BIOINFORMATICS

Vol. 00 no. 00
Pages 1–2

Supplementary Information for “LOGICOIL: Multi-state prediction of coiled-coil oligomeric state.”

Thomas. L. Vincent [1,2], Peter J. Green [3] and Derek N. Woolfson [1,4] *

[1]School of Chemistry, University of Bristol, Bristol, BS8 1TS.
[2]Bristol Centre for Complexity Science, University of Bristol, Bristol, BS8 1TR.
[3]School of Mathematics, University of Bristol, Bristol, BS8 1TW.
[4]School of Biochemistry, Medical Sciences Building, University of Bristol, Bristol, BS8 1TD.

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

Table S1. List of strongest pairwise interactions selected at positions $i \rightarrow i_{+3}$

a	b	c	d	e	f	g	a’
A	—	—	F	.	.	.	.
L	—	—	S	.	.	.	.
L	—	—	K	.	.	.	.
V	—	—	L	.	.	.	.
N	—	—	L	.	.	.	.
R	—	—	L	.	.	.	.
I	—	—	L	.	.	.	.
A	—	—	T	.	.	.	.
L	—	—	I	.	.	.	.
.	A	—	—	E	.	.	.
.	Q	—	—	N	.	.	.
.	R	—	—	Q	.	.	.
.	K	—	—	E	.	.	.
.	.	S	—	—	A	.	.
.	.	.	E	—	—	H	.
.	.	.	L	—	—	E	.
.	.	.	.	I	—	—	V
.	.	.	.	Y	—	—	A

Table S2. List of strongest pairwise interactions selected at positions $i \rightarrow i_{+4}$

a	b	c	d	e	f	g	a’
L	.	.	.	T			
Y	.	.	.	F			
N	.	.	.	R			
K	.	.	.	I			
K	.	.	.	Q			
L	.	.	.	E			
.	A	—	—	—	K	.	.
.	I	—	—	—	H	.	.
.	V	—	—	—	T	.	.
.	E	—	—	—	D	.	.
.	.	M	—	—	—	A	.
.	.	V	—	—	—	T	.
.	.	Q	—	—	—	E	.
.	.	E	—	—	—	H	.
.	.	H	—	—	—	E	.
.	.	D	—	—	—	R	.
.	.	E	—	—	—	I	.
.	.	.	I	—	—	—	I
.	.	.	I	—	—	—	L
.	.	.	I	—	—	—	S
.	.	.	L	—	—	—	V
.	.	.	L	—	—	—	N
.	.	.	L	—	—	—	R
.	.	.	Y	—	—	—	A
.	.	.	Y	—	—	—	T
.	.	.	V	—	—	—	I
.	.	.	K	—	—	—	I
.	.	.	L	—	—	—	K
.	.	.	N	—	—	—	I

*To whom correspondence should be addressed.

1 LOGICOIL AGAINST HOMOLOGY-BASED PREDICTIONS.

The results presented in the main manuscript show that LOGICOIL achieves a high discrimination rate between different coiled-coil architectures, but also equals or betters any other existing oligomeric state prediction algorithms. We also evaluate how homology-based predictions perform on mtuli-classification of coiled-coil oligomeric state. Through this, we hope to uncover whether there exists a redundancy cutoff for which homology-based predictions become efficient enough to be considered as more reliable than their counterpart *de novo* prediction algorithms when undertaking coiled-coil oligomeric state prediction. To quantify how well the 50% maximum sequence identity cutoff employed to generate the pristine dataset gives a representative and non-redundant set of sequences, BLAST (Altschul *et al.*, 1990) was used to search for homologues of each member sequence within that dataset. Default BLAST parameters were used, except for the low-complexity sequence filtering option, which was turned off. Of the 937 sequences tested, significant hits were obtained for only 18 sequences ($\sim$2% recall). The top hits for these 18 sequences are shown in Table (S1)

Query Sequence	Top Hit	E value	Oligomer state coincidence
1ECM	2VKL	0.005	Correct
1EUMC	1VLG	0.006	Incorrect
1GL2a	2NPSa	0.002	Correct
1GL2a	2NPSa	0.002	Correct
1GL2c	2NPSc	9e-08	Correct
1GL2d	2NPSd	6e-04	Correct
1X03	2Z0V	6e-04	Correct
1YBZ	2VKL	4e-06	Correct
2B9Ba	2WPQa	2e-08	Correct
2D8E	1YBZ	5e-04	Correct
2FXM	3BAS	3e-04	Correct
2WPQa	3BAS	5e-10	Incorrect
2Z0V	1X03	6e-04	Correct

Table S3. Summary of BLAST searches performed on the pristine dataset. In the last column, the result was recorded as being "correct" if the top hit and the query sequence shared the same experimental oligomeric state

These data show that simple homology methods are $\sim$85% accurate in predicting oligomer state when applied to our pristine dataset; that is $6/10$ recalled sequences had the same oligomer state as the query sequence. However, the low recall demonstrates that this method would be ineffective in practice, as only $\sim$3.5% of sequences in the pristine dataset would be assigned the correct oligomer state in this way. The performance of LOGICOIL was checked with each significant BLAST hit above omitted from the training set when scoring a test sequence (Figure S3). AUC analysis of the resulting ROC curves showed the performance to be very similar to the results of leave-one-out cross-validation studies reported in the previous sections: sequences $\geq$14 residues 0.75 vs 0.77; $\geq$21 residues 0.84 vs 0.86; $\geq$28 residues 0.86 vs 0.88.

These findings indicate that the 50% maximum identity cutoff used to generate the pristine dataset is sufficient to give a divergent set of coiled-coil sequences. Although this is somewhat higher than the identity cutoffs found for large globular proteins (often cited as 30%), coiled coils are usually encoded by short sequences exhibiting low complexity, meaning that higher levels of sequence identity are often shared between divergent sequences.

REFERENCES

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic Local Alignment Search Tool. *J Mol Biol*, **215**, 401–410.

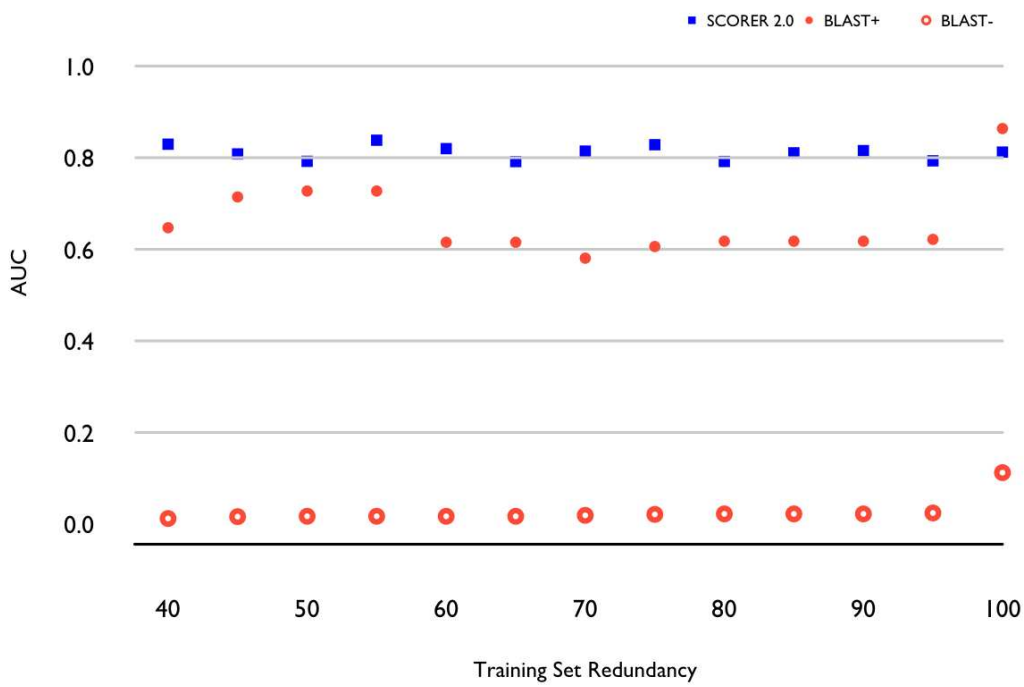

Fig. S1: Performance of LOGICOIL and BLAST when trained and tested with leave-one-out cross-validation on datasets with varying sequence redundancy.