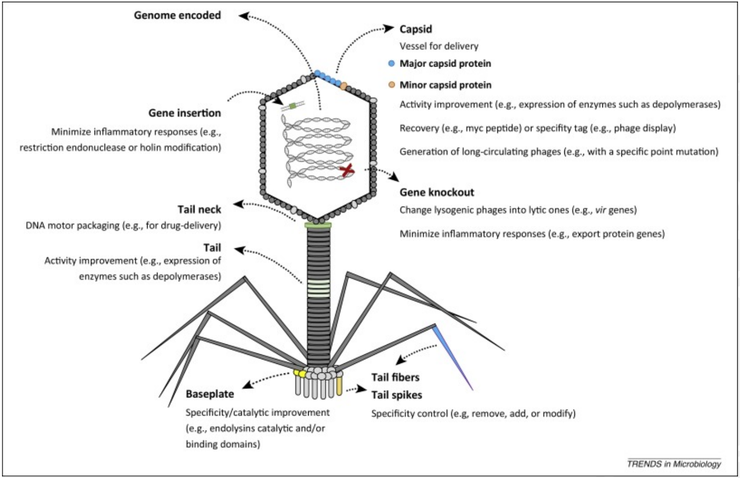
Artificial Neural networks for the prediction of phage protein function

Adrian Cantu

San Diego State University
Computational Science Research Center

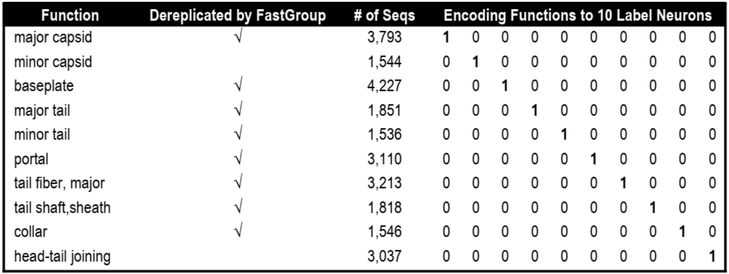
February 6th 2019

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)
Tail fibers
Tail spikes
Specificity control (e.g, remove, add, or modify)
Baseplate
Specificity/catalytic improvement (e.g., endolysins catalytic and/or binding domains)

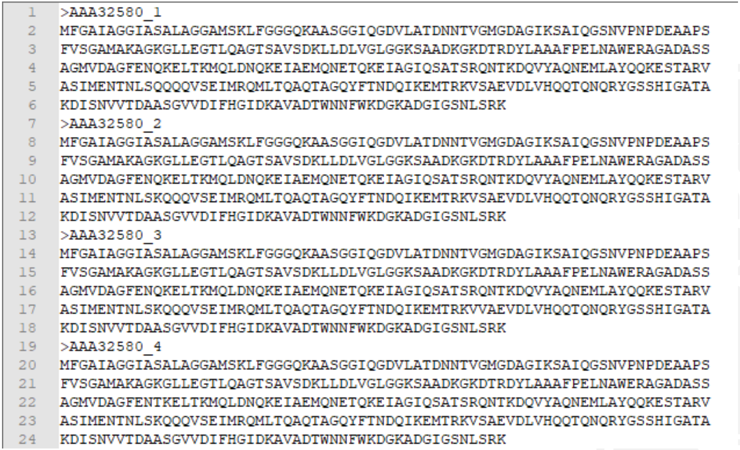
TRENDS in Microbiology

Databases



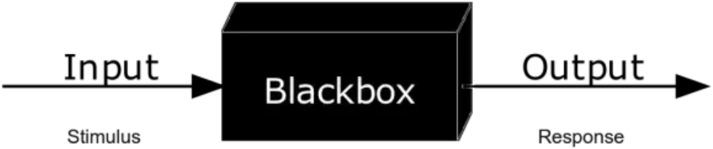
Function	Dereplicated by FastGroup	# of Seqs	Encoding Functions to 10 Label Neurons									
major capsid	√	3,793	1	0	0	0	0	0	0	0	0	0
minor capsid		1,544	0	1	0	0	0	0	0	0	0	0
baseplate	√	4,227	0	0	1	0	0	0	0	0	0	0
major tail	√	1,851	0	0	0	1	0	0	0	0	0	0
minor tail	√	1,536	0	0	0	0	1	0	0	0	0	0
portal	√	3,110	0	0	0	0	0	1	0	0	0	0
tail fiber, major	√	3,213	0	0	0	0	0	0	1	0	0	0
tail shaft,sheath	√	1,818	0	0	0	0	0	0	0	1	0	0
collar	√	1,546	0	0	0	0	0	0	0	0	1	0
head-tail joining		3,037	0	0	0	0	0	0	0	0	0	1

Protein Sequences

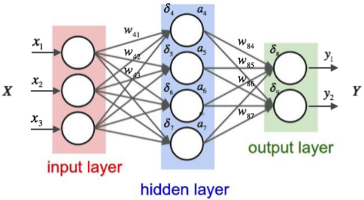


```
>AAA32580_1
MFGAIAGGIASALAGGAMSKLFGGGQKAASGGIQGDVLATDNNTVGMGDAGIKSAIQGSNVPNPDEAAPS
FVSGAMAKAGKGLLEGTLQAGTSAVSDKLLDLVGLGGKSAADKGKDTRDYLAAAFPELNAWERAGADASS
AGMVDAGFENQKELTKMQLDNQKEIAEMQNETQKEIAGIQSATSRQNTKDQVYAQNEMLAYQQKESTARV
ASIMENTNLSQQQQVSEIMRQMLTQAQTAGQYFTNDQIKEMTRKVSAEVDLVHQQTQNQRYGSSHIGATA
KDISNVVTDAASGVVDIFHGIDKAVADTWNNFWKDGKADGIGSNLSRK
>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 $1 < p < \infty$

Artificial Neural Networks

$$\begin{pmatrix} Z_1 \\ Z_2 \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ Z_{407} \end{pmatrix} = X$$

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

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

The 'black box' function

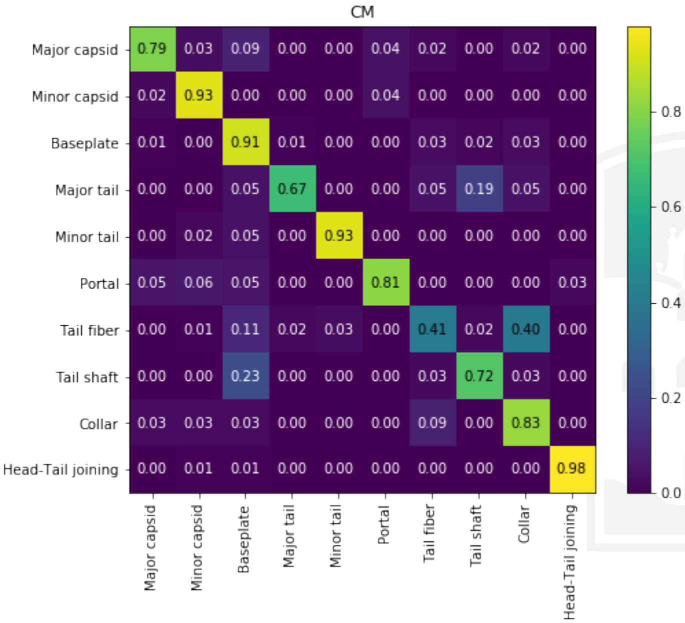
$$F(X) = \underbrace{[10 * 200]}_{W_3} \left(\underbrace{[200 * 200]}_{W_2} \left(\underbrace{[200 * 407]}_{W_1} \underbrace{[407 * 1]}_{X} + \underbrace{[200 * 1]}_{\delta_1} \right) + \underbrace{[200 * 10]}_{\delta_2} \right) + \underbrace{[10 * 1]}_{\delta_1}$$

289,866 Trainable parameters

Accuracy

	Precision	*Recall*	$f1 - score$	*Support*
Major capsid	0.91	0.79	0.85	95
Minor capsid	0.78	0.93	0.85	45
Baseplate	0.72	0.91	0.80	108
Major tail	0.91	0.67	0.77	43
Minor Tail	0.93	0.93	0.93	44
Portal	0.92	0.81	0.86	80
Tail Fiber	0.78	0.41	0.53	96
Tail shaft	0.70	0.72	0.71	39
Collar	0.39	0.83	0.53	53
$Head - Tail\ Joining$	0.98	0.98	0.98	90
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
Major capsid	0.79	0.03	0.09	0.00	0.00	0.04	0.02	0.00	0.02	0.00
Minor capsid	0.02	0.93	0.00	0.00	0.00	0.04	0.00	0.00	0.00	0.00
Baseplate	0.01	0.00	0.91	0.01	0.00	0.00	0.03	0.02	0.03	0.00
Major tail	0.00	0.00	0.05	0.67	0.00	0.00	0.05	0.19	0.05	0.00
Minor tail	0.00	0.02	0.05	0.00	0.93	0.00	0.00	0.00	0.00	0.00
Portal	0.05	0.06	0.05	0.00	0.00	0.81	0.00	0.00	0.00	0.03
Tail fiber	0.00	0.01	0.11	0.02	0.03	0.00	0.41	0.02	0.40	0.00
Tail shaft	0.00	0.00	0.23	0.00	0.00	0.00	0.03	0.72	0.03	0.00
Collar	0.03	0.03	0.03	0.00	0.00	0.00	0.09	0.00	0.83	0.00
Head-Tail joining	0.00	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.98

Conclusions

- ANN is slow to train but fast to run.
- Robots will rule the world
- ”Collar” proteins are not a real thing