bioRxiv preprint doi: https://doi.org/10.1101/311209; this version posted April 30, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

NetSurfP-2.0: improved prediction of protein structural features by integrated deep learning

Michael Schantz Klausen[1], Martin Closter Jespersen[2], Henrik Nielsen[2], Kamilla Kjærgaard Jensen[2], Vanessa Isabell Jurtz[2], Casper Kaae Sønderby[3], Morten Otto Alexander Sommer[1], Ole Winther[3], Morten Nielsen[2,4], Bent Petersen[2,5,*], Paolo Marcatili[2,*]
1: Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, Kgs. Lyngby, Denmark
2: Department of Bio and Health Informatics, Technical University of Denmark, Kgs. Lyngby, Denmark
3: Department of Applied Mathematics and Computer Science, Technical University of Denmark, Kgs. Lyngby, Denmark
4: Instituto de Investigaciones Biotecnológicas, Universidad Nacional de San Martín, Buenos Aires, Argentina
5: Centre of Excellence for Omics-Driven Computational Biodiscovery (COMBio), Faculty of Applied Sciences, AIMST University, Kedah, Malaysia
[*] To whom correspondence should be addressed. Tel: +45 4525 2489; Fax: +45 4593 1585
; email pamar@bioinformatics.dtu.dk
Correspondence may also be addressed to Bent Petersen. Email: bent@bioinformatics.dtu.dk
The authors wish it to be known that, in their opinion, the first 2 authors should be regarded as joint First Authors

ABSTRACT
The ability to predict a protein's local structural features from the primary sequence is of paramount importance for unravelling its function if no solved structures of the protein or its homologs are available. Here we present NetSurfP-2.0 (http://services.bioinformatics.dtu.dk/service.php?NetSurfP-2.0), an updated and extended version of the tool that can predict the most important local structural features with unprecedented accuracy and run-time. NetSurfP-2.0 is sequence-based and uses an architecture composed of convolutional and long short-term memory neural networks trained on solved protein structures. Using a single integrated model, NetSurfP-2.0 predicts solvent accessibility, secondary structure, structural disorder, interface residues and backbone dihedral angles for each residue of the input sequences.
We assessed the accuracy of NetSurfP-2.0 on several independent validation datasets and found it to consistently produce state-of-the-art predictions for each of its output features. In addition to improved prediction accuracy the processing time has been optimized to allow predicting more than 1,000 proteins in less than 2 hours, and complete proteomes in less than 1 day.

INTRODUCTION
The Anfinsen experiment, showing that the structural characteristics of a protein are encoded in its primary sequence alone, is more than 50 years old (1). As a practical application of it, several methods have been developed over the last decades to predict from sequence only several protein structural features, including solvent accessibility, secondary structure, backbone geometry, disorder, and interface residues (2-7). These tools have tremendously impacted biology and chemistry, and some are among the most cited works in the field. NetSurfP-1.0 (8) is a tool published in 2009 for prediction of solvent accessibility and secondary structure using a feed-forward neural network architecture. Since then, deep learning techniques have affected the application of machine learning in biology expanding the ability of prediction tools to produce more accurate results on complex datasets (9-16).

bioRxiv preprint doi: https://doi.org/10.1101/311209; this version posted April 30, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Here we present NetSurfP-2.0, a new extended version of NetSurfP, that uses a deep neural network approach to accurately predict absolute and relative solvent accessibility, secondary structure using both 3- and 8-class definitions (17), structural disorder (18), φ and ψ dihedral angles, and interface residues of any given protein from its primary sequence only. By having an integrated deep model with several outputs, NetSurfP-2.0 can not only significantly reduce the computational time, but also achieve an improved accuracy that could not be reached by having separate models for each feature. In fact, when assessed on different validation sets with less than 25% sequence identity to any protein used in the training, its accuracy was consistently at pair or better than that of other state-of-the-art tools (3,4,15,19,20). In particular, we observed a significant increase in the accuracy of solvent accessibility and secondary structure over all the other tested methods.

To further improve its efficiency, NetSurfP-2.0 uses a different approach to make predictions for small and large sets of sequences, without compromising its accuracy. It has a user friendly interface allowing non-expert users to obtain and analyse their results via a browser, thanks to its graphical output, or to download them in several common formats for further analysis.

MATERIALS AND METHODS

We describe briefly the dataset and method used for training NetSurfP-2.0, and the validations performed.

Structural dataset

A structural dataset consisting of 12,185 crystal structures was obtained from the Protein Data Bank (PDB) (21), culled and selected by the PISCES server (22) with 25% sequence similarity clustering threshold and a resolution of 2.5 Å or better. To avoid overfitting, any cluster containing sequences from the validation datasets (see later for details) was removed, as well as peptide chains with less than 20 residues, leaving 10,837 sequences. Finally, we randomly selected 500 sequences (test dataset) to be left out for early stopping and parameter optimization, leaving 10,337 sequences for training.

Structural Features

For each chain in the training dataset we calculated its absolute and relative solvent accessibility (ASA and RSA, respectively), 3- and 8-class secondary structure classification (SS3 and SS8, respectively), and the backbone dihedral angles φ and ψ using the DSSP software (17). Interface residues were defined as all the residues in multi-chain complexes with an observed difference of more than 1 $Å^2$ between the ASA calculated on the individual chain and the ASA calculated on the whole biological unit defined in the PDB database.
Finally, each residue that was present in the chain refseq sequence, but not in the solved structure, was defined as disordered. It is important to mention that disordered residues cannot be annotated with any of the other features, since no atomic coordinates are available for these residues.

Protein sequence profiles

NetSurfP-2.0, likewise its predecessor, exploits sequence profiles of the target protein for its prediction. We used two different ways of generating such profiles. The

bioRxiv preprint doi: https://doi.org/10.1101/311209; this version posted April 30, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

first exploits the HH-suite software (23), that runs quickly on individual sequences, while the second uses the MMseqs2 software (24), that is optimized for searches on large data sets. In both cases, the profile-generation tools are run with default parameters, except MMseqs2 which used 2 iterations with the `--max-seqs` parameter set to 2,000.

Deep Network architecture

The model was implemented using the Keras library. The input layer of the model consists of the one-hot (sparse) encoded sequences (20 features) plus the full HMM profiles from HH-suite (30 features in total, comprising 20 features for the amino acid profile, 7 features for state transition probabilities, and 3 features for local alignment diversity), giving a total of 50 input features. This input is then connected to two Convolutional Neural Network (CNN) layers, consisting of 32 filters each with size 129 and 257, respectively. The CNN output is concatenated with the initial 50 input features and connected to two bidirectional long short-term memory (LSTM) layers with 1024 nodes (figure 1, panel A).

Each output (RSA, SS8, SS3, φ, ψ, disorder, and interface) is calculated with a Fully Connected (FC) layer using the outputs from the final LSTM layer. RSA is encoded as a single output between 0 and 1. ASA output is not directly predicted, but calculated by multiplying RSA and ASAmax (25). SS8, SS3, disorder, and interface are encoded as 8, 3, or 2 outputs with the target encoded as a sparse vector (target is set to 1, while rest of the elements are 0). φ and ψ are each encoded as a vector of length 2, where the first element is the sine of the angle and the second element is the cosine. This encoding reduces the effect of the periodicity of the angles (26), and the predicted angle can be calculated with the arctan2 function.

Training

The training was performed using minibatches of size 15. The individual learning rate of each neuron was optimized using the Adam function (27). Early stopping was performed on the test dataset. Since the different target values for each output have different distributions, a weighted sum of different loss functions were used: SS8, SS3, disorder, and interface use cross entropy loss, while RSA, φ, and ψ use mean squared error loss. Weights were adjusted so each loss contribution was approximately equal and then fine-tuned for maximum overall performance. When the target value for a feature of a given residue was missing, e.g. for secondary structure of disordered residues, or φ angles of N-terminal residues, the loss for that output was set to 0.

Evaluation

The final two models trained with the HH-suite and MMseqs2 profiles, respectively, were tested on 3 independent datasets: the TS115 (115 proteins) and CB513 datasets (513 protein regions from 434 proteins) (28) and a dataset consisting of all the free-modeling targets (21 proteins) at the last CASP 12 experiment (29). No protein with more than 25% sequence identity to the proteins in these datasets was present in the training. Disorder prediction was not performed on the CB513 dataset, since it contains very few disordered regions.

We used different metrics to evaluate each feature: Pearson's correlation coefficient (PCC) for solvent accessibilities, Q3 and Q8 accuracy for SS3 and SS8 respectively

bioRxiv preprint doi: https://doi.org/10.1101/311209; this version posted April 30, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

(15), mean absolute error in degrees for φ and ψ angles (MAE), and Matthew's correlation coefficient (MCC) for interface and disorder.

WEB INTERFACE

To use NetSurfP-2.0 (http://services.bioinformatics.dtu.dk/service.php?NetSurfP-2.0), only the sequences of the proteins of interest in fasta format are required. Up to 4,000 sequences or 4,000,000 residues overall can be submitted per job. Only amino acid sequences are accepted. For submissions of less than 100 sequences, the HH-suite model is used, the MMseq2 model otherwise. Upon submission, a queuing page appears. The user can either wait until the job is finished and the results will automatically be displayed, or, for larger submissions which might take up to a few hours to complete, the user has the option to provide an email address and the result page link will be emailed when the job is completed.

The NetSurfP-2.0 output page (figure 2) contains a navigation bar with various tabs. The "Server Output" tab shows each individual chain result as a graphically annotated sequence. The protein sequence is on the top, and below it is the RSA value of each residue: residues with RSA of more than 25% are displayed in red and with positive values, residues with less than 25% is in blue and with negative values. Below this, there is a representation of 3 state secondary structure predictions with different symbols for helices (in red/orange), strands (purple) and loops (pink). The next line represents the disorder probability as a grey ribbon, with a thicker ribbon representation for residues with higher disorder score. Finally, the sequence numbering according to the submitted sequence comes last. When hovering on a specific residue in the sequence, all predictions are displayed as a tooltip. For larger jobs, not all the sequences are displayed at once. It is possible to browse through the results by either clicking on the page numbers at the bottom of the page, or by using the "Search protein ID's" textbox on the top right. Individual results can be downloaded by clicking on the grey "Export" button on top of each sequence. The button "Export all" on the top right allows exporting all the results at once. The results can be saved in Json, csv, or NetSurfP-1.0 format, or as a combined zip containing all of the above formats and all the files generated for the prediction.

All the most common browsers are supported. A more detailed description of the web server can be found on the NetSurfP-2.0 Help page.

RESULTS

We have compared the performance of NetSurfP-2.0 to other state-of-the-art tools with similar functionality: NetSurfP-1.0 (8), Spider3 (4), SPOT-Disorder (3), RaptorX (15,20), and JPred4 (19). It should be noticed that NetSurfP-2.0 did not include the validation datasets in its training.

In order to check whether the results of the methods are significantly different, we calculated a p-value for each feature by using a pairwise Student's t-test on the results of the two methods. Results are presented in table 1.

NetSurfP-2.0 outperforms all other methods in all the tests. The largest improvements are observed for solvent accessibility and disorder predictions.

Both the HH-suite and MMseqs2 models perform similarly on all datasets tested. However, they have very different running time: the runtime on a single protein sequence for the HH-suite model is approximately 2 minutes, but it scales linearly

bioRxiv preprint doi: https://doi.org/10.1101/311209; this version posted April 30, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

with the number of sequences. MMseqs2, conversely, is slower for small datasets, but on large datasets it provides a speed-up of up to 50 times and the ability to parallelise on multiple processor (figure 1, panel B). NetSurfP-2.0 uses the HH-suite model for searches of less than 100 sequences, and the MMseqs2 model otherwise, thus offering a good trade-off between computation time and resource demand, without sacrificing the method's accuracy.
An example of the ASA and SS3 predictions for the human Orotate phosphoribosyltransferase (OPRTase) domain, displayed on its solved structure (PDB id 2WNS) is illustrated in figure 3.

DISCUSSION

The NetSurfP-2.0 web server provides, to the best of our knowledge, the state-of-the-art sequence-based prediction for solvent accessibility, secondary structure, disorder, and backbone geometry. We believe that the high accuracy of the method is achieved as a result of the combination of the server architecture and the careful integration of different structural data.
By training a weight-sharing integrated model with several structural features, we improve the accuracy of disorder and interface residues with respect to models trained on individual features. We believe that this improvement is due to a more robust and informative internal state of the system, which is extremely valuable for features where only a few positives are present on average.
This integration was possible because we used a proper representation of the structural data. Other tools are not trained on the real protein sequences, but on the residues that are observed in the solved structure. In this way, the models are presented with cases that are neither physically nor biologically meaningful, such as residues divided by a disordered region, that are far apart in primary and tertiary structure but presented to the model as consecutive. In contrast, by using a recently developed strategy (12), we can train the model on all residues, including the disordered ones, thus increasing the accuracy of annotated features in the data and reduce the frustration during training.
Our model predicts also residues that are part of interfaces. In the last few years, a plethora of methods exploiting evolutionary-derived constraints to derive reside-residue interactions have been developed. These methods have good performance if sufficiently large alignment for both interacting proteins can be produced. If this is not the case, for example for proteins with few known homologs, or if only one of the interacting partners is available, our method presents a valuable alternative to identify interface residues.
Thanks to its accuracy, its fast computation time, and its easy and intuitive interface, we believe that NetSurfP-2.0 will become a valuable resource that will aid researchers both with and without extensive computational knowledge to analyse and understand protein structure and function.

ACKNOWLEDGEMENT

FUNDING: MSK and MOAS acknowledged funding from the Novo Nordisk Foundation.

REFERENCES

bioRxiv preprint doi: https://doi.org/10.1101/311209; this version posted April 30, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

1. Anfinsen, C.B., Haber, E., Sela, M. and White, F.H., Jr. (1961) The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. *Proc Natl Acad Sci U S A*, **47**, 1309-1314.
http://www.ncbi.nlm.nih.gov/pubmed/13683522

2. Cuff, J.A., Clamp, M.E., Siddiqui, A.S., Finlay, M. and Barton, G.J. (1998) JPred: a consensus secondary structure prediction server. *Bioinformatics*, **14**, 892-893.
http://www.ncbi.nlm.nih.gov/pubmed/9927721

3. Hanson, J., Yang, Y., Paliwal, K. and Zhou, Y. (2017) Improving protein disorder prediction by deep bidirectional long short-term memory recurrent neural networks. *Bioinformatics*, **33**, 685-692.
http://www.ncbi.nlm.nih.gov/pubmed/28011771
http://dx.doi.org/10.1093/bioinformatics/btw678

4. Heffernan, R., Yang, Y., Paliwal, K. and Zhou, Y. (2017) Capturing non-local interactions by long short-term memory bidirectional recurrent neural networks for improving prediction of protein secondary structure, backbone angles, contact numbers and solvent accessibility. *Bioinformatics*, **33**, 2842-2849.
http://www.ncbi.nlm.nih.gov/pubmed/28430949
http://dx.doi.org/10.1093/bioinformatics/btx218

5. Jones, D.T. (1999) Protein secondary structure prediction based on position-specific scoring matrices. *J Mol Biol*, **292**, 195-202.
http://www.ncbi.nlm.nih.gov/pubmed/10493868
http://dx.doi.org/10.1006/jmbi.1999.3091

6. McGuffin, L.J., Bryson, K. and Jones, D.T. (2000) The PSIPRED protein structure prediction server. *Bioinformatics*, **16**, 404-405.
http://www.ncbi.nlm.nih.gov/pubmed/10869041

7. Rost, B. (1996) PHD: predicting one-dimensional protein structure by profile-based neural networks. *Methods Enzymol*, **266**, 525-539.
http://www.ncbi.nlm.nih.gov/pubmed/8743704

8. Petersen, B., Petersen, T.N., Andersen, P., Nielsen, M. and Lundegaard, C. (2009) A generic method for assignment of reliability scores applied to solvent accessibility predictions. *BMC Struct Biol*, **9**, 51.
http://www.ncbi.nlm.nih.gov/pubmed/19646261
http://dx.doi.org/10.1186/1472-6807-9-51

9. Almagro Armenteros, J.J., Sonderby, C.K., Sonderby, S.K., Nielsen, H. and Winther, O. (2017) DeepLoc: prediction of protein subcellular localization using deep learning. *Bioinformatics*, **33**, 3387-3395.
http://www.ncbi.nlm.nih.gov/pubmed/29036616
http://dx.doi.org/10.1093/bioinformatics/btx431

bioRxiv preprint doi: https://doi.org/10.1101/311209; this version posted April 30, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

10. Esteva, A., Kuprel, B., Novoa, R.A., Ko, J., Swetter, S.M., Blau, H.M. and Thrun, S. (2017) Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, **542**, 115-118.
http://www.ncbi.nlm.nih.gov/pubmed/28117445
http://dx.doi.org/10.1038/nature21056

11. Hansen, C.S., Osterbye, T., Marcatili, P., Lund, O., Buus, S. and Nielsen, M. (2017) ArrayPitope: Automated Analysis of Amino Acid Substitutions for Peptide Microarray-Based Antibody Epitope Mapping. *PLoS One*, **12**, e0168453.
http://www.ncbi.nlm.nih.gov/pubmed/28095436
http://dx.doi.org/10.1371/journal.pone.0168453

12. Jurtz, V., Paul, S., Andreatta, M., Marcatili, P., Peters, B. and Nielsen, M. (2017) NetMHCpan-4.0: Improved Peptide-MHC Class I Interaction Predictions Integrating Eluted Ligand and Peptide Binding Affinity Data. *J Immunol*, **199**, 3360-3368.
http://www.ncbi.nlm.nih.gov/pubmed/28978689
http://dx.doi.org/10.4049/jimmunol.1700893

13. Jurtz, V.I., Johansen, A.R., Nielsen, M., Almagro Armenteros, J.J., Nielsen, H., Sonderby, C.K., Winther, O. and Sonderby, S.K. (2017) An introduction to deep learning on biological sequence data: examples and solutions. *Bioinformatics*, **33**, 3685-3690.
http://www.ncbi.nlm.nih.gov/pubmed/28961695
http://dx.doi.org/10.1093/bioinformatics/btx531

14. Min, S., Lee, B. and Yoon, S. (2017) Deep learning in bioinformatics. *Brief Bioinform*, **18**, 851-869.
http://www.ncbi.nlm.nih.gov/pubmed/27473064
http://dx.doi.org/10.1093/bib/bbw068

15. Wang, S., Peng, J., Ma, J. and Xu, J. (2016) Protein Secondary Structure Prediction Using Deep Convolutional Neural Fields. *Sci Rep*, **6**, 18962.
http://www.ncbi.nlm.nih.gov/pubmed/26752681
http://dx.doi.org/10.1038/srep18962

16. Wang, S., Sun, S., Li, Z., Zhang, R. and Xu, J. (2017) Accurate De Novo Prediction of Protein Contact Map by Ultra-Deep Learning Model. *PLoS Comput Biol*, **13**, e1005324.
http://www.ncbi.nlm.nih.gov/pubmed/28056090
http://dx.doi.org/10.1371/journal.pcbi.1005324

17. Kabsch, W. and Sander, C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577-2637.
http://www.ncbi.nlm.nih.gov/pubmed/6667333
http://dx.doi.org/10.1002/bip.360221211

bioRxiv preprint doi: https://doi.org/10.1101/311209; this version posted April 30, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

18. Linding, R., Jensen, L.J., Diella, F., Bork, P., Gibson, T.J. and Russell, R.B. (2003) Protein disorder prediction: implications for structural proteomics. *Structure*, **11**, 1453-1459.
http://www.ncbi.nlm.nih.gov/pubmed/14604535

19. Drozdetskiy, A., Cole, C., Procter, J. and Barton, G.J. (2015) JPred4: a protein secondary structure prediction server. *Nucleic Acids Res*, **43**, W389-394.
http://www.ncbi.nlm.nih.gov/pubmed/25883141
http://dx.doi.org/10.1093/nar/gkv332

20. Wang, S., Li, W., Liu, S. and Xu, J. (2016) RaptorX-Property: a web server for protein structure property prediction. *Nucleic Acids Res*, **44**, W430-435.
http://www.ncbi.nlm.nih.gov/pubmed/27112573
http://dx.doi.org/10.1093/nar/gkw306

21. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. and Bourne, P.E. (2000) The Protein Data Bank. *Nucleic Acids Res*, **28**, 235-242.
http://www.ncbi.nlm.nih.gov/pubmed/10592235

22. Dunbrack, R., Giguere, L.A. and St-Pierre, J.F. (2009) A comparison of gut evacuation models for larval mackerel (Scomber scombrus) using serial photography. *J Fish Biol*, **74**, 906-920.
http://www.ncbi.nlm.nih.gov/pubmed/20735607
http://dx.doi.org/10.1111/j.1095-8649.2008.02177.x

23. Remmert, M., Biegert, A., Hauser, A. and Soding, J. (2011) HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat Methods*, **9**, 173-175.
http://www.ncbi.nlm.nih.gov/pubmed/22198341
http://dx.doi.org/10.1038/nmeth.1818

24. Steinegger, M. and Soding, J. (2017) MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat Biotechnol*, **35**, 1026-1028.
http://www.ncbi.nlm.nih.gov/pubmed/29035372
http://dx.doi.org/10.1038/nbt.3988

25. Ahmad, S., Gromiha, M.M. and Sarai, A. (2003) Real value prediction of solvent accessibility from amino acid sequence. *Proteins*, **50**, 629-635.
http://www.ncbi.nlm.nih.gov/pubmed/12577269
http://dx.doi.org/10.1002/prot.10328

26. Lyons, J., Dehzangi, A., Heffernan, R., Sharma, A., Paliwal, K., Sattar, A., Zhou, Y. and Yang, Y. (2014) Predicting backbone Calpha angles and dihedrals from protein sequences by stacked sparse auto-encoder deep neural network. *J Comput Chem*, **35**, 2040-2046.
http://www.ncbi.nlm.nih.gov/pubmed/25212657
http://dx.doi.org/10.1002/jcc.23718

bioRxiv preprint doi: https://doi.org/10.1101/311209; this version posted April 30, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

27. Kingma, D.P. and Ba, J. (2014) Adam: A Method for Stochastic Optimization. *arXiv:1412.6980*.

28. Cuff, J.A. and Barton, G.J. (1999) Evaluation and improvement of multiple sequence methods for protein secondary structure prediction. *Proteins*, **34**, 508-519.
http://www.ncbi.nlm.nih.gov/pubmed/10081963

29. Abriata, L.A., Tamo, G.E., Monastyrskyy, B., Kryshtafovych, A. and Dal Peraro, M. (2017) Assessment of hard target modeling in CASP12 reveals an emerging role of alignment-based contact prediction methods. *Proteins*.
http://www.ncbi.nlm.nih.gov/pubmed/29139163
http://dx.doi.org/10.1002/prot.25423

TABLE AND FIGURES LEGENDS

CASP12	RSA [PCC]	ASA [PCC]	SS3 [Q3]	SS8 [Q8]	Disorder [MCC]	Phi [MAE]	Psi [MAE]	Interface [MCC]
NetSurfP-2.0 (mmseqs)	**0.726**	*0.735*	*0.820*	*0.703*	**0.660**	*20.3*	*31.8*	**0.063**
NetSurfP-2.0 (hhblits)	*0.725*	**0.737**	**0.824**	**0.711**	*0.604*	**20.0**	**31.2**	*0.038*
NetsurfP-1.0	0.617	0.641	0.709					
Spider3		0.688	0.791		*0.582*	21.6	33.2	
RaptorX			0.786	0.661	*0.621*			
Jpred4			0.760					
TS115	**RSA [PCC]**	**ASA [PCC]**	**SS3 [Q3]**	**SS8 [Q8]**	**Disorder [MCC]**	**Phi [MAE]**	**Psi [MAE]**	**Interface [MCC]**
NetSurfP-2.0 (mmseqs)	**0.778**	**0.797**	**0.857**	**0.750**	*0.656*	**17.2**	**25.8**	*0.311*
NetSurfP-2.0 (hhblits)	*0.775*	*0.795*	*0.853*	0.744	**0.663**	*17.5*	*26.5*	**0.319**
NetsurfP-1.0	0.661	0.691	0.779					
Spider3		0.769	0.839		*0.575*	18.5	27.3	
RaptorX			0.822	0.716	*0.567*			
Jpred4			0.767					
CB513	**RSA [PCC]**	**ASA [PCC]**	**SS3 [Q3]**	**SS8 [Q8]**	**Disorder [MCC]**	**Phi [MAE]**	**Psi [MAE]**	**Interface [MCC]**
NetSurfP-2.0 (mmseqs)	**0.794**	**0.807**	**0.854**	**0.723**		**20.1**	**28.0**	0.283
NetSurfP-2.0 (hhblits)	*0.788*	*0.803*	*0.853*	*0.720*		*20.2*	28.6	**0.321**
NetsurfP-1.0	0.700	0.723	0.788					
Spider3		0.797	0.845			*20.4*	*28.2*	
RaptorX			0.827	0.706				
Jpred4			0.779					

bioRxiv preprint doi: https://doi.org/10.1101/311209; this version posted April 30, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Table 1. Results of the method's validation on independent datasets. The performance of NetSurfP-2.0 (using HH-suite and MMSeqs2 profiles), NetSurfP-1.0, Spider3, SPOT-disorder, RaptorX, and JPred4, is displayed for the CASP12, TS115, and CB513 datasets. SPOT-disorder and Spider3 predictions are reported as a single row. The following performance metrics are used: Pearson Correlation Coefficient (PCC), Matthew's Correlation Coefficient (MCC), Q3 and Q8 accuracy, and mean absolute error (MAE) in degrees. The different predicted features are reported in the column header, together with the corresponding performance metric. For each feature and each dataset, the best score is reported in bold. Scores in italics are the ones for which no significant difference with respect to the top scoring method (calculated using a 2-tailed paired Student's t-test and a significance threshold of 0.05) is observed. Greyed-out cells represent predictions that were not performed, either because not part of a method's output, or because the feature was not present in the corresponding dataset.

Figure 1. Network architecture and computation time plot. In panel the Network architecture is shown. N is the number of amino acids in the target protein sequence. Each box represents a different layer of the network, from the input (bottom) to the output (top), and the corresponding number of neurons/filters. The arrows represent the features that are passed between consecutive layers. The computation time per sequence of NetSurfP-2.0 is reported in Panel B. The x-axis represents the number of input sequences (logarithmic scale), the y-axis the average computation time in seconds per sequence. The method implementation using HH-suite profiles is reported as a grey dashed line, and the one using MMSeqs2 profiles is reported as a solid black line.

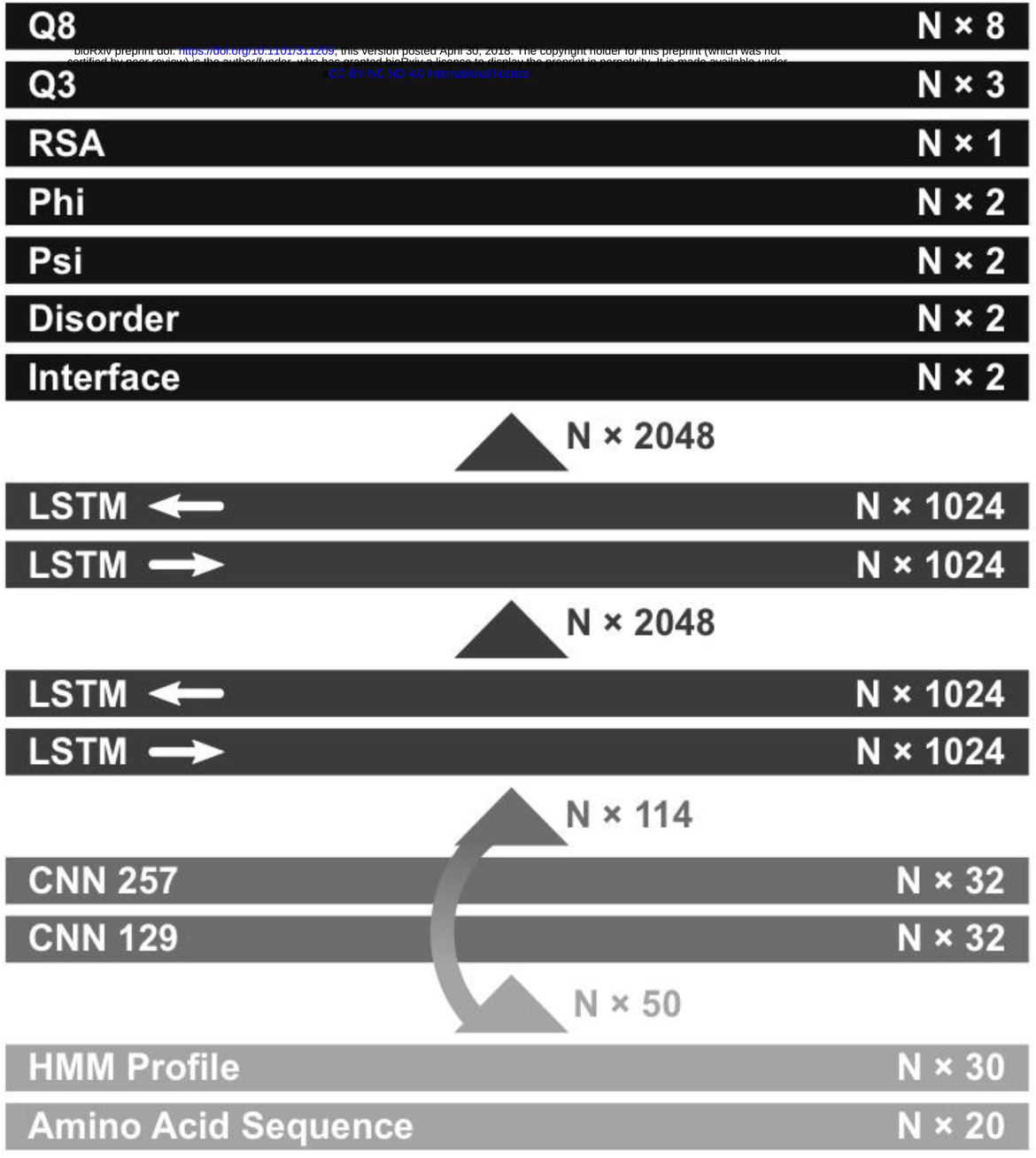
Figure 2. NetSurfP-2.0 web server result page.

bioRxiv preprint doi: https://doi.org/10.1101/311209; this version posted April 30, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

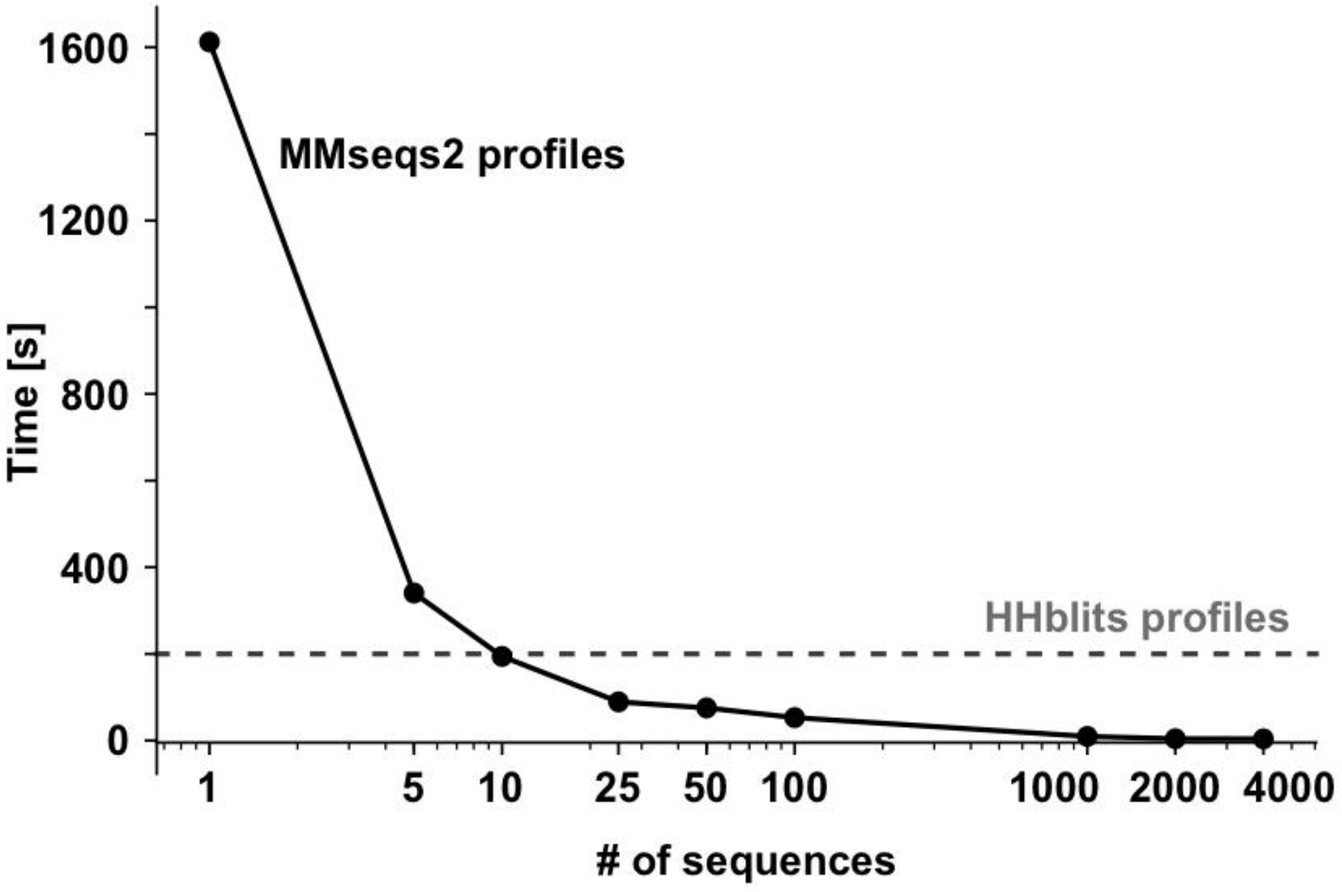
Figure 3. NetSurfP-2.0 predictions mapped on the OPRTase domain structure. Panel A represents the predicted ASA in a color gradient from blue (low) to red (high). Panel B represents SS3 Helix, Strand, and Coil classes in orange, purple, and pink, respectively. The actual secondary structure of the protein is displayed in the carton representation of the structure. Both color codings are consistent with the web server graphical output.

bioRxiv preprint doi: https://doi.org/10.1101/311209; this version posted April 30, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under CC-BY-NC-ND 4.0 International license

A) Deep model architecture overview



B) Computational time per sequence



NetSurfP-2.0

bioRxiv preprint doi: https://doi.org/10.1101/311209; this version posted April 30, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

NetSurfP server predicts the surface accessibility, secondary structure, disorder, phi/psi dihedral angles and protein-protein binding propensity of amino acids in an amino acid sequence.

Submission | Help | Abstract/Cite | Data | Server Output | Export All

Search Protein ID's

Showing 2 Predictions

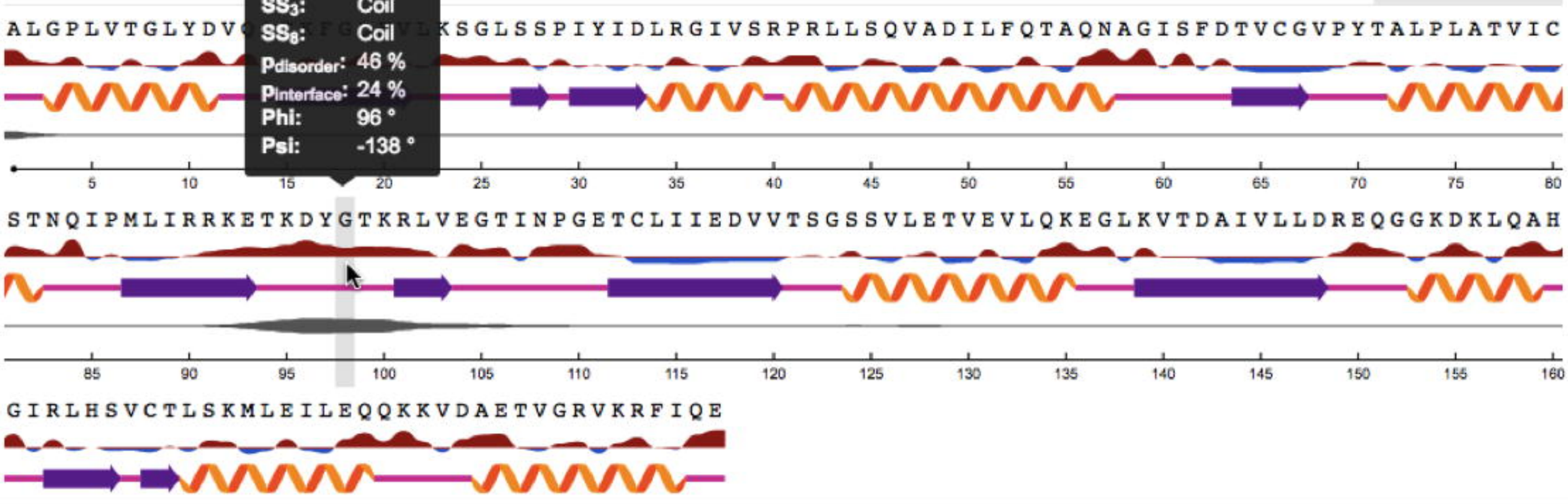
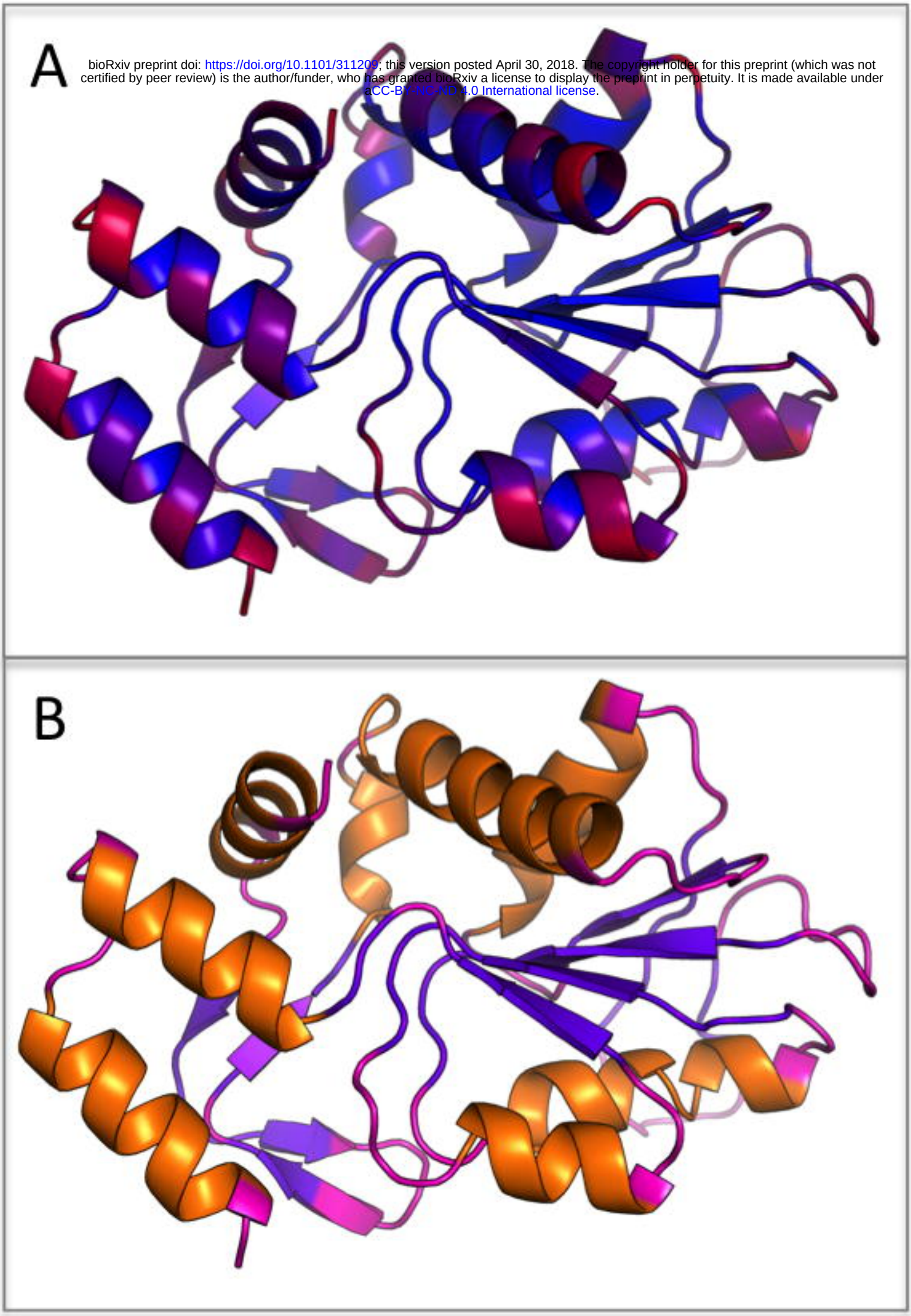
Below is a graphical representation of 359 residue predictions across 2 sequences. Running time was 215 seconds (108 seconds per sequence). Hover your mouse over a sequence position to see all outputs.

Relative Surface Accessibility: Red is exposed and blue is buried, thresholded at 25%.
Secondary Structure: Helix, Strand, Coil.
Disorder: Thickness of line equals probability of disordered residue.

2WNS.A.1 TRANSFER

Export 2WNS.A.1

RSA:	55 %
ASA:	43 Å
SS_3:	Coil
SS_8:	Coil
$P_{disorder}$:	46 %
$P_{interface}$:	24 %
Phi:	96 °
Psi:	-138 °



bioRxiv preprint doi: https://doi.org/10.1101/311209; this version posted April 30, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.