



qNABpredict: Quick, accurate, and taxonomy-aware sequence-based prediction of content of nucleic acid binding amino acids

Zhonghua Wu¹  | Sushmita Basu² | Xuantai Wu¹  | Lukasz Kurgan² 

¹School of Mathematical Sciences and LPMC, Nankai University, Tianjin, China

²Department of Computer Science, Virginia Commonwealth University, Richmond, Virginia, USA

Correspondence

Lukasz Kurgan, Department of Computer Science, Virginia Commonwealth University, Richmond, VA, USA.
Email: lkurgan@vcu.edu

Review Editor: Nir Ben-Tal

Abstract

Protein sequence-based predictors of nucleic acid (NA)-binding include methods that predict NA-binding proteins and NA-binding residues. The residue-level tools produce more details but suffer high computational cost since they must predict every amino acid in the input sequence and rely on multiple sequence alignments. We propose an alternative approach that predicts content (fraction) of the NA-binding residues, offering more information than the protein-level prediction and much shorter runtime than the residue-level tools. Our first-of-its-kind content predictor, qNABpredict, relies on a small, rationally designed and fast-to-compute feature set that represents relevant characteristics extracted from the input sequence and a well-parametrized support vector regression model. We provide two versions of qNABpredict, a taxonomy-agnostic model that can be used for proteins of unknown taxonomic origin and more accurate taxonomy-aware models that are tailored to specific taxonomic kingdoms: archaea, bacteria, eukaryota, and viruses. Empirical tests on a low-similarity test dataset show that qNABpredict is 100 times faster and generates statistically more accurate content predictions when compared to the content extracted from results produced by the residue-level predictors. We also show that qNABpredict's content predictions can be used to improve results generated by the residue-level predictors. We release qNABpredict as a convenient webserver and source code at <http://biomine.cs.vcu.edu/servers/qNABpredict/>. This new tool should be particularly useful to predict details of protein–NA interactions for large protein families and proteomes.

KEYWORDS

prediction, protein function, protein–nucleic acids interactions, protein sequence

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial](https://creativecommons.org/licenses/by-nc/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

© 2022 The Authors. *Protein Science* published by Wiley Periodicals LLC on behalf of The Protein Society.

1 | INTRODUCTION

Interactions between proteins and nucleic acids are instrumental for a broad range of cellular functions that include transcription, translation, gene expression and regulation, DNA repair, RNA splicing and regulation, and chromatin packing and remodeling, to name just a few (Charoensawan et al., 2010; Glisovic et al., 2008; Kelaini et al., 2021; Malhotra & Sowdhamini, 2013; Wang et al., 2015). Molecular-level details of these interactions are studied using the X-ray crystallography, electron microscopy, and nuclear magnetic resonance, with over 11,000 structures of protein–DNA and protein–RNA complexes in the Protein Data Bank (PDB) (Berman et al., 2000; Burley et al., 2021). However, these investigations are unable to keep up with the rapid accumulation of protein and nucleic acid (NA) sequence data (O'Leary et al., 2016). Consequently, faster and more cost-effective computational tools that predict protein–NA interactions from protein sequences are being developed to support efforts related to the functional characterization of proteins (Emamjomeh et al., 2019; Gromiha & Nagarajan, 2013; Miao & Westhof, 2015; Puton et al., 2012; Si et al., 2015a, 2015b; Walia et al., 2012; Wang et al., 2020; Yan et al., 2016; Zhang et al., 2019b; Zhao et al., 2013). They are trained using the limited amount of experimental molecular-level data and are applied to identify putative protein–NA interactions in a high-throughput manner for the uncharacterized sequences. These methods are divided into two distinct groups: protein-level tools that predict whether a target protein sequence interacts with DNA and RNA vs. residue-level tools that predict DNA and RNA binding residues in a given protein sequence. A few recently released tools that predict DNA/RNA-binding proteins include iDRBP-ECHF (Feng et al., 2022), iDRBP_MMC (Zhang et al., 2020), iDRBP-EL (Wang et al., 2021), and a method by Du and Hu (2022). The residue-level tools produce more detailed information when compared to the protein-level approaches. Representative methods that belong to this category include BindN (Wang & Brown, 2006), BindN+ (Wang et al., 2010), DRNAPred (Yan & Kurgan, 2017), ProNA2020 (Qiu et al., 2020), and HybridNAP (Zhang et al., 2019b). While we focus on the predictions from the protein sequences, we also acknowledge efforts to identify protein binding regions in the NA sequences (Laverty et al., 2022; Ma et al., 2022; Park et al., 2014; Su et al., 2019; Zhang et al., 2022).

While residue-level tools provide useful putative data concerning presence and positions of NA-binding residues in the protein chains, they suffer a relatively high computational cost since they must make predictions for each amino acid in the input sequence and they utilize

computationally costly multiple sequence alignment. An alternative approach is to predict the content of the NA-binding residues, that is, the fraction of the residues in the input sequence that bind DNA or RNA, which ranges between 0 and 1. This type of prediction should be faster since it is done at the protein-level and potentially could be realized without utilizing sequence alignment, while providing substantially more information than the protein-level prediction, that is, it identifies which proteins interact with NAs and what is the relative size of the binding interface. The advantage of the short runtime is particularly useful for large-scale applications that analyze protein families and even entire proteomes. While to the best of our knowledge there are no content predictors for the NA-binding, these predictors are available in related areas including prediction of the secondary structures content (Chen et al., 2011; Homaeian et al., 2007; Horne, 1988; Krigbaum & Knutton, 1973; Lee et al., 2006; Lin & Pan, 2001; Liu & Chou, 1999; Mizianty et al., 2011; Ruan et al., 2005; Yu et al., 2022) and the intrinsic disorder content (Yan et al., 2013). The first secondary structure content predictor was developed in 1973 (Krigbaum & Knutton, 1973), while the latest was released in 2022 (Yu et al., 2022). They are applied in the large-scale studies with several illustrative example for the recent disorder content predictor, RAPID (Yan et al., 2013), which was recently used to perform analysis in several protein families (Gawron et al., 2016; Skupien-Rabian et al., 2016; Warren & Shechter, 2017) and whole proteomes (Boone et al., 2021; Choura et al., 2020).

To this end, we propose first-of-its-kind, fast and accurate predictor of the content of NA-binding residues in protein sequences, qNABpredict (quick Nucleic Acids Binding content predictor). We conceptualize, design, test and release two versions of this tool, one that is taxonomy-aware and relies on models that are tailored for proteins from the four taxonomic kingdoms: eukaryota, archaea, bacteria, and viruses; and the second that provides taxonomy-agnostic predictions. The first model provides more accurate results but requires the knowledge of the underlying organism(s), while the latter can be used for any protein sequence.

2 | RESULTS AND DISCUSSION

2.1 | Comparative assessment of NA-binding residues content predictions

We compare the NA-binding residues content predictions generated by qNABpredict, in both the taxonomy-aware and taxonomy-agnostic modes, with the results produced by the predictors of the residue-level NA-binding residues

TABLE 1 Results of the comparative analysis on the low-similarity test dataset

Predictor	MAE	MAE for the non-NA-binding proteins	SCC
Baseline (random_correct_range)	0.421 ^{+/+}	0.489 ^{+/+}	0.000 ^{+/+}
Baseline (random_resampled_content)	0.160 ^{+/+}	0.109 ^{+/+}	0.053 ^{+/+}
BindN+	0.152 ^{+/+}	0.170 ^{+/+}	0.529 ^{+/+}
DRNAPred	0.147 ^{+/+}	0.112 ^{+/+}	0.295 ^{+/+}
Baseline (Pfam domains)	0.136 ^{+/+}	0.022 ^{-/-}	0.088 ^{+/+}
BLAST (sequence similarity)	0.101 ^{+/+}	0.051 ^{+/+}	0.372 ^{+/+}
HybridNAP	0.098 ^{+/+}	0.076 ^{+/+}	0.477 ^{+/+}
qNABpredict (taxonomy-agnostic)	0.077 ^{+/}	0.033 ^{+/}	0.577 ^{+/}
qNABpredict (taxonomy-aware)	0.074	0.032	0.602

Note: We report average values over 100 random samplings of 50% of proteins from the test dataset. The sampling evaluates robustness of the differences between predictions of various methods that are measured over substantially different protein sets. Predictors are sorted by their overall MAE values. We quantify statistical significance of the differences over the sampled tests, which is shown next to the values using the x/y format where $x = \{+, -, =\}$ denotes that the taxonomy-aware qNABpredict is statistically better (+), worse (-) and not different (=) than the result produced by a given predictor, respectively, at p -value = 0.01; y uses the same nomenclature while comparing against the taxonomy-agnostic version of qNABpredict.

Abbreviations: MAE, mean absolute error; NA, nucleic acid; SCC, Spearman's rank correlation coefficient.

using the low-similarity test dataset. Given the relatively large size of the test dataset (600 proteins), we select methods that are relatively runtime efficient (<5 min to predict an average size protein) and available to the end users as either webserver or source code, which include BindN+ (Wang et al., 2010), DRNAPred (Yan & Kurgan, 2017), and HybridNAP (Zhang et al., 2019b). We convert their residue-level predictions into the content of NA-binding residues by dividing the number of predicted DNA and RNA binding residues by the length of the corresponding protein sequence. We also include three baselines. The first, which we name “random_correct_range”, predicts a random value in the [0, 1] range, which is the actual range of the possible content values. The second, which we call “random_resampled_content”, uses a randomly selected content value from the design dataset. The latter baseline produces the same distribution of the content values as the distribution in the design dataset. The third baseline relies on domain annotations from InterPro (Paysan-Lafosse et al., 2022), which incorporates data from multiple sources including Pfam (El-Gebali et al., 2019). We collected 4324 domains for the test proteins and used a subset of 436 domains that are annotated as DNA, RNA, nucleotide and dinucleotide-binding (including by mapping to these domain annotations as ancestors) to make predictions. We use residues from the regions that correspond to these 436 domains to compute the content values. Moreover, we compute predictions based on sequence similarity. Using the popular BLAST program (Altschul et al., 1997), we identify the most similar protein from the design dataset (based on the lowest e -value) and use its content as the prediction. We quantify the predictive performance by computing the mean

absolute error (MAE) and Spearman's rank correlation coefficient (SCC) values; Section 4.2 defines these error- and correlation-based metrics. We also compute MAE value for the non-NA-binding proteins in the test dataset, to evaluate the extend of the over-prediction of the content values for the proteins that do not interact with the nucleic acids. Table 1 reports average values over 100 random samplings of 50% of proteins from the test dataset. This allows us to evaluate robustness of the differences between predictions of various methods that are measured over substantially different protein sets.

The qNABpredict models specific to the taxonomic kingdoms obtain MAE = 0.074 and SCC = 0.60, which implies that these predictions are accurate and highly correlated with the native content. Moreover, these models only modestly overpredict the NA-binding content for the proteins that do not bind NAs, with the corresponding MAE = 0.032. We find that these taxonomy-aware qNABpredict models statistically outperform the taxonomy-agnostic qNABpredict model (p -value < 0.01), which produces higher MAE = 0.077 and lower SCC = 0.58. This agrees with the observations on the design dataset (Figure 1). However, the latter model can be used for any input protein sequence, even if its taxonomic classification is unknown.

Both versions of the qNABpredict are statistically more accurate than the other considered alternatives including the three residue-level predictors of the NA-binding residues, the BLAST-based prediction and the random baselines (p -value < 0.01). The InterPro-based baseline generates statistically worse MAE = 0.136 and SCC = 0.09 (p -value < 0.01) while producing statistically better MAE for the non-NA-binding-proteins (p -

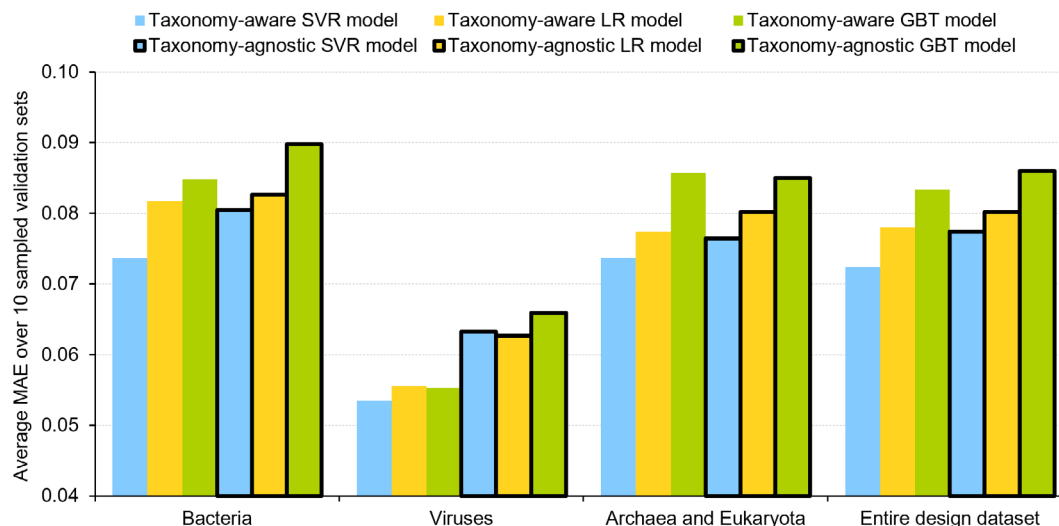


FIGURE 1 Comparison of the average MAE values computed over the 10 sampled validation sets for the four models and the three regressor types on the data from bacteria, viruses, combined archaea and eukaryota, and all kingdoms together. The four models include the taxonomy-agnostic model and the taxonomy-aware models optimized for bacteria, viruses and combined archaea and eukaryota. The three regressor types are linear regression (LR), gradient boosted regression trees (GBT), and support vector regression (SVR).

value < 0.01), suggesting that sequence-based prediction provides a more accurate option to identify NA-binding interfaces. The best alternative, HybridNAP (Zhang et al., 2019b), secures MAE = 0.098 and SCC = 0.48. As expected, predictions of the random baselines lack correlation with the native content values (SCC < 0.06) and their MAEs estimate an expected ceiling of the error values, which is at around 0.160. This means that qNABpredict's results reduce the baseline error by $(0.160 - 0.074)/0.160 = 54\%$. Furthermore, qNABpredict significantly outperforms BLAST alignment that produces only modestly accurate results, with MAE = 0.101 and SCC = 0.37 (p -value < 0.01). This is due to the low similarity between the design and test proteins and indicates that qNABpredict can be used to make accurate predictions for protein sequences irrespective of their similarity to the known NA-binding proteins.

2.2 | Comparative assessment of predictions of NA-binding proteins

The NA-binding content predictions can be used to identify NA-binding proteins. We classify a given sequence as NA-binding if the putative content predicted by qNABpredict is higher than a threshold. We identify a suitable threshold using predictions on the design dataset, as the average of the median content values predicted for the NA-binding and the non-NA-binding proteins, excluding extreme/outlier predictions (i.e., top 5% of the highest and bottom 5% of the lowest predictions). The medians

for the NA-binding and the non-NA-binding proteins are 0.14 and 0.02, respectively, which results in the threshold = 0.08. We test performance of this prediction and compare it to the InterPro-based baseline and results produced by the newest predictor of the NA-binding proteins, iDRBP-ECHF (Feng et al., 2022). This method utilizes a sophisticated ensemble model that combines two deep neural networks, random forest and extremely randomized trees, and was shown to outperform a comprehensive collection of earlier methods (Feng et al., 2022), such as iDRBP-EL (Wang et al., 2021), iDRBP_MMC (Zhang et al., 2020), AIRBP (Mishra et al., 2021), TriPepSVM (Bressin et al., 2019), DeepRBPPred (Zheng et al., 2018), RBPPred (Zhang & Liu, 2017), RNAPred (Kumar et al., 2011), StackDPPred (Mishra et al., 2019) and DNABinder (Kumar et al., 2007). We ensure that the test proteins that we use for this assessment share low, <25% similarity to the design/training datasets of both qNABpredict and iDRBP-ECHF. The test dataset is already dissimilar to the design dataset of qNABpredict and so we apply procedure explained in Section 4.1 to remove sequences similar with the training dataset of iDRBP-ECHF. More specifically, we combine the test proteins with the training dataset of iDRBP-ECHF, cluster the combined dataset using BLASTCLUST (default parameters except for $-S$ 25) and remove all proteins that are in clusters with the training proteins. The resulting dataset includes 413 proteins. The baseline utilizes domain annotations from InterPro (Paysan-Lafosse et al., 2022) described in Section 2.1, where a given sequence is predicted as NA-binding if it

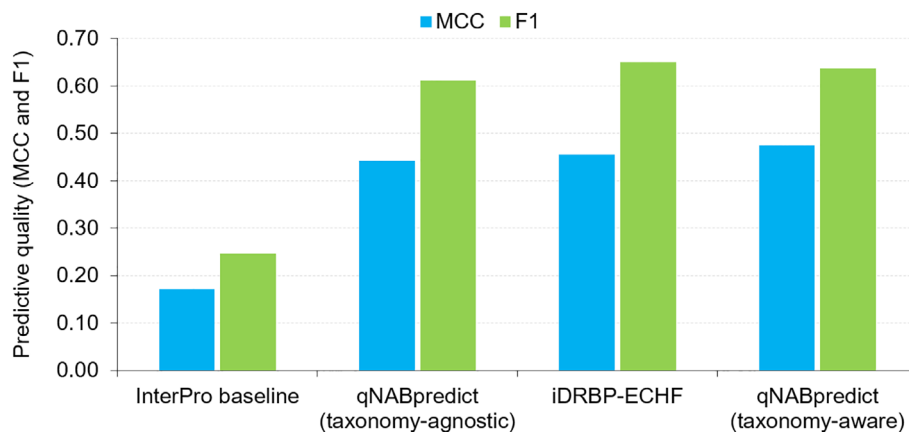


FIGURE 2 Comparison of quality of protein-level predictions of the NA-binding proteins on the subset of 413 proteins from the test dataset that have low, below 25% sequence similarity to the training datasets of qNABpredict and iDRBP-ECHF. iDRBP-ECHF is the newest sequence-based predictor of the NA-binding proteins. qNABpredict predicts a given sequence as NA-binding when its content prediction >0.081 , which corresponds to the average of the median content values predicted for NA-binding and non-NA-binding proteins in the design dataset. The baseline is based on inclusion of NA-binding domains collected from InterPro.

includes at least one domain that is DNA, RNA, nucleotide, and dinucleotide-binding. Figure 2 summarizes the results. The iDRBP-ECHF method is slightly more accurate (Matthews correlation coefficient (MCC) = 0.455) than the taxonomy-agnostic version of qNABpredict (MCC = 0.442), while being slightly worse than the taxonomy aware qNABpredict (MCC = 0.474). These results are substantially better than a baseline, suggesting that sequence-based predictions provide useful clues to identify NA-binding proteins. They also confirm that taxonomy-aware predictions from qNABpredict are more accurate than the taxonomy-agnostic results. Altogether, we find that qNABpredict offers similar levels of predictive quality in identifying NA-binding proteins when compared to modern protein-level predictors, while it provides additional information in the form of the content of NA-binding residues.

2.3 | Application to improve predictions of NA-binding residues

We assess possibility of using qNABpredict's predictions to improve quality of predictions of the NA-binding residues. We consider the three runtime-efficient and available to the end users predictors of the NA-binding residues: BindN+ (Wang et al., 2010), DRNAPred (Yan & Kurgan, 2017), and HybridNAP (Zhang et al., 2019b). We compare quality of their original binary predictions versus binary predictions that are adjusted to match the disorder content predicted by qNABpredict. In other words, we use the real-valued propensities that are output by these methods to predict the number of residues with the

highest scores that matches the putative disorder content output by qNABpredict for the same test protein. Figure 3 compares results on the test dataset. We find that the original predictions are relatively accurate, with MCC values ranging between 0.151 (for HybridNAP) and 0.260 (for BindN+); orange bars in Figure 3. These values are in agreement with previously published assessments (Yan & Kurgan, 2017; Zhang et al., 2019b). Interestingly, using the taxonomy-agnostic version of qNABpredict to adjust these predictions results in a substantial increase in accuracy, with MCC = 0.219 for predictions from HybridNAP and MCC = 0.289 for BindN+; blue bars in Figure 3. Moreover, the taxonomy-aware qNABpredict provides content values that generate further and more modest levels of improvements, MCC = 0.226 for HybridNAP and MCC = 0.293 for BindN+; green bars in Figure 3. In short, this experiment demonstrates that qNABpredict can be used to adjust results produced by predictors of the NA-binding residues, providing a more accurate identification of the location of binding interfaces.

2.4 | Evaluation of runtime

We measure and compare runtime of the qNABpredict with the three other predictors, DRNAPred, hybridNAP and BindN+ using the 600 test proteins and the same computer system, Linux OS (Ubuntu v14.04.5) with 48 64-bit Intel processors and 128 GM RAM. To accommodate for potentially variability of the background workload, we measure the runtime three times for each predictor and record average of the three replicates,

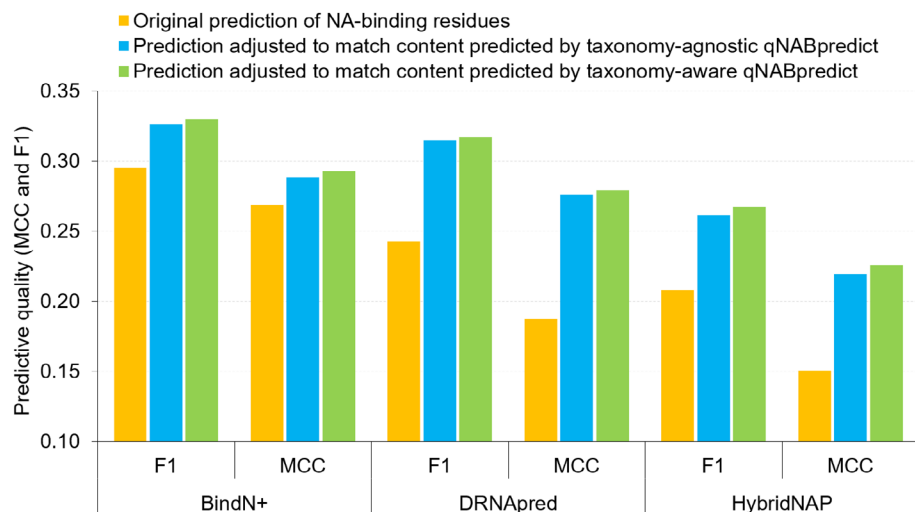


FIGURE 3 Comparison of quality of prediction of NA-binding residues on the low-similarity test dataset. The original results produced by the three predictors are compared to their predictions that are adjusted to match the content predicted by qNABpredict.

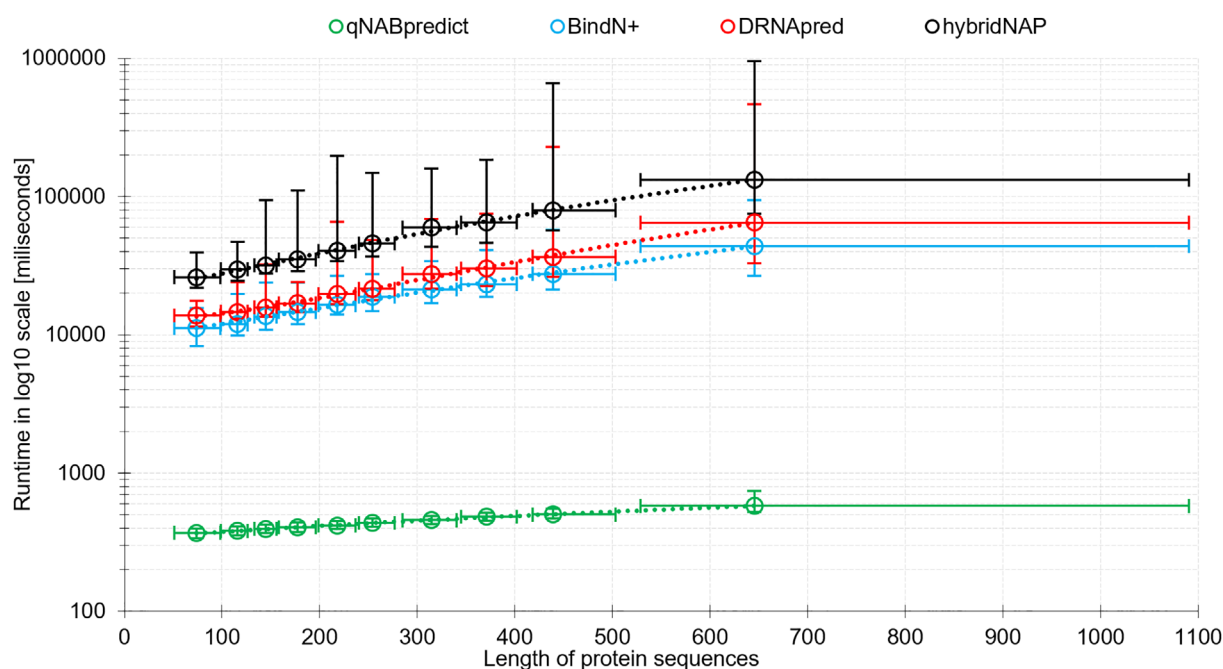


FIGURE 4 Comparison of runtime of qNABpredict (in green) with the three residue-level predictors of NA-binding residues (DRNAPred in red; HybridNAP in black; BindN+ in blue) measured on proteins from the test dataset. We perform the measurements in triplicate using the same computer system. We sort the test sequences by their length and divide them into 10 equally-sized sets of 60 proteins, resulting in sets of progressively larger proteins. The y-axis reports median execution time for each protein set in milliseconds using base 10 logarithmic scale. The x-axis shows the corresponding median sequence length (x-axis). The error bars along both axes denote the 5th and 95th percentiles of the values in a given protein set. The dotted lines show quadratic functions that we fit into the experimental values.

allowing for a break between each experiment. We also study effect of the sequence length on the runtime. We sort the test sequences by their length and divide them into 10 equally-sized sets of 60 proteins, resulting in sets of progressively larger proteins. Figure 4 plots the relation between the median execution time for each protein set measured in milliseconds (y-axis in base 10 logarithmic scale) against the median sequence length (x-axis).

We observe that the runtime of qNABpredict is substantially lower than the runtime of the other tools, approximately by two orders of magnitude. This can be explained by the fact that the other methods perform predictions for each residue in the input sequence and they utilize computationally costly multiple-sequence alignments while qNABpredict generates a single content value and uses an alignment-free (single-sequence)

Please follow steps below to make the predictions:

1. Select a predictor based on sequence information:

Select the appropriate kingdom of the protein, if known:

Bacteria Archaea Eukaryota Viruses

OR

Select, if source of the protein is not known

Unknown

2. Upload a file with protein sequences, or paste them into text area

Please enter [FASTA formatted](#) protein sequence(s) (see [Help](#) section for details of input format).

Each protein length should have minimum 20 and maximum 30,000 residues. For batch file submission, number of sequences in each batch file should not exceed 2000. For larger submission please contact the authors.

No file chosen

3. Provide your email address (optional)

Please enter your email address in the following text area. A link to results of assessment will be sent to your email address once they are ready. The results will be also available in the browser window.

4. Predict

Click the Run button to launch qNABpredict.

FIGURE 5 Screenshot of the user interface of the qNABpredict webserver

ASAquick predictor. We also find that the increase in the runtime for longer protein sequence is relatively small for qNABpredict, that is, the median runtime is 1.6 times larger when comparing the longest against the shortest proteins while their length is 8.7 time larger. To compare, the runtimes of BindN+, DRNAPred, and HybridNAP grow by a factor of 3.9, 4.7, and 5.1, respectively.

The relation between the measured runtime values and the sequence length can be accurately approximated

with a quadratic function; we illustrate this fit with the dotted and color-coded lines in Figure 4. We use these functions to approximate and compare the runtime needed to make predictions for the complete set of 79,740 proteins from the reference human proteome from UniProt (proteome UP000005640) (UniProt, 2021). We estimate that qNABpredict would take approximately 10.3 h to complete these predictions compared to 799 h (33 days) for BindN+, 1315 h for DRNAPred (55 days)

and 2452 h (102 days) for HybridNAP. This demonstrates a clear advantage of using fast content predictors, such as qNABpredict, in the context of large, proteome-scale studies.

2.5 | Webserver and standalone code

Motivated by the relatively high levels of predictive performance and low computational cost of qNABpredict, we implemented its taxonomy-aware and taxonomy-agnostic versions as a convenient and free for academic use webserver at <http://biomine.cs.vcu.edu/servers/qNABpredict/>. Users need to follow a simple four-step process to setup and run the prediction, see Figure 5. First, users should select the taxonomic kingdoms of the protein set they want to predict or leave it as unknown. The former selection leads to using the corresponding taxonomy-aware model while the latter applies the taxonomy-agnostic prediction. Second, users should copy their FASTA-formatted proteins into the input field or upload them as a text file. We allow for batch jobs, with up to 2000 proteins in a single run, to facilitate large-scale analyses. We limit the dataset size to ensure that access to the underlying server infrastructure, which hosts a variety of other predictive server, is balanced between users. Third, users can optionally provide email address where we send a notification and link to the results, once they become available. The results are also produced in the browser window. Lastly, users need to click “Run qNABpredict” to start the prediction. The entire prediction process is automated and executed on the server side. Users do not need to install any software. The webserver outputs the putative content values in an easily parsable text file. We store results on the server for at least 3 months. The files with results are accessible via a direct link that we provide in the return email and on the web page, upon completion of the prediction.

We also offer standalone code for users who would like to run qNABpredict on their local hardware and potentially incorporate our tool into broader bioinformatics platforms. This option also facilitates large-scale predictions, beyond the 2000 proteins submission limit of the webserver. We distribute code as a convenient container that includes all necessary applications, scripts and data. The code and the installation instructions are available via the webserver website at <http://biomine.cs.vcu.edu/servers/qNABpredict/>.

3 | SUMMARY

The current protein sequence-based predictors of NA-binding include protein-level methods (Du & Hu, 2022;

Wang et al., 2021; Zhang et al., 2019a, 2020) and residue-level methods (Miao & Westhof, 2015; Qiu et al., 2020; Wang & Brown, 2006; Wang et al., 2010; Yan & Kurgan, 2017; Yan et al., 2016; Zhang et al., 2019b). The residue-level tools produce more details but suffer much higher computational cost since they rely on computationally expensive sequence alignment. We consider an alternative prediction that produces content of the NA-binding residues, offering more information than the protein-level prediction and substantially shorter runtime compared to the residue-level tools. We note that there are many content predictors for secondary structure (Chen et al., 2011; Homaeian et al., 2007; Horne, 1988; Krigbaum & Knutton, 1973; Lee et al., 2006; Lin & Pan, 2001; Liu & Chou, 1999; Mizianty et al., 2011; Ruan et al., 2005; Yu et al., 2022) and intrinsic disorder (Yan et al., 2013).

Our first-of-its-kind NA-binding content predictor, qNABpredict, relies on two innovations: (1) small, rationally designed and fast-to-compute feature set that represents relevant characteristics extracted from the input sequence; and (2) two versions that include the taxonomy-agnostic model that can be used for proteins of unknown taxonomic origin and the more accurate taxonomy-aware models that are tailored for proteins from specific taxonomic kingdoms: bacteria, viruses, and combined eukaryota and archaea. Using low-similarity dataset of test proteins, we demonstrate that qNABpredict generates statistically more accurate content predictions than the content computed from the residue-level predictors at a small fraction of the computational cost. More specifically, qNABpredict is about 100 times faster and capable of generating predictions for full proteomes in a matter of hours. Moreover, we find that qNABpredict's predictions can be used to improve binary predictions produced by residue-level predictors of NA-binding proteins, and match predictive quality of modern protein-level predictors of NA-binding proteins. We make both versions of qNABpredict available as a convenient webserver and standalone code at <http://biomine.cs.vcu.edu/servers/qNABpredict/>. The taxonomy-aware version requires the user to specify the taxonomic kingdom of the submitted sequences and on average produces more accurate results when compared to the taxonomy-agnostic version. This new tool should be particularly useful when analyzing protein–NA binding for large protein families and proteomes.

4 | MATERIALS AND METHODS

4.1 | Datasets

We follow the data collection protocols from a recent study that investigates DNA, RNA, and protein binding

residues in protein sequences (Zhang et al., 2019b). This protocol ensures a more complete annotation of the binding residues by mapping data across multiple protein–NA complexes that involve the same protein. We utilize the BioLip resource (Yang et al., 2013) to collect annotations of the NA-binding residues. BioLip provides high-quality semi-manually curated annotations of amino acids that bind a wide variety of ligands and which are extracted from the PDB structures. We map the chains collected from PDB using BioLip into full protein sequences from UniProt (2021) using SIFTS (Dana et al., 2019). Next, we delete UniProt IDs that correspond to protein fragments and combine annotations of binding residues from all PDB structures that are mapped to the same protein (UniProt ID). Consequently, we obtain 23,458 proteins that are annotated with binding residues, which include 817 DNA-binding and 1040 RNA-binding proteins. Next, we remove proteins for which the coverage by the PDB structures is below 80% to ensure that the experimental data is sufficiently complete to accurately compute the native content of the NA-binding residues. This results in 1066 well-annotated NA-binding proteins that include 152 proteins from archaea, 482 from bacteria, 362 from eukaryota and 70 viral proteins. Next, we randomly sample the same number of proteins from each of the four taxonomic kingdoms that are annotated to bind ligands that exclude NAs. This protein set constitutes our “negative” data where the content of NA-binding residues is set to 0.

We use the corresponding collection of 2132 proteins to establish design and test datasets. We cluster the 2132 proteins using BLASTCLUST (default parameters except for $-S$ 25) and place entire clusters into either design or test datasets. We include 1532 proteins into the design dataset and the remaining 600 proteins into the test dataset. We use the design dataset to train and optimize the predictive model. We set aside the test dataset during the model design process and use it exclusively to comparatively evaluate the already optimized model. The clustering ensures that the proteins in the design and test dataset share low, $<25\%$ sequence similarity, which means that sequence alignment should not be able to perform accurate predictions on the test proteins. The two datasets, including the annotations of the NA-binding residues, are available at <http://biomine.cs.vcu.edu/servers/qNABpredict/>.

4.2 | Assessment setup

Our predictor takes the amino acid sequence as a sole input and generates a real value in the $[0, 1]$ interval that quantifies the putative content of the NA-binding

residues in that sequence. We evaluate the content predictions by comparing the putative content values against the corresponding native values using the MAE and SCC. Given a set of n proteins with the native content values a_1, a_2, \dots, a_n and the predicted content values x_1, x_2, \dots, x_n and the difference in ranks of the i th value d_i , the MAE and SCC are defined as:

$$MAE = \frac{1}{n} \sum_{i=1}^n |x_i - a_i| \text{ and } SCC = 1 - \frac{6 \sum_i d_i^2}{n(n^2 - 1)}.$$

We also evaluate whether qNABpredict’s results can be used to make accurate predictions of NA-binding proteins and to improve residue-level predictions generated by predictors of NA-binding residues. For both scenarios, we evaluate binary predictions (NA-binding vs. non-NA-binding proteins/residues) using two popular metrics:

$$F1(\text{harmonic mean of precision and sensitivity}) = \frac{2 * TP}{2 * TP + FP + FN},$$

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP) * (TP + FN) * (TN + FP) * (TN + FN)}},$$

where TP and TN denote numbers of correctly predicted NA-binding proteins/residues and non-NA-binding proteins/residues, respectively; FP is the number of the non-NA-binding proteins/residues incorrectly predicted as the NA-binding proteins/residues; and FN is the number of the NA-binding proteins/residues incorrectly predicted as the non-NA-binding proteins/residues.

Given the limited number of the NA-binding proteins in our dataset, we perform sampling to evaluate whether the predictive models are robust and produce accurate results across a diverse collection of protein sets. More specifically, we sample proteins in the design dataset at random and without replacement 10 times, each time selecting two disjoint sets of 600 proteins, $train_j$ and $valid_j$, where $j = 1, 2, \dots, 10$. We use $train_j$ to train the predictive model and $valid_j$ to estimate predictive performance of this model.

4.3 | Distributions of the NA-binding content values across taxonomic kingdoms

We investigate a hypothesis that taxonomical kingdoms have different distributions of the NA-binding content. If true then this would motivate the development of kingdom-specific predictors. We plot the cumulative

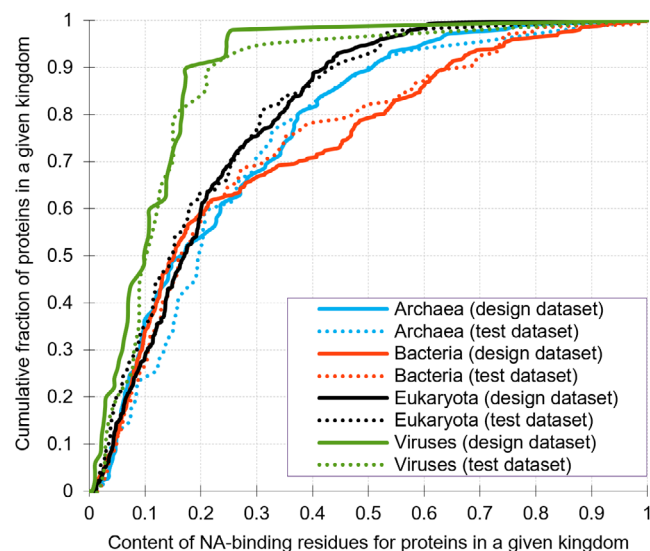


FIGURE 6 Cumulative distribution functions of the content of NA-binding residues for the proteins in the four taxonomic kingdoms: archaea, eukaryota, bacteria, and viruses

distributions functions of the content for the archaea, bacteria, eukaryota and viruses in the design and the test datasets in Figure 6. We find that the distributions for a given kingdom are consistent across the two datasets while being substantially different between some of the kingdoms. The fact that proteins in the two datasets share low <25% similarity suggests that these trends are robust. The viral and bacterial proteins are characterized by the lowest and the highest content values, respectively, while eukaryotic and archaeal proteins are in the middle. We analyze statistical significance of the differences in the distributions of the NA-binding content values for each of the six pairs of kingdoms using the Kolmogorov-Smirnov test. The difference between eukaryota and archaea is not statistically significant (p -value = 0.33), whereas the differences for all other pairs are statistically significant (p -values <0.05). Consequently, we identify three distinct types of distributions, for the viruses, the bacteria, and the combined set of the archaeal and eukaryotic proteins. Correspondingly, we examine two types of designs, a single taxonomy-agnostic model and a combination of three models that target predictions for the viral proteins, the bacterial proteins and the archaeal and eukaryotic proteins. The latter option should generate better predictive performance since these models can be better optimized into the underlying distributions. However, this option requires knowledge of the taxonomic kingdom of the predicted proteins, while the former option allows predicting protein sequence without knowing their taxonomic classification.

4.4 | Architecture of the predictor

Prediction of the content of the NA-binding residues is a two-step process. First, an input protein sequence must be converted into a fixed-size vector of numerical features. Second, these features are processed by a predictive model that generates the putative content values. We follow two key design principles: (1) the model should produce accurate estimates of the content; and (2) it must be very fast. The former requires a careful selection and empirical parametrization of the predictive models while the latter can be accomplished by designing and using a small and fast-to-compute feature space. Moreover, we consider two version of the predictor, the taxonomy-agnostic model versus the combination of the three kingdom-specific models. The design process relies exclusively on the design dataset and follows the robust statistical sampling approach explained in Section 4.2.

4.5 | Formulation and selection of features

We craft a small collection of features that quantify two key characteristics of the NA-binding residues. These residues are located on the protein surface and they are enriched in amino acid types that facilitate the interaction. For instance, the NA-binding residues typically include His, Arg, and Lys, given the presence of the ionic interactions between these positively charged amino acids and the phosphate group of DNA and RNA (Ellis et al., 2007; Lejeune et al., 2005; Li et al., 2014). Moreover, they are also typically enriched in the aromatics, such as Phe, Trp, and Tyr, which is due to the π - π stacking interactions in complexes with NAs (Duh et al., 2015; Wilson et al., 2014). Consequently, we utilize two indices that quantify the propensities of amino acids for the interactions with DNA and RNA that were developed in a recent study (Zhang et al., 2019b). These indices can be used to quantify statistically significant enrichment of positively charged and aromatic amino acids (Arg, His, Lys, Phe, Trp, and Tyr) and statistically significant depletion of negatively charged amino acids (Asp and Glu) in the NA-binding interface (Table 2). We set the propensities for the remaining residues to zero. We identify putative surface residues using a very fast ASAquick method that makes accurate predictions of the relative solvent accessible surface area (rASA) from protein sequences without using multiple sequence alignment (Faraggi et al., 2014, 2017).

We multiply the propensity values by the corresponding putative rASA values to produce scores that quantify likelihood for binding NAs for the amino acid on the

TABLE 2 Propensity of amino acids for interaction with DNA and RNA

Amino acid	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
DNA-binding propensity	0	1.40	0	-0.63	0	0	-0.70	0	0.52	0	0	0.49	0.40	0.51	-0.46	0	0	1.14	0.92	0
RNA-binding propensity	0	1.33	0	-0.62	0	0	-0.64	0	0.54	0	0	0.51	0.55	0	0	0	0	0	0	0

Note: Positive/negative values indicate the enrichment/depletion of a given amino acid type among the NA-binding residues. Abbreviation: NA, nucleic acid.

putative protein surface, that is, higher predicted rASA suggests higher chance for being on the protein surface. We separately apply the DNA-binding and the RNA-binding propensities for enrichment (i.e., positive values) and depletion (i.e., negative values that quantify propensity for exclusion of NA-binding), which results in four sets of the product-based scores. Moreover, we apply a commonly used sliding window-based averaging to smooth out these scores and convert the amino-acid level values into the protein-level aggregates. We enumerate, describe and name the resulting 10 features in Appendix S1. Given the high similarity between the DNA-binding and the RNA-binding propensities (Table 2), we perform feature selection that eliminates highly correlated/redundant features. This is crucial in order to effectively optimize regression-based models that we describe in the next sub-section. We compute Pearson correlation coefficient (PCC) between all pairs of features using the design dataset. Next, we eliminate one feature from each pair for which $PCC > 0.9$, which suggests that they are highly correlated. As a result, we eliminate three redundant features, resulting in a final feature set that includes $rASA_{med}$, $rASA_{max}$, $medianDNA_{positive}$, $maxDNA_{positive}$, $maxDNA_{negative}$, $medianDNA_{negative}$, and $medianRNA_{negative}$ features.

4.6 | Selection and optimization of predictive models

Given that the predictive output is a real-valued content, we consider a broad range of regressors including a simple linear regression (LR) and more sophisticated gradient boosted regression trees (GBT) (Friedman, 2001) and support vector regression (SVR) (Smola & Scholkopf, 2004); deep neural networks are not suitable given the relatively limited size of the design dataset. These models are widely used in related predictions problems. LR was used to predict secondary structure content (Chen et al., 2011; Homaeian et al., 2007; Lin & Pan, 2001; Liu & Chou, 1999), relative solvent accessibility (Qin et al., 2005; Wagner et al., 2005; Wang et al., 2005), B-factor (Zhang et al., 2009), and folding rates (Gao et al., 2010; Jiang et al., 2009). SVR was applied to produce putative intrinsic disorder content (Yan et al., 2013), secondary structure content (Lee et al., 2006), relative solvent accessibility (Chang et al., 2008), isoelectric point and pKa dissociation constants (Kozlowski, 2021), and folding rates (Cheng et al., 2013). GBT was recently applied to predict relative solvent accessibility (Fan et al., 2016) and protein–RNA binding affinity (Deng et al., 2019).

We parametrize the SVR and GBT models by performing a grid search that minimizes average MAE over the 10 training/validation results that are based on statistical sampling approach explained in Section 4.2. LR does not require parametrization. For GBT we consider a wide range of depth values = {1,2,3,4} (i.e., we did not consider deeper tree since we utilize seven features) and numbers of trees = {1,2,3,4,5,6,7,8}. For SVR, we use the popular and robust Radial Basis Function (RBF) kernel and parametrize values of gamma (spread of the RBF function) and C by using 10^x values where $x \in [-2, 2]$ for C and $[-4, 0]$ for gamma; gamma values are small since values of the features are also relatively small. We repeat this parametrization for the four considered models: the taxonomy-agnostic model and the taxonomy-aware models for the viruses, the bacteria, and the combined set of the archaeal and eukaryotic proteins.

Figure 1 compares predictive performance quantified with the average MAE over the 10 sampled validation sets for the four models on data from bacteria, viruses, combined archaea and eukaryota, and all proteins together. This facilitates investigating whether the taxonomy-aware models in fact surpass the taxonomy-agnostic model and to assess differences between regressors. The results reveal that the taxonomy-aware models (bars without borders) outperform the taxonomy-agnostic model (bars with black borders), across all of the color-coded regressor types. This validates our hypothesis that the development of the taxonomy-aware models should result in improvements. Moreover, the SVR model performs the best for the three taxonomy-aware models, while for the taxonomy-agnostic model SVR produces similar results to LR for viruses while generating the lowest errors for bacteria, archaea, eukaryota and the collection of all proteins. Consequently, our predictor, qNABpredict, relies exclusively on the SVR algorithm that uses the RBF kernel with gamma = 0.0162 and $C = 1.528$ for the bacteria model; with gamma = 0.0001 and $C = 4.548$ for the virus model; with gamma = 0.0695 and $C = 0.083$ for the eukaryota and archaea model; and with gamma = 0.0428 and $C = 0.739$ for the taxonomy-agnostic model. SVR produces the lowest errors likely because it uses a non-linear optimization algorithm (due to the use of the RBF kernel), compared to a linear optimization and a heuristic search that are utilized by LR and GBT, respectively. We also observe that MAEs are smaller for the viral proteins when compared to the other kingdoms, which stems from the fact that these proteins have on average lowest native content values and more narrow range of native content values (Figure 6).

Finally, we assess whether developing four taxonomy-aware models (i.e., creating separate models for archaea and eukaryota instead of combining these taxonomic

kingdoms together) would lead to further improvements. Figure S1 reveals that the combined archaea and eukaryota model produces virtually the same quality of predictions as when using an archaea-specific model for archaea proteins and a eukaryota-specific model for eukaryotic proteins. This is true across the three types of the regressors. In particular, for the selected/best SVR, the results are MAE = 0.086 (archaea-specific model) vs. 0.087 (for archaea and eukaryota model) for the archaea proteins, and MAE = 0.069 (eukaryota-specific model) vs. 0.069 (for archaea and eukaryota model) for the eukaryotic proteins. This experiment justifies our approach to combine these two taxonomic kingdoms together.

AUTHOR CONTRIBUTIONS

Zhonghua Wu: Conceptualization (supporting); data curation (equal); formal analysis (equal); investigation (supporting); methodology (lead); resources (equal); software (supporting); validation (equal); writing – original draft (supporting); writing – review and editing (supporting). **Sushmita Basu:** Data curation (equal); formal analysis (equal); methodology (supporting); software (lead); validation (equal); writing – original draft (supporting). **Xuantai Wu:** Investigation (supporting); methodology (supporting); writing – original draft (supporting). **Lukasz Kurgan:** Conceptualization (lead); formal analysis (equal); funding acquisition (lead); investigation (lead); methodology (supporting); project administration (lead); resources (equal); supervision (lead); writing – original draft (lead); writing – review and editing (lead).

CONFLICT OF INTEREST

Authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

Datasets used in this work are freely available at <http://biomine.cs.vcu.edu/servers/qNABpredict/>

ORCID

Zhonghua Wu  <https://orcid.org/0000-0002-4191-0241>

Xuantai Wu  <https://orcid.org/0000-0002-3694-2011>

Lukasz Kurgan  <https://orcid.org/0000-0002-7749-0314>

REFERENCES

- Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 1997;25:3389–402. PMID: 9254694.
- Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The Protein Data Bank. *Nucleic Acids Res.* 2000;28:235–42. PMID: 10592235.

- Boone M, Ramasamy P, Zuallaert J, Bouwmeester R, Van Moer B, Maddelein D, et al. Massively parallel interrogation of protein fragment secretability using SECRIFY reveals features influencing secretory system transit. *Nat Commun.* 2021;12:6414. PMID: WOS:000714972500026.
- Bressin A, Schulte-Sasse R, Figini D, Urdaneta EC, Beckmann BM, Marsico A. TriPepSVM: de novo prediction of RNA-binding proteins based on short amino acid motifs. *Nucleic Acids Res.* 2019;47:4406–17. PMID: 30923827.
- Burley SK, Bhikadiya C, Bi C, Bittrich S, Chen L, Crichlow GV, et al. RCSB protein data Bank: powerful new tools for exploring 3D structures of biological macromolecules for basic and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering and energy sciences. *Nucleic Acids Res.* 2021;49:D437–51. PMID: 33211854.
- Chang DTH, Huang HY, Syu YT, Wu CP. Real value prediction of protein solvent accessibility using enhanced PSSM features. *BMC Bioinformatics.* 2008;9:S12. PMID: WOS:00026215430012.
- Charoensawan V, Wilson D, Teichmann SA. Genomic repertoires of DNA-binding transcription factors across the tree of life. *Nucleic Acids Res.* 2010;38:7364–77.
- Chen K, Stach W, Homaian L, Kurgan L. iFC(2): an integrated web-server for improved prediction of protein structural class, fold type, and secondary structure content. *Amino Acids.* 2011;40:963–73. PMID: 20730460.
- Cheng X, Xiao X, Wu ZC, Wang P, Lin WZ. Swfoldrate: predicting protein folding rates from amino acid sequence with sliding window method. *Proteins.* 2013;81:140–8. PMID: 22933332.
- Choura M, Rebai A, Hanin M. Proteome-wide analysis of protein disorder in *Triticum aestivum* and *Hordeum vulgare*. *Comput Biol Chem.* 2020;84:107138 PMID: 31767506.
- Dana JM, Gutmanas A, Tyagi N, Qi G, O'Donovan C, Martin M, et al. SIFTS: updated structure integration with function, taxonomy and sequences resource allows 40-fold increase in coverage of structure-based annotations for proteins. *Nucleic Acids Res.* 2019;47:D482–9. PMID: 30445541.
- Deng L, Yang W, Liu H. PredPRBA: prediction of protein-RNA binding affinity using gradient boosted regression trees. *Front Genet.* 2019;10:637 PMID: 31428122.
- Du X, Hu J. Deep multi-label joint learning for RNA and DNA-binding proteins prediction. *IEEE/ACM Trans Comput Biol Bioinform.* 2022; PP. PMID: 35148267. <https://doi.org/10.1109/TCBB.2022.3150280>
- Duh Y, Hsiao YY, Li CL, Huang JC, Yuan HS. Aromatic residues in RNase T stack with nucleobases to guide the sequence-specific recognition and cleavage of nucleic acids. *Protein Sci.* 2015;24:1934–41. PMID: 26362012.
- El-Gebali S, Mistry J, Bateman A, Eddy SR, Luciani A, Potter SC, et al. The Pfam protein families database in 2019. *Nucleic Acids Res.* 2019;47:D427–32. PMID: 30357350.
- Ellis JJ, Broom M, Jones S. Protein–RNA interactions: structural analysis and functional classes. *Proteins.* 2007;66:903–11. PMID: 17186525.
- Emamjomeh A, Choobineh D, Hajieghrari B, MahdiNezhad N, Khodavirdipour A. DNA–protein interaction: identification, prediction and data analysis. *Mol Biol Rep.* 2019;46:3571–96.
- Fan C, Liu D, Huang R, Chen Z, Deng L. PredRSA: a gradient boosted regression trees approach for predicting protein solvent accessibility. *BMC Bioinformatics.* 2016;17(Suppl 1):8 PMID: 26818760.
- Faraggi E, Kouza M, Zhou YQ, Kloczkowski A. Fast and accurate accessible surface area prediction without a sequence profile. *Methods Mol Biol.* 2017;1484:127–36. PMID: WOS:000400734600011.
- Faraggi E, Zhou YQ, Kloczkowski A. Accurate single-sequence prediction of solvent accessible surface area using local and global features. *Proteins.* 2014;82:3170–6. PMID: WOS:00034437830026.
- Feng J, Wang N, Zhang J, Liu B. iDRBP-ECHF: identifying DNA- and RNA-binding proteins based on extensible cubic hybrid framework. *Comput Biol Med.* 2022;149:105940 PMID: 36044786.
- Friedman JH. Greedy function approximation: a gradient boosting machine. *Ann Stat.* 2001;29:1189–232. PMID: WOS:000173361700001.
- Gao J, Zhang T, Zhang H, Shen S, Ruan J, Kurgan L. Accurate prediction of protein folding rates from sequence and sequence-derived residue flexibility and solvent accessibility. *Proteins.* 2010;78:2114–30. PMID: 20455267.
- Gawron D, Ndah E, Gevaert K, Van Damme P. Positional proteomics reveals differences in N-terminal proteoform stability. *Mol Syst Biol.* 2016;12:858 PMID: 26893308.
- Glisovic T, Bachorik JL, Yong J, Dreyfuss G. RNA-binding proteins and post-transcriptional gene regulation. *FEBS Lett.* 2008;582:1977–86.
- Gromiha MM, Nagarajan R. Computational approaches for predicting the binding sites and understanding the recognition mechanism of protein–DNA complexes. *Adv Protein Chem Struct Biol.* 2013;91:65–99. PMID: 23790211.
- Homaian L, Kurgan LA, Ruan J, Cios KJ, Chen K. Prediction of protein secondary structure content for the twilight zone sequences. *Proteins.* 2007;69:486–98. PMID: 17623861.
- Horne DS. Prediction of protein helix content from an autocorrelation analysis of sequence hydrophobicities. *Biopolymers.* 1988;27:451–77. PMID: 3359010.
- Jiang Y, Iglinski P, Kurgan L. Prediction of protein folding rates from primary sequences using hybrid sequence representation. *J Comput Chem.* 2009;30:772–83. PMID: 18752216.
- Kelaini S, Chan C, Cornelius VA, Margariti A. RNA-binding proteins hold key roles in function, dysfunction, and disease. *Biology (Basel).* 2021;10:366. PMID: 33923168.
- Kozłowski LP. IPC 2.0: prediction of isoelectric point and pKa dissociation constants. *Nucleic Acids Res.* 2021;49:W285–92. PMID: 33905510.
- Krigbaum WR, Knutton SP. Prediction of the amount of secondary structure in a globular protein from its amino acid composition. *Proc Natl Acad Sci USA.* 1973;70:2809–13. PMID: 4355367.
- Kumar M, Gromiha MM, Raghava GP. Identification of DNA-binding proteins using support vector machines and evolutionary profiles. *BMC Bioinformatics.* 2007;8:463. PMID: WOS:000252790200001.
- Kumar M, Gromiha MM, Raghava GPS. SVM based prediction of RNA-binding proteins using binding residues and evolutionary information. *J Mol Recognit.* 2011;24:303–13. PMID: 20677174.
- Laverty KU, Jolma A, Pour SE, Zheng H, Ray D, Morris Q, et al. PRIESTESS: interpretable, high-performing models of the sequence and structure preferences of RNA-binding proteins. *Nucleic Acids Res.* 2022;50:e111 PMID: 36018788.

- Lee S, Lee BC, Kim D. Prediction of protein secondary structure content using amino acid composition an evolutionary information. *Proteins-Struct Funct Bioinform.* 2006;62:1107–14. PMID: WOS:000235872700026.
- Lejeune D, Delsaux N, Charlotheaux B, Thomas A, Brasseur R. Protein–nucleic acid recognition: statistical analysis of atomic interactions and influence of DNA structure. *Proteins.* 2005;61:258–71. PMID: 16121397.
- Li SL, Yamashita K, Amada KM, Standley DM. Quantifying sequence and structural features of protein–RNA interactions. *Nucleic Acids Res.* 2014;42:10086–98. PMID: WOS:000343220300051.
- Lin Z, Pan XM. Accurate prediction of protein secondary structural content. *J Protein Chem.* 2001;20:217–20. PMID: 11565901.
- Liu W, Chou KC. Prediction of protein secondary structure content. *Protein Eng.* 1999;12:1041–50. PMID: 10611397.
- Ma H, Wen H, Xue Z, Li G, Zhang Z. RNANetMotif: identifying sequence-structure RNA network motifs in RNA–protein binding sites. *PLoS Comput Biol.* 2022;18:e1010293 PMID: 35819951.
- Malhotra S, Sowdhamini R. Genome-wide survey of DNA-binding proteins in *Arabidopsis thaliana*: analysis of distribution and functions. *Nucleic Acids Res.* 2013;41:7212–9. PMID: 23775796.
- Miao Z, Westhof E. A large-scale assessment of nucleic acids binding site prediction programs. *PLoS Comput Biol.* 2015;11:e1004639.
- Mishra A, Khanal R, Kabir WU, Hoque T. AIRBP: accurate identification of RNA-binding proteins using machine learning techniques. *Artif Intell Med.* 2021;113:102034 PMID: 33685590.
- Mishra A, Pokhrel P, Hoque MT. StackDPPred: a stacking based prediction of DNA-binding protein from sequence. *Bioinformatics.* 2019;35:433–41. PMID: 30032213.
- Mizianty MJ, Zhang T, Xue B, Zhou Y, Dunker AK, Uversky VN, et al. In-silico prediction of disorder content using hybrid sequence representation. *BMC Bioinformatics.* 2011;12:245 PMID: 21682902.
- O'Leary NA, Wright MW, Brister JR, Ciuffo S, Haddad D, McVeigh R, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* 2016;44:D733–45. PMID: 26553804.
- Park B, Im J, Tuvshinjargal N, Lee W, Han K. Sequence-based prediction of protein-binding sites in DNA: comparative study of two SVM models. *Comput Meth Prog Biomed.* 2014;117:158–67. PMID: WOS:000343091400011.
- Paysan-Lafosse T, Blum M, Chuguransky S, Grego T, Pinto BL, Salazar GA, et al. InterPro in 2022. *Nucleic Acids Res.* 2022; PMID: 36350672.
- Puton T, Kozlowski L, Tuszyńska I, Rother K, Bujnicki JM. Computational methods for prediction of protein–RNA interactions. *J Struct Biol.* 2012;179:261–8. PMID: 22019768.
- Qin S, He Y, Pan XM. Predicting protein secondary structure and solvent accessibility with an improved multiple linear regression method. *Proteins.* 2005;61:473–80. PMID: 16152601.
- Qiu J, Bernhofer M, Heinzinger M, Kemper S, Norambuena T, Melo F, et al. ProNA2020 predicts protein–DNA, protein–RNA, and protein–protein binding proteins and residues from sequence. *J Mol Biol.* 2020;432:2428–43. PMID: 32142788.
- Ruan J, Wang K, Yang J, Kurgan LA, Cios K. Highly accurate and consistent method for prediction of helix and strand content from primary protein sequences. *Artif Intell Med.* 2005;35:19–35. PMID: 16081261.
- Si J, Cui J, Cheng J, Wu R. Computational prediction of RNA-binding proteins and binding sites. *Int J Mol Sci.* 2015a;16:26303–17. PMID: 26540053.
- Si J, Zhao R, Wu R. An overview of the prediction of protein DNA-binding sites. *Int J Mol Sci.* 2015b;16:5194–215.
- Skupien-Rabian B, Jankowska U, Swiderska B, Lukasiewicz S, Ryszawy D, Dziedzicka-Wasylewska M, et al. Proteomic and bioinformatic analysis of a nuclear intrinsically disordered proteome. *J Proteomics.* 2016;130:76–84. PMID: 26376097.
- Smola AJ, Scholkopf B. A tutorial on support vector regression. *Stat Comput.* 2004;14:199–222. PMID: WOS:000222770200003.
- Su Y, Luo Y, Zhao X, Liu Y, Peng J. Integrating thermodynamic and sequence contexts improves protein–RNA binding prediction. *PLoS Comput Biol.* 2019;15:e1007283. PMID: 31483777.
- UniProt C. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.* 2021;49:D480–9. PMID: 33237286.
- Wagner M, Adamczak R, Porollo A, Meller J. Linear regression models for solvent accessibility prediction in proteins. *J Comput Biol.* 2005;12:355–69. PMID: 15857247.
- Walia RR, Caragea C, Lewis BA, Towfic F, Terribilini M, El-Manzalawy Y, et al. Protein–RNA interface residue prediction using machine learning: an assessment of the state of the art. *BMC Bioinformatics.* 2012;13:89 PMID: 22574904.
- Wang C, Lv Y, Wang B, Yin C, Lin Y, Pan L. Survey of protein–DNA interactions in *Aspergillus oryzae* on a genomic scale. *Nucleic Acids Res.* 2015;43:4429–46. PMID: 25883143.
- Wang JY, Lee HM, Ahmad S. Prediction and evolutionary information analysis of protein solvent accessibility using multiple linear regression. *Proteins-Struct Funct Bioinform.* 2005;61:481–91. PMID: WOS:000233029800004.
- Wang K, Hu G, Wu Z, Su H, Yang J, Kurgan L. Comprehensive survey and comparative assessment of RNA-binding residue predictions with analysis by RNA type. *Int J Mol Sci.* 2020;21:6879. <https://doi.org/10.3390/ijms21186879>
- Wang L, Huang C, Yang MQ, Yang JY. BindN+ for accurate prediction of DNA and RNA-binding residues from protein sequence features. *BMC Syst Biol.* 2010;4(Suppl 1):S3 PMID: 20522253.
- Wang LJ, Brown SJ. BindN: a web-based tool for efficient prediction of DNA and RNA binding sites in amino acid sequences. *Nucleic Acids Res.* 2006;34:W243–8. PMID: WOS:000245650200051.
- Wang N, Zhang J, Liu B. iDRBP-EL: identifying DNA- and RNA-binding proteins based on hierarchical ensemble learning. *IEEE/ACM Trans Comput Biol Bioinform.* 2021; PP. PMID: 34932484. <https://doi.org/10.1109/TCBB.2021.3136905>
- Warren C, Shechter D. Fly fishing for histones: catch and release by histone chaperone intrinsically disordered regions and acidic stretches. *J Mol Biol.* 2017;429:2401–26. PMID: 28610839.
- Wilson KA, Kellie JL, Wetmore SD. DNA–protein pi-interactions in nature: abundance, structure, composition and strength of contacts between aromatic amino acids and DNA nucleobases or deoxyribose sugar. *Nucleic Acids Res.* 2014;42:6726–41. PMID: WOS:000338768100056.

- Yan J, Friedrich S, Kurgan L. A comprehensive comparative review of sequence-based predictors of DNA- and RNA-binding residues. *Brief Bioinform.* 2016;17:88–105.
- Yan J, Kurgan L. DRNAPred, fast sequence-based method that accurately predicts and discriminates DNA- and RNA-binding residues. *Nucleic Acids Res.* 2017;45:e84 PMID: 28132027.
- Yan J, Mizianty MJ, Filipow PL, Uversky VN, Kurgan L. RAPID: fast and accurate sequence-based prediction of intrinsic disorder content on proteomic scale. *Biochim Biophys Acta.* 2013;1834:1671–80. PMID: 23732563.
- Yang J, Roy A, Zhang Y. BioLiP: a semi-manually curated database for biologically relevant ligand-protein interactions. *Nucleic Acids Res.* 2013;41:D1096–103. PMID: 23087378.
- Yu CH, Chen W, Chiang YH, Guo K, Martin Moldes Z, Kaplan DL, et al. End-to-end deep learning model to predict and design secondary structure content of structural proteins. *ACS Biomater Sci Eng.* 2022;8:1156–65. PMID: 35129957.
- Zhang H, Zhang T, Chen K, Shen S, Ruan J, Kurgan L. On the relation between residue flexibility and local solvent accessibility in proteins. *Proteins.* 2009;76:617–36. PMID: 19274736.
- Zhang J, Chen Q, Liu B. DeepDRBP-2L: a new genome annotation predictor for identifying DNA-binding proteins and RNA-binding proteins using convolutional neural network and long short-term memory. *IEEE/ACM Trans Comput Biol Bioinform.* 2019a;18:1451–63.
- Zhang J, Chen Q, Liu B. iDRBP_MMC: identifying DNA-binding proteins and RNA-binding proteins based on multi-label learning model and motif-based convolutional neural network. *J Mol Biol.* 2020;432:5860–75. PMID: 32920048.
- Zhang J, Ma Z, Kurgan L. Comprehensive review and empirical analysis of hallmarks of DNA-, RNA- and protein-binding residues in protein chains. *Brief Bioinform.* 2019b;20:1250–68. PMID: 29253082.
- Zhang X, Liu S. RBPPred: predicting RNA-binding proteins from sequence using SVM. *Bioinformatics.* 2017;33:854–62. PMID: 27993780.
- Zhang Y, Bao WZ, Cao Y, Cong HH, Chen BT, Chen YH. A survey on protein–DNA-binding sites in computational biology. *Brief Funct Genomics.* 2022;21:357–75. PMID: WOS:000804727900001.
- Zhao H, Yang Y, Zhou Y. Prediction of RNA binding proteins comes of age from low resolution to high resolution. *Mol Biosyst.* 2013;9:2417–25. PMID: 23872922.
- Zheng J, Zhang X, Zhao X, Tong X, Hong X, Xie J, et al. Deep-RBPPred: predicting RNA binding proteins in the proteome scale based on deep learning. *Sci Rep.* 2018;8:15264 PMID: 30323214.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Wu Z, Basu S, Wu X, Kurgan L. qNABpredict: Quick, accurate, and taxonomy-aware sequence-based prediction of content of nucleic acid binding amino acids. *Protein Science.* 2023;32(1):e4544. <https://doi.org/10.1002/pro.4544>