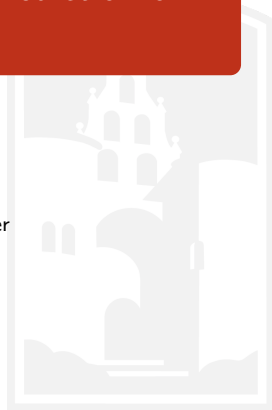


Artificial Neural networks for the prediction of phage protein function

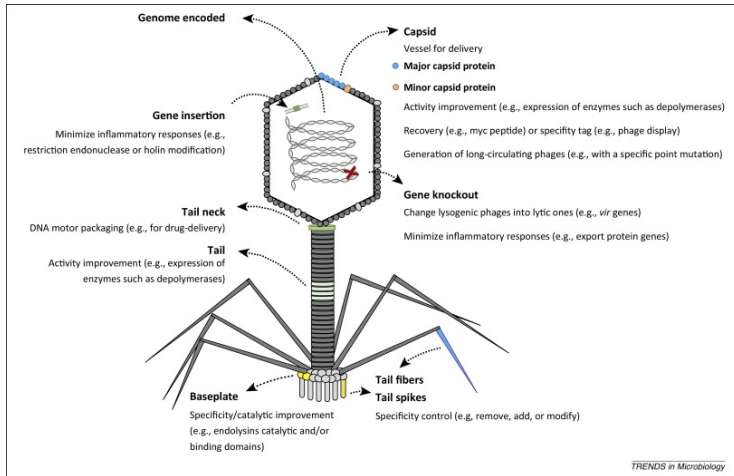
Adrian Cantu

San Diego State University
Computational Science Research Center

March 23th 2019



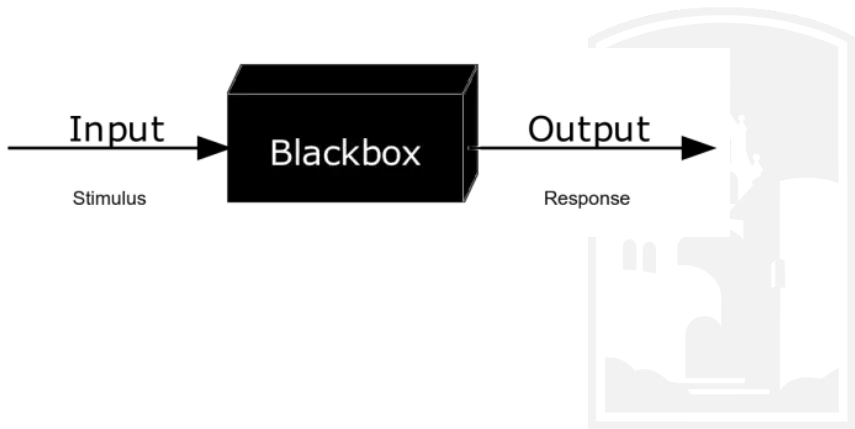
BacterioPhage



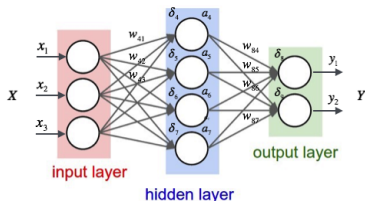
Function	Dereplicated by FastGroup	# of Seqs	Encoding Functions to 10 Label Neurons									
major capsid	✓	3,793	1	0	0	0	0	0	0	0	0	0
minor capsid		1,544	0	1	0	0	0	0	0	0	0	0
baseplate	✓	4,227	0	0	1	0	0	0	0	0	0	0
major tail	✓	1,851	0	0	0	1	0	0	0	0	0	0
minor tail	✓	1,536	0	0	0	0	1	0	0	0	0	0
portal	✓	3,110	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

```
1 >AAA32580_1
2 MFGAIAGGIASALAGGAMSKLFGGGQKAASGGIQQGDVLATDNNVTGMDAGIKSAIQGSNVPNPDEAAPS
3 FVSGAMAKAGKGLLEGLTQAGTSAVSDKLLDLVGLGGKSAADKGKDRDYLAFAPELNAWERAGADASS
4 AGMVDAGFENQKELTKMQLDNQKEIAEMQNETQKEIAGIQSATSQRNTKQDVYAQNEMLAYQQKESTARV
5 ASIMENTNLSQQQQVSEIMRQMLTQAQTAGQYFTNDQIKEMTRKVSAEVDLVHQQTQNQRYGSSSHIGATA
6 KDISNVVTDAAAGVVDIFHGDIDKAVADTWNNFWKDGKADGIGSNLSRK
7 >AAA32580_2
8 MFGAIAGGIASALAGGAMSKLFGGGQKAASGGIQQGDVLATDNNVTGMDAGIKSAIQGSNVPNPDEAAPS
9 FVSGAMAKAGKGLLEGLTQAGTSAVSDKLLDLVGLGGKSAADKGKDRDYLAFAPELNAWERAGADASS
10 AGMVDAGFENQKELTKMQLDNQKEIAEMQNETQKEIAGIQSATSQRNTKQDVYAQNEMLAYQQKESTARV
11 ASIMENTNLSKQQQVSEIMRQMLTQAQTAGQYFTNDQIKEMTRKVSAEVDLVHQQTQNQRYGSSSHIGATA
12 KDISNVVTDAAAGVVDIFHGDIDKAVADTWNNFWKDGKADGIGSNLSRK
13 >AAA32580_3
14 MFGAIAGGIASALAGGAMSKLFGGGQKAASGGIQQGDVLATDNNVTGMDAGIKSAIQGSNVPNPDEAAPS
15 FVSGAMAKAGKGLLEGLTQAGTSAVSDKLLDLVGLGGKSAADKGKDRDYLAFAPELNAWERAGADASS
16 AGMVDAGFENQKELTKMQLDNQKEIAEMQNETQKEIAGIQSATSQRNTKQDVYAQNEMLAYQQKESTARV
17 ASIMENTNLSKQQQVSEIMRQMLTQAQTAGQYFTNDQIKEMTRKVSAEVDLVHQQTQNQRYGSSSHIGATA
18 KDISNVVTDAAAGVVDIFHGDIDKAVADTWNNFWKDGKADGIGSNLSRK
19 >AAA32580_4
20 MFGAIAGGIASALAGGAMSKLFGGGQKAASGGIQQGDVLATDNNVTGMDAGIKSAIQGSNVPNPDEAAPS
21 FVSGAMAKAGKGLLEGLTQAGTSAVSDKLLDLVGLGGKSAADKGKDRDYLAFAPELNAWERAGADASS
22 AGMVDAGFENTKELTKMQLDNQKEIAEMQNETQKEIAGIQSATSQRNTKQDVYAQNEMLAYQQKESTARV
23 ASIMENTNLSKQQQVSEIMRQMLTQAQTAGQYFTNDQIKEMTRKVSAEVDLVHQQTQNQRYGSSSHIGATA
24 KDISNVVTDAAAGVVDIFHGDIDKAVADTWNNFWKDGKADGIGSNLSRK
```

F:Sequence \rightarrow 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 $1 < p < \infty$

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

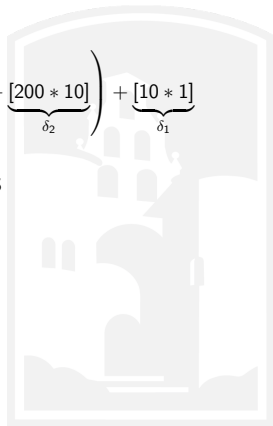
$$\begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \\ Y_5 \\ Y_6 \\ Y_7 \\ Y_8 \\ Y_9 \\ Y_{10} \end{pmatrix} = Y$$

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

The 'black box' function

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

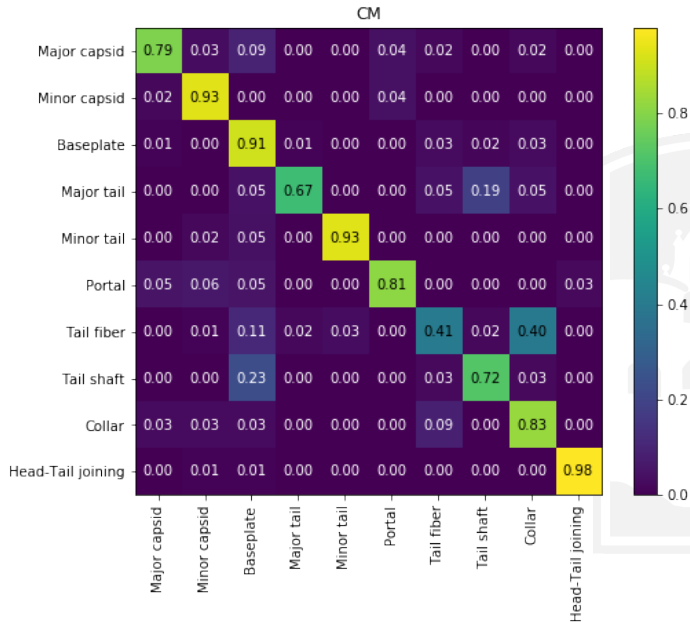
289,866 Trainable parameters

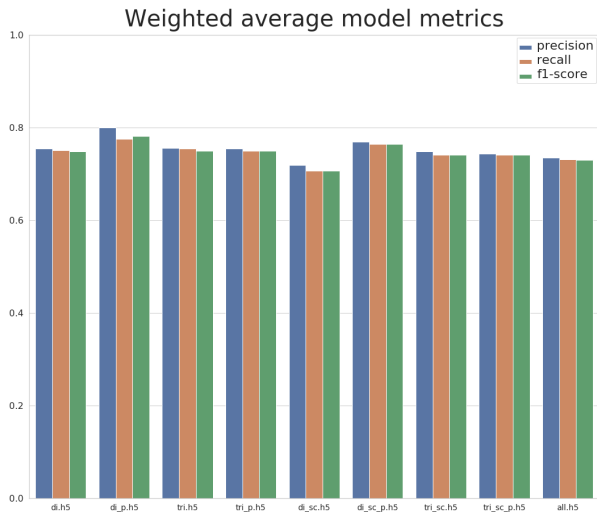


Accuracy

	<i>Precision</i>	<i>Recall</i>	<i>f1 – score</i>	<i>Support</i>
<i>Major capsid</i>	0.91	0.79	0.85	95
<i>Minor capsid</i>	0.78	0.93	0.85	45
<i>Baseplate</i>	0.72	0.91	0.80	108
<i>Major tail</i>	0.91	0.67	0.77	43
<i>Minor Tail</i>	0.93	0.93	0.93	44
<i>Portal</i>	0.92	0.81	0.86	80
<i>Tail Fiber</i>	0.78	0.41	0.53	96
<i>Tail shaft</i>	0.70	0.72	0.71	39
<i>Collar</i>	0.39	0.83	0.53	53
<i>Head – Tail Joining</i>	0.98	0.98	0.98	90
<i>weighted avg</i>	0.82	0.79	0.79	675

Results Confusion matrix





Per class f1-score





https://edwards.sdsu.edu/adrian_net



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

