Pacific Northwest
NATIONAL LABORATORY

PNNL-32088

Structural- and Functional-Informed Machine Learning for Protein Function Prediction

September 2021

Jason E McDermott
Song Feng
Christine H Chang
Darren J Schmidt
Vincent G Danna

U.S. DEPARTMENT OF ENERGY

Prepared for the U.S. Department of Energy
under Contract DE-AC05-76RL01830

DISCLAIMER

PACIFIC NORTHWEST NATIONAL LABORATORY
operated by
BATTELLE
for the
UNITED STATES DEPARTMENT OF ENERGY
under Contract DE-AC05-76RL01830

Printed in the United States of America

Available to DOE and DOE contractors from the Office of Scientific and Technical Information,
P.O. Box 62, Oak Ridge, TN 37831-0062
www.osti.gov
ph: (865) 576-8401
fax: (865) 576-5728
email: reports@osti.gov

Available to the public from the National Technical Information Service
5301 Shawnee Rd., Alexandria, VA 22312
ph: (800) 553-NTIS (6847)
or (703) 605-6000
email: info@ntis.gov
Online ordering: http://www.ntis.gov

Structural- and Functional-Informed Machine Learning for Protein Function Prediction

September 2021

Jason E McDermott
Song Feng
Christine H Chang
Darren J Schmidt
Vincent G Danna

Prepared for
the U.S. Department of Energy
under Contract DE-AC05-76RL01830

Pacific Northwest National Laboratory
Richland, Washington 99354

Overview
In this project we aimed to extend methods for protein function prediction to include structural prediction data, and benchmark methods against existing tools. We proposed to apply the method to large metagenome datasets, and develop approaches to examine activity-based protein profiling results for protein function-structure patterns.

Nitrogen cycle protein families were previously identified and are used here to provide a proof-of-principle for use of structure prediction in protein function classification. The protein families are depicted in Figure 1.

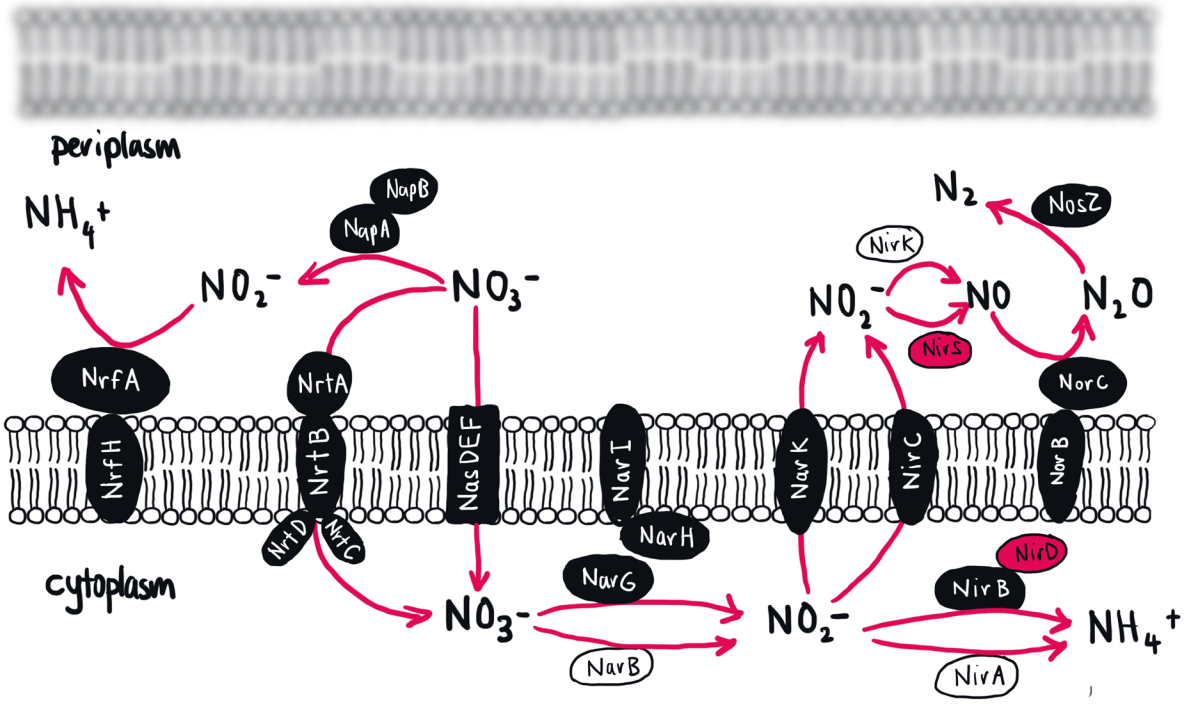

Figure 1. Nitrogen Cycle Families

Implementation
To accomplish these aims we have implemented the following:

1. We have applied secondary structure prediction and post-translational modification prediction algorithms to our existing nitrogen cycling gene family set. This set includes 28 families and over 4300 protein sequences.
2. We have benchmarked the performance of our Snekmer algorithm to develop classification models for all the protein families using native protein sequence (benchmarking against existing tools).
3. We have applied the Snekmer pipeline to the secondary structure prediction results and to the post-translational modification prediction results.
4. We have begun the process of installing the AlphaFold prediction software on PNNL computers. This will allow us to accurately predict 3-D structures of proteins from sequence and compare structural similarity with kmer-based sequence similarity for ability to discriminate between protein families.

5. We have assembled resources on translating protein tertiary structure to features for machine learning. Coupled with AlphaFold this will provide a generalizable approach to use of protein structure in function prediction from sequence.

Results

We first applied the Porter 4.0 secondary structure prediction program and the MuSite-DEEP post-translational modification prediction program to the nitrogen cycling family protein sequences. The MuSite-DEEP analysis is ongoing and is not discussed here.

Secondary structure prediction results in two levels of predictions: a three state prediction (SS3) that designates each amino acid as part of an unstructured coil (C), helix (H), or sheet (S); and an eight state prediction (SS8) which expands on the SS3 to include 3_{10} helix (G), alpha helix (H), π-helix (I), β-strand (E), bridge (B), turn (T), bend (S), and other unstructured coil (C).

We treated each state as an individual character (i.e. no further encoding or grouping of states was performed) in our kmer approach. Snekmer, our pipeline for automatic generation of protein function models, was applied to the 'native' protein sequence, the SS3, and the SS8 predictions using a kmer size of 14. For each of the native, SS3, and SS8 "encodings" probability scores were calculated for each kmer in each family model. Models were then built using logistic regression and 10-fold cross-validation used to assess performance. Performance was assessed using the standard receiver-operator characteristic curve (ROC) area under the curve (AUC). AUCROC will be 1.0 for perfect classification (all positive examples and negative examples correctly classified) and 0.5 for random classification.

Results from the models are shown in Table 1. Missing values indicate models for which the process failed. This failure is currently being debugged but we assume that the remainder of the model values are correct.

Importantly, the overall average of the AUC values across all protein family models shows that the secondary structure predictions perform as well or better than the native protein sequence. For a number of protein families (e.g. NirK1, NirK2, NrfA) the native protein sequence models failed to classify the protein families above random chance, but the SS3 and SS8 models performed quite well. This demonstrates that, for this set of parameters at least, predicted secondary structure can provide valuable information in terms of protein function classification.

Conclusions and Future Directions

Though our approach produced very promising results we acknowledge a number of limitations of our current work. First, the nitrogen cycling families are heterogenous: some are very limited in their intra-family sequence diversity, which makes them much easier to predict. However, we also observe that some families are more diverse in sequence and more difficult to predict overall. We will also choose several difficult protein families to assess for future work, including our previously characterized ubiquitin ligase mimic family. A second, related issue, is that we know from other work that using different lengths of kmer can change the results. In the future we will further parameterize this approach to better understand the limitations of the models. We believe that the results presented here make a strong case for the importance of considering protein structure in protein function prediction, and plan to examine the utility of three-dimensional protein structure (predicted by AlphaFold or experimentally determined, as available) in protein function prediction for future work.

Table 1. Results of Protein Family Models Using Predicted Secondary Structure by Snekmer

		ROC AUC			Predicted Secondary Structure		
Family	**Number**	**Sequence**	**SS3**	**SS8**	**C**	**H**	**S**
AmoA	23	0.98	0.99	1.00	0.3	0.7	0.0
NapB	306	0.96	1.00	1.00	0.7	0.2	0.0
NapC-NirT	111	1.00		0.95	0.4	0.6	0.0
NapE	45	0.81	0.93	0.99	0.3	0.7	0.0
NapF	117	0.84	0.43		0.6	0.3	0.1
NapG	117	1.00	0.99	1.00	0.6	0.3	0.1
NapH	143	0.99	1.00	1.00	0.4	0.6	0.0
NarH	390	1.00		0.99	0.6	0.3	0.1
NarJ	380	0.97	1.00	1.00	0.4	0.6	0.0
NarK	268	1.00	0.72	1.00	0.2	0.8	0.0
NasB-NirB	650	1.00	0.41	0.91	0.5	0.3	0.2
NifD	197	1.00	0.96	0.99	0.4	0.5	0.1
NirC	280	1.00	1.00	1.00	0.2	0.7	0.0
NirK1	49	0.50	1.00	1.00	0.6	0.1	0.3
NirK2	52	0.49	1.00	1.00	0.5	0.2	0.3
NirS	49	1.00	0.90	1.00	0.5	0.1	0.3
NorB	134	0.74	0.99	1.00	0.6	0.1	0.3
NosZI	91	1.00			0.5	0.1	0.4
NosZII	71	1.00		0.99	0.5	0.1	0.4
NrfA	55	0.50	1.00		0.4	0.5	0.0
NrfB	44	0.99		1.00	0.6	0.4	0.0
NrfC	39	1.00			0.6	0.3	0.2
NrfE	13	1.00	1.00	1.00	0.3	0.6	0.1
NrfF	46	0.99	0.98	1.00	0.3	0.7	0.0
NrfH	**101**	0.94		1.00	0.5	0.5	0.0
NtrB-NasE	215	1.00			0.3	0.7	0.0
NtrC-NasD	324	1.00	1.00	0.98	0.4	0.4	0.2
NxrA	13	1.00	0.96	1.00	0.6	0.3	0.1
Average		**0.917**	**0.912**	**0.991**			

Pacific Northwest National Laboratory

902 Battelle Boulevard
P.O. Box 999
Richland, WA 99354
1-888-375-PNNL (7665)

www.pnnl.gov