BIOINFORMATICS

Vol. 00 no. 00 xxxx
Pages 1–3

Supplementary Information for "SCORER 2.0: An algorithm for distinguishing parallel dimeric and trimeric coiled-coil sequences."

Craig T. Armstrong [1 †] and 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]Department 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

1 EVALUATION OF THE PRISTINE DATASET THROUGH BLAST SEARCHES.

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 166 sequences tested, significant hits were obtained for only 10 sequences (∼6% recall). The top hits for these 10 sequences are shown in Table S1.

Query Sequence	Top Hit	E value	Oligomer state coincidence
1ci6_0_0	2oqq_0_1	0.002	Correct
1coi_3_2	1fmh_0_1	0.008	Inorrect
1fmh_0_0	3bas_0_1	0.004	Correct
1fmh_0_1	1coi_3_2	0.008	Incorrect
2b9b_3_32	3bas_0_1	5e-06	Incorrect
2efr_0_0	3bas_0_1	2e-11	Correct
2fxm_0_0	3bas_0_1	6e-5	Correct
2oqq_0_1	1ci6_0_0	0.002	Correct
1ci6_0_0	3bas_0_1	2e-07	Incorrect
3bas_0_1	2efr_0_0	2e-11	Correct

Table S1. 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 ∼60% 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 ∼3.5% of sequences in the pristine dataset would be assigned the correct oligomer state in this way.

The performance of SCORER 2.0 was checked with each significant BLAST hit above omitted from the training set when scoring a test sequence. 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 accompanying manuscript: 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.

*To whom correspondence should be addressed.
†Authors contributed equally to this paper.

2 SCORER 2.0 PREDICTIONS FOR THE PRISTINE DATASET WHEN TRAINED ON COILED-COIL DATA AT DIFFERENT REDUNDANCY CUTOFFS.

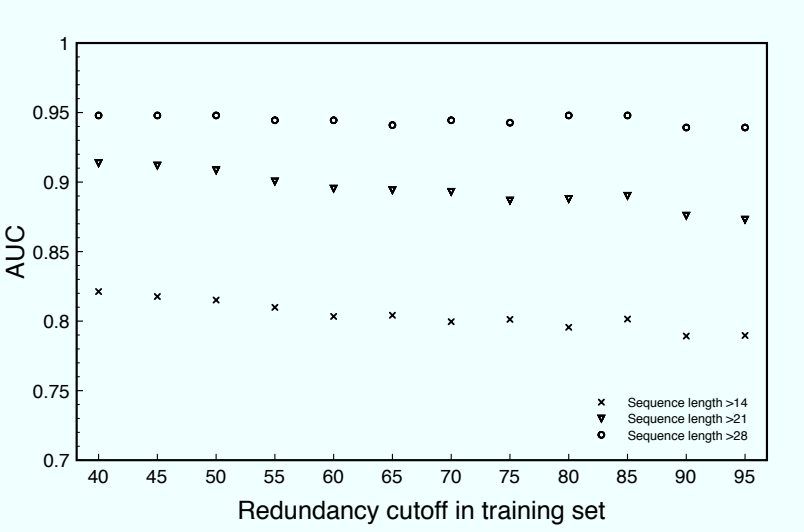

Fig. S1: AUC values of SCORER 2.0 when tested on the divergent dataset and trained on coiled-coil data at different redundancy cutoffs.

3 SCORER 2.0 AND SCORER PERFORMANCE FOR LEAVE-ONE-OUT CROSS-VALIDATION ON THE PRISTINE DATASET.

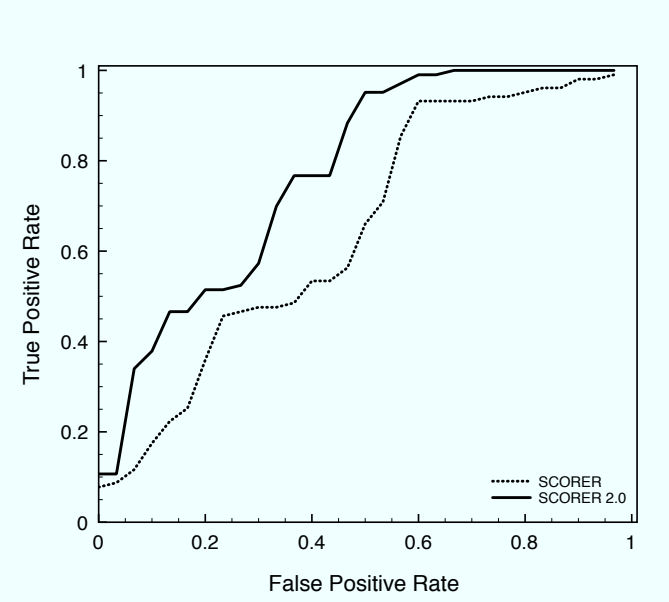

Fig. S2: ROC curves for SCORER and SCORER 2.0 when trained and tested under a leave-one-out cross-validation scheme for all sequences in the pristine dataset.

4 SCORER 2.0 PREDICTIONS ON ANTIPARALLEL DIMERS AND TETRAMERS.

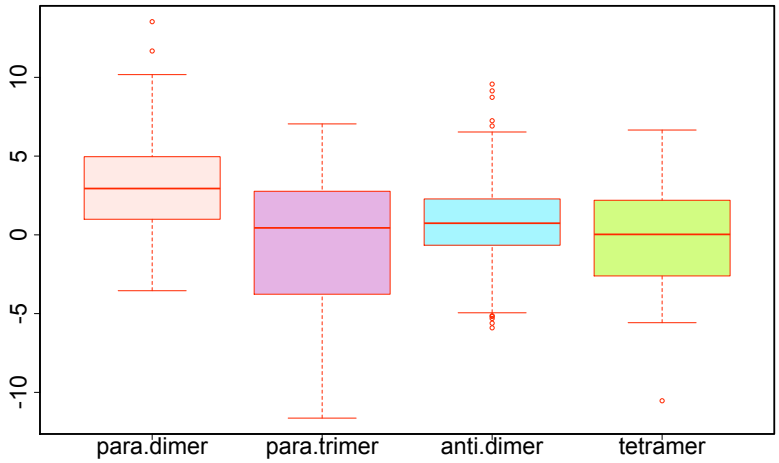

Fig. S3: Box-and-whiskers plots of the SCORER 2.0 scores for coiled-coil sequences with different oligomeric states. These plots give a five-number summary in which the sample minimum, the lower quartile, the value of the median prediction scores, the upper quartile and the sample maximum are graphically represented. Included oligomeric states are parallel dimers, parallel trimers, antiparallel dimers and tetramers. This was performed in order to investigate the behaviour of SCORER 2.0 on coiled-coil sequences known to be neither parallel dimers or parallel trimers.

REFERENCES

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, **215**(3), 403–410.