biomolecules

MDPI

Review

Single-Stranded DNA Binding Proteins and Their Identification Using Machine Learning-Based Approaches

Jun-Tao Guo * and Fareeha Malik

Department of Bioinformatics and Genomics, University of North Carolina at Charlotte, Charlotte, NC 28223, USA
* Correspondence: jguo4@uncc.edu

Abstract: Single-stranded DNA (ssDNA) binding proteins (SSBs) are critical in maintaining genome stability by protecting the transient existence of ssDNA from damage during essential biological processes, such as DNA replication and gene transcription. The single-stranded region of telomeres also requires protection by ssDNA binding proteins from being attacked in case it is wrongly recognized as an anomaly. In addition to their critical roles in genome stability and integrity, it has been demonstrated that ssDNA and SSB–ssDNA interactions play critical roles in transcriptional regulation in all three domains of life and viruses. In this review, we present our current knowledge of the structure and function of SSBs and the structural features for SSB binding specificity. We then discuss the machine learning-based approaches that have been developed for the prediction of SSBs from double-stranded DNA (dsDNA) binding proteins (DSBs).

Keywords: single-stranded DNA; ssDNA; single-stranded DNA binding protein; SSB; binding specificity

Citation: Guo, J.-T.; Malik, F. Single-Stranded DNA Binding Proteins and Their Identification Using Machine Learning-Based Approaches. *Biomolecules* **2022**, *12*, 1187. https://doi.org/10.3390/biom12091187

Academic Editors: Prakash Kulkarni, Vladimir N. Uversky, Philippe Urban and Alessandro Paiardini

Received: 18 July 2022
Accepted: 24 August 2022
Published: 26 August 2022

1. Introduction

Since Watson and Crick's Nobel Prize winning discovery in 1953, the canonical representation of genomic DNA conformation has been the double-stranded DNA (dsDNA) helical structure [1]. However, there are many instances that single-stranded DNA (ssDNA) exists either transiently or consistently. For example, dsDNA unwinds to form ssDNA during essential biological processes such as DNA replication, transcription, recombination, and repair [2–6]. Unlike the double-helical structure of dsDNA that is stabilized by base pairing and base stacking, ssDNA is more flexible and less stable, making it vulnerable to chemical or enzymatic attacks [3,7]. The 3′ single-stranded DNA overhang, a key component of telomere structure at the end of eukaryotic chromosomes, is susceptible to "unauthorized" processing or misrecognition as DNA damage [3,8]. The lack of protection of these ssDNA regions can pose serious problems to genomic stability and integrity and may cause diseases [8,9].

The critical importance of maintaining genome stability and the need of protecting vulnerable ssDNA from damages require a special family of proteins, called ssDNA binding proteins (SSBs), which are ubiquitous in all living organisms [2,4]. The first SSB was discovered and characterized in the bacteriophage T4 in 1970 [10]. Shortly after that, the *E. coli* SSB protein was identified [11,12]. *E. coli* SSB, which functions as a homotetramer, is the most widely studied prokaryotic SSB. Replication protein A (RPA), originally identified as a key component for simian virus 40 (SV40) replication, represents the first eukaryotic SSB that was found to be directly involved in DNA metabolism [13–16]. Human RPA is a heterotrimeric complex composed of three subunits, RPA70, RPA32, and RPA14, and is involved in a variety of DNA repair pathways including mismatch repair, double-stranded break repair, and recombination repair [17]. Two other human SSBs, termed hSSB1 and hSSB2, were identified in 2008 [18]. It has been demonstrated that hSSB1 plays important roles in genome stability as well as in cell cycle regulation and transcription [7,18].

Another group of SSBs that are of paramount importance in maintaining genome stability is telomere end-binding protein (TEBP), a sequence-specific ssDNA binding protein. At the very end of eukaryotic chromosomes of the 3′ termini, there exist single-stranded DNA overhangs. These overhangs vary in length in different species, ranging from several to hundreds of bases [5,9]. To safeguard these vulnerable 3′ overhangs from inappropriate processing, TEBP binds and acts as a cap to sequester the ssDNA in a sequence-specific manner [5,8]. Failure of such protection can lead to destabilization of the genome and early onset of cellular senescence [8,19,20].

2. SSB Structure, Binding Specificity and Function

2.1. Structural Folds of ssDNA Binding Domains

There are four main structural folds in ssDNA binding domains: oligonucleotide/oligosaccharide-binding (OB) folds, K homology (KH) domains, RNA recognition motifs (RRMs), and whirly domains [2]. The OB fold is a well-known ssDNA binding domain found in many SSBs, including *E. coli* SSB, bacteriophage T4 SSB, human RPA, human mtSSB, human SSB1 and SSB2, and the telomere-end protection family, such as TEBP from *Oxytricha nova*, Cdc13 from *Saccharomyces cerevisiae*, and Pot1 from humans.

The OB fold consists of a five-stranded antiparallel β-sheet that forms a characteristic β-barrel core with various lengths of loops connecting the strands [2,21]. The narrow ssDNA binding cleft on the surface of the β-barrel is close to strand 2 and strand 3 [2]. ssDNA binds its interacting surface with its bases facing the protein. The number and organization of the OB domains that participate in ssDNA interaction vary quite differently [22]. RPA is a heterotrimer with six OB folds, two for subunit interaction and four for ssDNA binding. *E. coli* SSB is a homotetramer with each unit having one OB domain. The SSB from *Sulfolobus solfataricus* (ssoSSB) has a single OB domain and binds ssDNA with a footprint of five bases and a defined binding polarity [22].

The diversity and versatility of the OB fold can be understood from two different angles. First, OB fold proteins have different ssDNA binding specificity (Figure 1). Even though binding specificity is a relative term and is difficult to define with a hard cutoff, proteins can be grouped based on the differences of binding affinity among a large number of sequences [23]. In both protein-dsDNA interaction and protein-ssDNA interaction, hydrogen bonds (HBs) between DNA bases and protein sidechain atoms are considered the major contributor to binding specificity [23–31]. Some OB fold proteins bind ssDNA with very high sequence specificity. The prime example of sequence-specific OB fold is TEBP proteins that recognize specific sequences at the 3′ ssDNA overhang at the end of chromosomes [2]. The complex structure between human POT1 OB fold domain and telomeric ssDNA (TTAGGGTTAG) reveals more sidechain-base HBs than non-sidechain-base HBs (Figure 1A) [32]. Some other OB fold proteins are sequence independent or non-specific, meaning they can bind different ssDNA sequences. The non-specific ssDNA binding OB folds include the highly conserved eukaryotic RPA and bacteria SSBs [33]. Figure 1B shows the non-specific binding between *Bacillus subtilis* SsbB OB fold and ssDNA, which has a fewer percentage of sidechain-base HBs [34]. The binding specificity also depends on the sequence/structure context. For example, the C-terminal of the ORF6 of *Enterobacter Phage* Enc 34, is important for specific binding to ssDNA. The removal of the C-terminal from Enc 34 makes it less specific as it can bind both ssDNA and dsDNA [35]. The second aspect of the OB fold diversity lies in its lack of sequence conservation. While the sequence diversity of the OB fold provides a variety of interaction surface for recognition of different ssDNA sequences, especially for the sequence-specific OB-fold SSBs, it presents challenges in identification of new SSBs based on sequence information alone.

The KH domains are characterized by three α-helices that are closely packed with a three-stranded β- sheet. ssDNA binds to KH domains with conserved polarity and with bases facing the protein. Examples of ssDNA binding KH domains include hnRNP k, FBP, poly (C)-binding protein (PCBP) 1 and 2 [2]. All of these KH domains bind ssDNA at sites upstream of promoter regions and affect transcription. Although the RRM domains are

abundant in annotated human genomes, there are only five proteins with RRM domains that have known structures complexed with ssDNA, including human UP1, RBM-45, FBP-interacting repressor (FIR), and SUP-12 and MEC-8 from *Caenorhabditis elegans* [36–39]. Compared to other three ssDNA binding domains, whirly domains are relatively large, comprising about 180 amino acids. They have 2 four-stranded β-sheets arranged in almost parallel along with some helical fragments. The whirly domains are mostly found in mitochondria and chloroplasts in plants and perform diverse functions in transcriptional activation, splicing, and DNA repair.

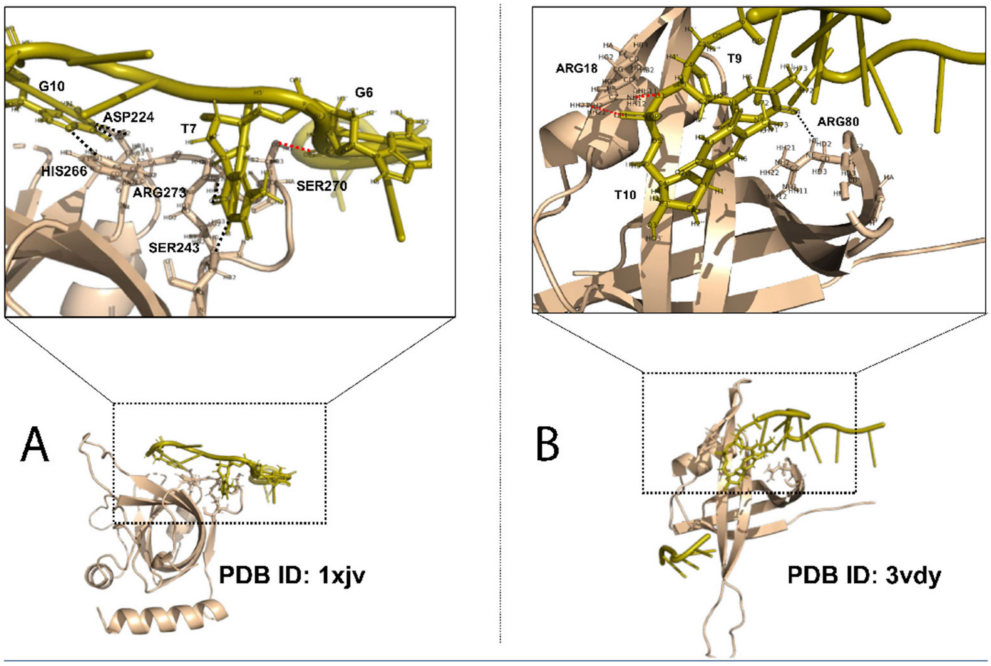

Figure 1. OB fold of SSBs and binding specificity. (**A**) Specific OB fold domain of human POT1 bound to telomeric single-stranded DNA (TTAGGGTTAG), PDB ID: 1xjv, chain: A, domain: 151–299. (**B**) Non-specific OB fold of *Bacillus subtilis* SsbB, PDB ID: 3vdy, chain: A. Black dashed lines, sidechain-base hydrogen bonds; red dashed lines, non-sidechain-base hydrogen bonds.

2.2. Structural Features in SSB Binding Specificity

As described previously, SSBs can bind ssDNA specifically or non-specifically. While there are quite a number of studies on structural features in protein-dsDNA binding specificity [23,24,26,40–44], the specificity investigation of protein-ssDNA interaction is underexplored due to the limited availability of protein-ssDNA complex structures. We recently performed a comparative study of protein–ssDNA interactions in terms of binding specificity [27]. A non-redundant dataset of SSB–ssDNA complexes with high structural quality was generated and classified into two groups: specific and non-specific, based on their binding specificity, which was manually annotated by searching the primary references of these SSB–ssDNA complexes and their homologs in Protein Data Bank (PDB) [45,46] as well as relevant information in UniProt [47]. We then compared the key structural features in protein-ssDNA interaction, including binding propensities and secondary structure types of ssDNA base-interacting residues, hydrogen bonds, π–π interactions between residue side chains and DNA bases, interaction interfaces, and protein conformational changes upon ssDNA binding [27].

Hydrogen bonds and amino acid binding propensities were found to be the key discriminating features between specific and non-specific ssDNA binding proteins [27] (Figure 2 for comparison of sidechain-base hydrogen bonds). As for amino acid binding propensities, while aromatic and positively charged amino acids, phenylalanine, tryptophan, tyrosine, histidine, lysine, and arginine are enriched in both specific and non-specific

protein-ssDNA complexes, three amino acids (histidine, tyrosine, and arginine) are more enriched in the specific group than those in the non-specific group. The positively charged lysine and arginine are capable of forming both hydrogen bonds with DNA bases and ionic interactions with the negatively charged DNA backbone atoms. Aliphatic amino acids alanine, isoleucine, leucine, and valine show low propensity in both binding specificity groups. Interestingly, the negatively charged aspartate is also enriched in the specific protein-ssDNA interactions as we demonstrated in specific protein-dsDNA interactions [23,27]. π–π interactions may primarily contribute to binding affinity as there are no apparent difference of π–π interactions between the two binding specificity groups. No significant differences were found between specific and non-specific groups with respect of conformational changes upon ssDNA binding, suggesting that the flexibility of SSBs plays a lesser role than that of double-stranded DNA-binding proteins in conferring binding specificity. The features that are different between specific and non-specific ssDNA binding proteins can be applied to machine learning algorithms for prediction of binding specificity if the protein–ssDNA complex structures are available.

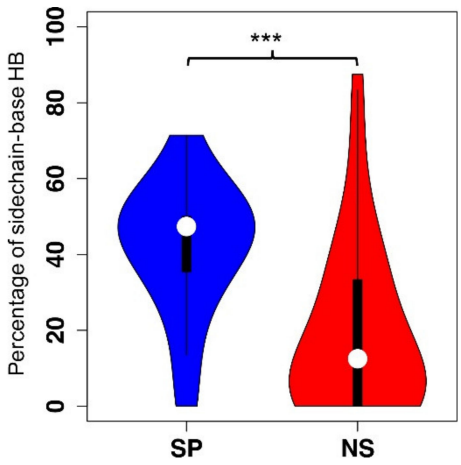

Figure 2. Comparison of the distribution of the percentage of the sidechain-base hydrogen bonds in each protein chain–ssDNA complex between the specific (SP) and the non-specific (NS) protein–ssDNA complexes. Statistical analysis of the comparison between the NS and SP groups was done with Wilcoxon rank sum test. *** = p-value $\leq$ 0.001.

Wang et al. carried out a comparative analysis of structural features that can be used to differentiate SSBs and DSBs [48]. The analysis was based on a dataset of 238 DSBs and 97 SSBs in complex with dsDNA and ssDNA respectively. The features include surface shapes that are grouped into three categories: peak, flat, and valley, and the surrounding environment, such as amino acid composition and electrostatic charge of the interface. They demonstrated that the distributions of the above features are significantly different between DSBs and SSBs, which can be used for structure-based classification between DSBs and SSBs [48].

2.3. Role of ssDNA and SSBs in Transcription Regulation

In addition to their critical roles in genome stability and integrity, it has been demonstrated that ssDNA and SSBs play important roles in transcriptional regulation in all three domains of life and viruses [49–60]. In the traditional model of transcriptional regulation, transcription factors recognize and bind to their cognate DNA binding sites in double helical form to either activate or repress gene expression [24]. However, it has been shown that the presence of single-stranded DNA regions in a number of transcriptionally active promoters is much more than previously thought, which warranted an investigation of their role in the regulation of gene transcription [49]. Attempts to unravel the mechanism of ssDNA in transcriptional regulation not only revealed that they are important for optimal

transcription but also led to the identification of a number of SSBs. For example, c-myc transcription is regulated by the binding of a sequence-specific SSB, FBP (FUSE-binding protein), to FUSE (far upstream element) [50,61]. Heterogeneous nuclear ribonucleoprotein (hnRNP) K binds to single-stranded DNA with a CT element upstream of the c-myc promoter and activates transcription [62]. In mice, sequence-specific SSBs were reported to be involved in the transcriptional regulation of μ-opioid receptor gene and timp-1 gene (54,55). A number of unique ssDNA binding proteins that participate in transcriptional regulation have also been identified in plants [56–59]. Desveaux et al. found a novel ssDNA binding protein that regulates the expression of the PR-10a gene in potato [56]. The protein, called PBF-2 for PR-10a binding factor 2, binds to a 30-bp promoter sequence with high affinity in a sequence-specific manner. PBF-2 consists of four p24 subunits that belong to a novel family of ubiquitous plant-specific whirly family [58]. In Arabidopsis, the Whirly1 (Why1) protein not only binds to transcriptional response elements, it also binds to telomeric DNA and modulates telomere length homeostasis [63].

2.4. SSBs and Diseases

Aberrant expressions and changes in SSBs can result in DNA damage and lead to cancer [17,64]. Conditional deletion of mouse homolog of hSSB1 in mice showed increased susceptibility to several types of tumors [65]. It has been shown that defective regulation and expression of APOBEC3A (apolipoprotein B messenger RNA-editing enzyme, catalytic polypeptide-like), a single-stranded DNA deoxycytidine deaminase, is involved in the development of cancer [66,67]. BRCA2, a breast cancer tumor suppressor, promotes direct interactions between RAD51 recombinase and ssDNA. Loss-of-function mutations in BRCA2 increase susceptibility to breast and ovarian cancers owing to genome instability [68–70]. Moreover, small molecules that inhibit binding activity between ssDNA and RPA can prevent cell cycle progression and enhance chemosensitivity in cancer cells, suggesting a therapeutic value for targeting SSBs [17,71].

3. Machine Learning-Based Methods for SSB Prediction

Despite the critical roles of ssDNA binding proteins in essential biological processes, investigation of protein–ssDNA interactions clearly lags behind other types of protein-nucleic acids interactions, such as protein–dsDNA interactions, due to the limited number of protein–ssDNA complex structures. Another key limiting factor in studying protein–ssDNA interactions is that new SSBs have been identified at a very slow pace. As indicated in a recent call of an open invitation to the Understudied Proteins Initiative [72,73], in the life science field, research efforts focus on a group of well-studied proteins, while the biological function of the majority of proteins are understudied or remain unexplored due to a variety of practical reasons. These proteins with unknown functions are termed "the dark matter of the sequence universe", and it is practically impossible to experimentally characterize each of them from the dark [74–77]. Therefore, it is of particular importance to develop efficient computational methods for predicting new SSBs from the uncharacterized proteins. Not only can it expand the landscape of SSBs, it can also narrow down the cases to a reasonable number for follow-up experimental validations and potentially increase the number of ssDNA–SSB complex structures for structural studies.

Machine learning methods have been widely used in a number of research topics in structural bioinformatics. The application of machine learning methods to bioinformatics problems culminated in the success of the development of AlphaFold for protein structure prediction, a scientific breakthrough that was built upon years of previous efforts [78–81]. Various machine learning methods have also been developed for prediction of nucleic acid-binding or DNA binding proteins from protein sequences [82–94]. Compared to the large number of publications on prediction of DNA binding proteins, the investigation on ssDNA binding protein prediction is limited so far. To our knowledge, currently there are only four published studies on SSB prediction using machine learning-based approaches [95–98]. These methods typically consist of four major steps as shown in Figure 3: (1) dataset

generation for training and testing; (2) features for learning and prediction; (3) classification models; and (4) performance evaluation. We discuss below each of these four steps in machine learning-based methods for SSB prediction.

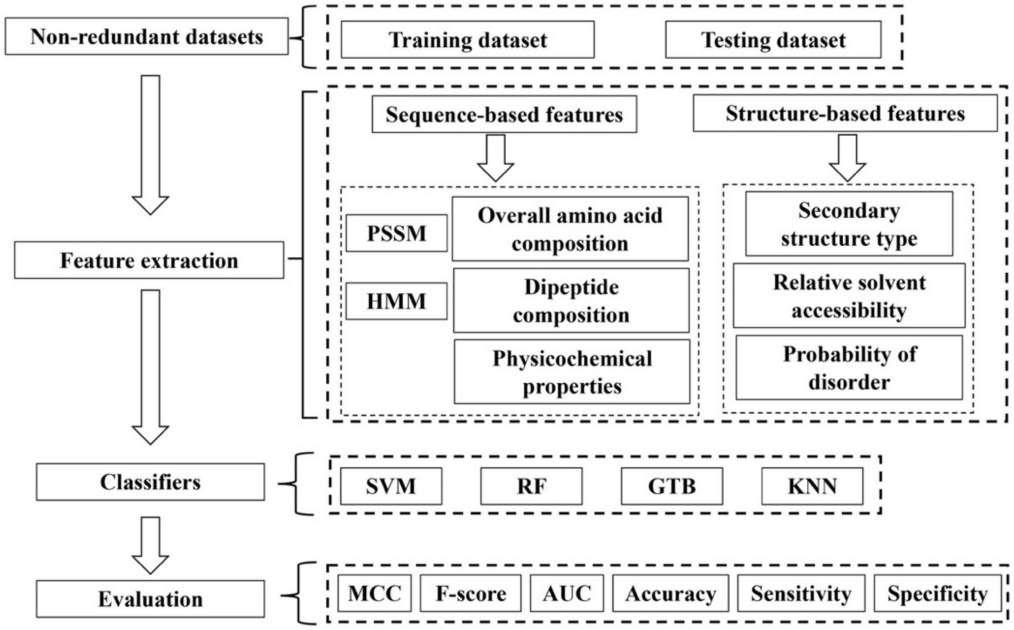

Figure 3. The flowchart for machine learning-based prediction of SSBs.

3.1. Datasets

Prediction performance of machine learning-based approaches, especially when applying the learning models to real-world biological problems, relies on carefully generated, robust data sets for training and testing [94]. So far, the prediction of SSBs has been generally carried out for differentiation between SSBs and dsDNA binding proteins (DSBs). Essentially all the methods use the same training and independent test sets from the work by Wang et al. [95–98]. The training set (Uniprot1065) consists of 873 DSBs and 183 SSBs, which were culled from UniProtKB/SwissProt with a sequence identity cutoff at 0.7. The independent non-redundant dataset of SSBs and DSBs was derived from PDB with 125 DSBs and 41 SSBs, which share less than 30% sequence identity [95]. These protein structures were either solved by X-ray crystallography with a resolution of less than 3 Å or by NMR.

It is understandable that a higher sequence identity cutoff was adopted for the training dataset construction due to limited availability of annotated SSBs. However, it may cause an overfitting problem in machine learning models as evidenced by the unbalanced performance between the training set and the independent dataset, resulting in a big drop of MCC (Matthew's correlation coefficient) from the training set to the independent set, regardless of the type of models or features used for the prediction [95–98]. Another potential issue that may contribute to the unbalanced performance is the phenomenon of protein moonlighting since some DNA binding proteins can bind both ssDNA and dsDNA. For example, in the independent dataset, *Arabidopsis* cryptochrome 3 (2VTBE) and the tumor repressor p53 (2B3GB) are annotated to bind both ssDNA and dsDNA [99–102]. In addition, some proteins may be able to bind both RNA and DNA or they are RNA binding proteins, but due to technical challenges in solving RNA structures, the structures were solved with ssDNA instead of ssRNA.

Two other datasets were mentioned in SSB prediction studies. Sharma et al. compiled an additional independent test set with 64 SSBs and 53 DSBs, as shown in their GitHub site after filtering out cases with more than 70% sequence identity with the above mentioned training set and independent set by Wang et al. [96]. Tests were performed with three sequence identity cutoffs at 0.9, 0.7, and 0.3 for this new independent test set [96]. We

performed a CD-HIT analysis of this new set with a 0.3 sequence identity cutoff and found that there are only 18 SSBs and 14 DSBs, while 13 out of the 47 entries were annotated as double-stranded RNA binding proteins and 2 were considered as ssDNA and dsDNA binding proteins in UniProt [70]. Therefore, the low prediction performance may be a result of this dataset with mixed annotations. Tan et al. mentioned a dataset with 1271 SSBs and 2252 DSBs in their study, which includes the training set from Wang.et al. [98]. However, it is not clear how the proteins in the list were selected and the number of false positive SSBs as the proteins of this dataset are not published.

3.2. Features

Both sequence features and sequence-derived structural features have been applied for SSB and DSB prediction (Table 1 and Figure 3). The widely-used sequence features in protein structure and function prediction, such as PSSM (position specific scoring matrix) and HMM (hidden Markov model) profiles, which consider evolution information and are generated from multiple sequence alignments, have been explored [95,97,98]. PSSM is typically generated with PSI-BLAST through searching a protein database to detect increasingly divergent members of a protein in consecutive iterations [103,104].

Table 1. Summary of machine learning-based SSB prediction methods.

References	Predictor (If Any)	Features	Classifiers
Wang et al. [95]	NA	-OAAC -Dipeptide composition -Physicochemical properties -PSSM	SVM RF
Ali et al. [97]	SDBP-Pred	-PSSM -CS-PSSM -CSS-PSSM2 -CSS-PSSM3	SVM
Tan et al. [98]	PredPSD	-OAAC -Dipeptide composition -Physicochemical properties -PSSM -Structural features from NetSurfP and DisEMBL	GTB
Sharma et al. [96]	NA	-HMM -normalized profile-monogram -normalized profile-bigram	SVM RF KNN

Both the original PSSM or PSSM derivatives have been used for prediction of SSBs and DSBs. For example, a feature descriptor called consensus sequence-based K-segmentation PSSM (CSKS-PSSM) was developed in SDBP-Pred [97]. To generate the consensus sequence, a position in a protein sequence is replaced with a residue with the maximum substitution probability at this position. For K-segmentation PSSM, the PSSM is split into K-segments of equal sizes. Ali et al. implemented two- and three-segmented PSSM in their model [97]. Sharma et al. used HMM profiles produced from HHblits to capture evolutionary information [96]. More specifically, they applied the normalized profile-monogram and normalized profile-bigram feature extraction methods, which have been demonstrated to be useful in protein fold prediction and intrinsically disordered region prediction [96].

Other sequence features that have been explored for SSB and DSB prediction include overall amino acid composition (OAAC), dipeptide composition, and physicochemical properties [95,98]. OAAC is a 20-dimensional descriptor with a frequency value for each of the 20 standard amino acids for a given protein sequence. It has been indicated that

using the square root of the frequencies can result in a better performance than the raw frequencies [95,105]. Dipeptide compositions represent the frequencies of two consecutive residues, or two residues separated by one or two residues in a protein sequence. Wang et al. and Tan et al. used all three types of dipeptide compositions that are separated by zero, one, or two residues [95,98]. As for the physicochemical properties, 28 descriptors have been selected from over 500 physicochemical and biochemical properties in the AAindex database for classification between SSBs and DSBs [95,98,106].

In addition to the sequence-based descriptors, some sequence-derived structural features have been used for classification between SSBs and DSBs, such as the predicted secondary structure types, relative solvent accessibility, and the probability being disordered [98]. The prediction accuracy of secondary structure types improved greatly with deep learning approaches. DeepCNF, a deep learning-based secondary structure prediction method, has achieved a three-state accuracy of 84% [107]. With the availability of AlphaFold, we expect to see that structural features will play more roles in future prediction of DNA binding proteins.

3.3. Classification Models

Several traditional machine learning methods, such as random forests (RF) [108] and support vector machines (SVM) [109] have been used in SSB predictions [95–97] (Table 1 and Figure 3). RF and SVM are two very popular traditional machine learning methods. They have been extensively used for protein structure and function prediction with good performance before the era of deep learning. Random forest is an ensemble learning method with many decision trees for both classification and regression problems. Compared to the simple decision tree method, it has a low risk of overfitting. Random forest methods have been applied for predicting protein–protein interactions, protein structure prediction, and function prediction, such as prediction of DNA binding proteins [88,89,110–112]. Support vector machines use hyperplane separation and kernel function optimization to classify a given dataset [113]. SVM methods have been used in improving the prediction of protein secondary structure types, protein fold recognition, and functional prediction [83,85,114–116]. There are different types of kernel functions that can be used for SVM classification, including linear, polynomial, radial basis function (RBF), and sigmoid. In SSB prediction, while some studies only used one kernel function, such as using the default parameters in Wang et al. or RBF in Sharma et al., Ali et al. tried three kernel functions: linear, polynomial, and RBF [95–97]. For the RF methods, the number of trees was set at 3000 in both RF models [95,96]). Tan et al. used gradient tree boosting (GTB) for classification and the parameters were determined by a 10-fold cross-validation based on the training set with a grid search approach [97]. The performances of these classifiers are summarized in the next section.

3.4. Performance Evaluation

The performances of the predictive models, a combination of classifiers and weighted features, were evaluated with the widely used measurements for classification problems (Figure 3). These measurements include accuracy, sensitivity, specificity, MCC, AUC (area under the ROC curve), and F-score (or F1-score). These values were calculated on both the training set and the independent test set. To estimate the performance of machine learning models, usually a 10-fold cross validation method was applied to the training dataset before assessing the models on the independent test set [95–98]. Among the four aforementioned classification models for SSB prediction, there is a mixed performance even with the same set of feature descriptors, datasets, and classifier models. With HMM profiles, it seems that SVM performs slightly better than the RF and KNN methods [96]. However, with a combination of OAAC, dipeptide composition, physicochemical properties, and PSSM, Wang et al. found that RF is a better predictor than SVM [95].

4. Challenges and Perspective

In structure-based studies of SSBs, one major challenge is the limited number of known protein–ssDNA complexes in PDB. Machine learning-based approaches can play an important role in identifying novel SSBs and accordingly increasing the number of structures of SSBs and SSB–ssDNA complexes in follow-up studies. For sequence-based SSB prediction, currently all the published studies predict SSBs from DSBs with an assumption that we already know that a protein being predicted is a DNA-binding protein. However, to be practically useful, the more challenging problem is to predict if any given protein is a SSB or not. To improve prediction performance, as discussed above, well-designed, carefully annotated, and robust DNA binding protein datasets for training and testing are necessary. In addition, a carefully considered balance between the number of cases and redundancy is critical for developing a better prediction model. Novel sequence-based features including sequence-based structural features need to be developed for better classification or prediction. Lastly, deep learning methods have been successful recently in structural bioinformatics, such as in prediction of protein secondary structure types, solvent accessibility and disorder regions, and protein structure prediction [78,79,81,107]. However, the deep learning models typically require a large number of training protein sequences to avoid overfitting. Therefore, novel strategies need to be developed for creative use of deep learning technology for SSB prediction.

Author Contributions: Writing-original draft preparation, J.-T.G.; writing-review and editing, J.-T.G. and F.M.; figure preparation, F.M.; funding acquisition, J.-T.G. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Institutes of Health [R15GM132846 to J.G].

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Acknowledgments: The authors would like to thank the anonymous reviewers for their constructive comments.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Watson, J.D.; Crick, F.H. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature* **1953**, *171*, 737–738. [CrossRef] [PubMed]
2. Dickey, T.H.; Altschuler, S.E.; Wuttke, D.S. Single-stranded DNA-binding proteins: Multiple domains for multiple functions. *Structure* **2013**, *21*, 1074–1084. [CrossRef] [PubMed]
3. Mishra, G.; Levy, Y. Molecular determinants of the interactions between proteins and ssDNA. *Proc. Natl. Acad. Sci. USA* **2015**, *112*, 5033–5038. [CrossRef] [PubMed]
4. Eoff, R.L.; Raney, K.D. A catch and release program for single-stranded DNA. *J. Biol. Chem.* **2017**, *292*, 13085–13086. [CrossRef]
5. Ashton, N.W.; Bolderson, E.; Cubeddu, L.; O'Byrne, K.J.; Richard, D.J. Human single- stranded DNA binding proteins are essential for maintaining genomic stability. *BMC Mol. Biol.* **2013**, *14*, 9. [CrossRef]
6. Mortusewicz, O.; Evers, B.; Helleday, T. PC4 promotes genome stability and DNA repair through binding of ssDNA at DNA damage sites. *Oncogene* **2016**, *35*, 761–770. [CrossRef]
7. Croft, L.V.; Bolderson, E.; Adams, M.N.; El-Kamand, S.; Kariawasam, R.; Cubeddu, L.; Gamsjaeger, R.; Richard, D.J. Human single-stranded DNA binding protein 1 (hSSB1, OBFC2B), a critical component of the DNA damage response. *Semin. Cell Dev. Biol.* **2018**, *86*, 121–128. [CrossRef]
8. Croy, J.E.; Wuttke, D.S. Themes in ssDNA recognition by telomere-end protection proteins. *Trends Biochem. Sci.* **2006**, *31*, 516–525. [CrossRef]
9. Lloyd, N.R.; Dickey, T.H.; Hom, R.A.; Wuttke, D.S. Tying up the Ends: Plasticity in the Recognition of Single-Stranded DNA at Telomeres. *Biochemistry* **2016**, *55*, 5326–5340. [CrossRef]
10. Alberts, B.M.; Frey, L. T4 bacteriophage gene 32: A structural protein in the replication and recombination of DNA. *Nature* **1970**, *227*, 1313–1318. [CrossRef]
11. Sigal, N.; Delius, H.; Kornberg, T.; Gefter, M.L.; Alberts, B. A DNA-unwinding protein isolated from Escherichia coli: Its interaction with DNA and with DNA polymerases. *Proc. Natl. Acad. Sci. USA* **1972**, *69*, 3537–3541. [CrossRef]
12. Overman, L.B.; Lohman, T.M. Linkage of pH, anion and cation effects in protein-nucleic acid equilibria. Escherichia coli SSB protein-single stranded nucleic acid interactions. *J. Mol. Biol.* **1994**, *236*, 165–178. [CrossRef]

13. Wobbe, C.R.; Weissbach, L.; Borowiec, J.A.; Dean, F.B.; Murakami, Y.; Bullock, P.; Hurwitz, J. Replication of simian virus 40 origin-containing DNA in vitro with purified proteins. *Proc. Natl. Acad. Sci. USA* **1987**, *84*, 1834–1838. [CrossRef]
14. Wold, M.S.; Kelly, T. Purification and characterization of replication protein A, a cellular protein required for in vitro replication of simian virus 40 DNA. *Proc. Natl. Acad. Sci. USA* **1988**, *85*, 2523–2527. [CrossRef]
15. Fairman, M.P.; Stillman, B. Cellular factors required for multiple stages of SV40 DNA replication in vitro. *EMBO J.* **1988**, *7*, 1211–1218. [CrossRef]
16. Dean, F.B.; Bullock, P.; Murakami, Y.; Wobbe, C.R.; Weissbach, L.; Hurwitz, J. Simian virus 40 (SV40) DNA replication: SV40 large T antigen unwinds DNA containing the SV40 origin of replication. *Proc. Natl. Acad. Sci. USA* **1987**, *84*, 16–20. [CrossRef]
17. Wu, Y.; Lu, J.; Kang, T. Human single-stranded DNA binding proteins: Guardians of genome stability. *Acta Biochim. Biophys. Sin.* **2016**, *48*, 671–677. [CrossRef]
18. Richard, D.J.; Bolderson, E.; Cubeddu, L.; Wadsworth, R.I.; Savage, K.; Sharma, G.G.; Nicolette, M.L.; Tsvetanov, S.; McIlwraith, M.J.; Pandita, R.K.; et al. Single-stranded DNA-binding protein hSSB1 is critical for genomic stability. *Nature* **2008**, *453*, 677–681. [CrossRef]
19. Bunch, J.T.; Bae, N.S.; Leonardi, J.; Baumann, P. Distinct requirements for Pot1 in limiting telomere length and maintaining chromosome stability. *Mol. Cell Biol.* **2005**, *25*, 5567–5578. [CrossRef]
20. Veldman, T.; Etheridge, K.T.; Counter, C.M. Loss of hPot1 function leads to telomere instability and a cut-like phenotype. *Curr. Biol.* **2004**, *14*, 2264–2270. [CrossRef]
21. Murzin, A.G. OB (oligonucleotide/oligosaccharide binding)-fold: Common structural and functional solution for non-homologous sequences. *EMBO J.* **1993**, *12*, 861–867. [CrossRef]
22. Gamsjaeger, R.; Kariawasam, R.; Gimenez, A.X.; Touma, C.; McIlwain, E.; Bernardo, R.E.; Shepherd, N.E.; Ataide, S.F.; Dong, Q.; Richard, D.J.; et al. The structural basis of DNA binding by the single-stranded DNA-binding protein from Sulfolobus solfataricus. *Biochem. J.* **2015**, *465*, 337–346. [CrossRef]
23. Corona, R.I.; Guo, J.T. Statistical analysis of structural determinants for protein-DNA-binding specificity. *Proteins* **2016**, *84*, 1147–1161. [CrossRef]
24. Rohs, R.; Jin, X.; West, S.M.; Joshi, R.; Honig, B.; Mann, R.S. Origins of specificity in protein-DNA recognition. *Annu. Rev. Biochem.* **2010**, *79*, 233–269. [CrossRef] [PubMed]
25. Corona, R.I.; Sudarshan, S.; Aluru, S.; Guo, J.T. An SVM-based method for assessment of transcription factor-DNA complex models. *BMC Bioinform.* **2018**, *19*, 506. [CrossRef] [PubMed]
26. Lin, M.; Guo, J.T. New insights into protein-DNA binding specificity from hydrogen bond based comparative study. *Nucleic Acids Res.* **2019**, *47*, 11103–11113. [CrossRef] [PubMed]
27. Lin, M.; Malik, F.K.; Guo, J.T. A comparative study of protein-ssDNA interactions. *NAR Genom. Bioinform.* **2021**, *3*, lqab006. [CrossRef]
28. Malik, F.K.; Guo, J.T. Insights into protein-DNA interactions from hydrogen bond energy-based comparative protein-ligand analyses. *Proteins* **2022**, *90*, 1303–1314. [CrossRef]
29. Angarica, V.E.; Perez, A.G.; Vasconcelos, A.T.; Collado-Vides, J.; Contreras-Moreira, B. Prediction of TF target sites based on atomistic models of protein-DNA complexes. *BMC Bioinform.* **2008**, *9*, 436. [CrossRef]
30. Luscombe, N.M.; Laskowski, R.A.; Thornton, J.M. Amino acid-base interactions: A three-dimensional analysis of protein-DNA interactions at an atomic level. *Nucleic Acids Res.* **2001**, *29*, 2860–2874. [CrossRef]
31. Seeman, N.C.; Rosenberg, J.M.; Rich, A. Sequence-specific recognition of double helical nucleic acids by proteins. *Proc. Natl. Acad. Sci. USA* **1976**, *73*, 804–808. [CrossRef]
32. Lei, M.; Podell, E.R.; Cech, T.R. Structure of human POT1 bound to telomeric single-stranded DNA provides a model for chromosome end-protection. *Nat. Struct. Mol. Biol.* **2004**, *11*, 1223–1229. [CrossRef]
33. Bochkarev, A.; Pfuetzner, R.A.; Edwards, A.M.; Frappier, L. Structure of the single-stranded-DNA-binding domain of replication protein A bound to DNA. *Nature* **1997**, *385*, 176–181. [CrossRef]
34. Yadav, T.; Carrasco, B.; Myers, A.R.; George, N.P.; Keck, J.L.; Alonso, J.C. Genetic recombination in Bacillus subtilis: A division of labor between two single-strand DNA-binding proteins. *Nucleic Acids Res.* **2012**, *40*, 5546–5559. [CrossRef]
35. Cernooka, E.; Rumnieks, J.; Tars, K.; Kazaks, A. Structural Basis for DNA Recognition of a Single-stranded DNA-binding Protein from Enterobacter Phage Enc34. *Sci. Rep.* **2017**, *7*, 15529. [CrossRef]
36. Crichlow, G.V.; Zhou, H.; Hsiao, H.H.; Frederick, K.B.; Debrosse, M.; Yang, Y.; Folta-Stogniew, E.J.; Chung, H.J.; Fan, C.; De la Cruz, E.M.; et al. Dimerization of FIR upon FUSE DNA binding suggests a mechanism of c-myc inhibition. *EMBO J.* **2008**, *27*, 277–289. [CrossRef]
37. Myers, J.C.; Shamoo, Y. Human UP1 as a model for understanding purine recognition in the family of proteins containing the RNA recognition motif (RRM). *J. Mol. Biol.* **2004**, *342*, 743–756. [CrossRef]
38. Soufari, H.; Mackereth, C.D. Conserved binding of GCAC motifs by MEC-8, couch potato, and the RBPMS protein family. *RNA* **2017**, *23*, 308–316. [CrossRef]
39. Amrane, S.; Rebora, K.; Zniber, I.; Dupuy, D.; Mackereth, C.D. Backbone-independent nucleic acid binding by splicing factor SUP-12 reveals key aspects of molecular recognition. *Nat. Commun.* **2014**, *5*, 4595. [CrossRef]
40. Joshi, R.; Passner, J.M.; Rohs, R.; Jain, R.; Sosinsky, A.; Crickmore, M.A.; Jacob, V.; Aggarwal, A.K.; Honig, B.; Mann, R.S. Functional specificity of a Hox protein mediated by the recognition of minor groove structure. *Cell* **2007**, *131*, 530–543. [CrossRef]

41. Rohs, R.; West, S.M.; Liu, P.; Honig, B. Nuance in the double-helix and its role in protein-DNA recognition. *Curr. Opin. Struct. Biol.* **2009**, *19*, 171–177. [CrossRef]
42. Rohs, R.; West, S.M.; Sosinsky, A.; Liu, P.; Mann, R.S.; Honig, B. The role of DNA shape in protein-DNA recognition. *Nature* **2009**, *461*, 1248–1253. [CrossRef] [PubMed]
43. Yang, L.; Orenstein, Y.; Jolma, A.; Yin, Y.; Taipale, J.; Shamir, R.; Rohs, R. Transcription factor family-specific DNA shape readout revealed by quantitative specificity models. *Mol. Syst. Biol.* **2017**, *13*, 910. [CrossRef]
44. Luscombe, N.M.; Thornton, J.M. Protein-DNA interactions: Amino acid conservation and the effects of mutations on binding specificity. *J. Mol. Biol.* **2002**, *320*, 991–1009. [CrossRef]
45. Berman, H.M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T.N.; Weissig, H.; Shindyalov, I.N.; Bourne, P.E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235–242. [CrossRef] [PubMed]
46. Burley, S.K.; Berman, H.M.; Christie, C.; Duarte, J.M.; Feng, Z.; Westbrook, J.; Young, J.; Zardecki, C. RCSB Protein Data Bank: Sustaining a living digital data resource that enables breakthroughs in scientific research and biomedical education. *Protein Sci.* **2018**, *27*, 316–330. [CrossRef] [PubMed]
47. Apweiler, R.; Bairoch, A.; Wu, C.H.; Barker, W.C.; Boeckmann, B.; Ferro, S.; Gasteiger, E.; Huang, H.; Lopez, R.; Magrane, M.; et al. UniProt: The Universal Protein knowledgebase. *Nucleic Acids Res.* **2004**, *32*, D115–D119. [CrossRef]
48. Wang, W.; Liu, J.; Sun, L. Surface shapes and surrounding environment analysis of single- and double-stranded DNA-binding proteins in protein-DNA interface. *Proteins* **2016**, *84*, 979–989. [CrossRef]
49. Swamynathan, S.K.; Nambiar, A.; Guntaka, R.V. Role of single-stranded DNA regions and Y-box proteins in transcriptional regulation of viral and cellular genes. *FASEB J.* **1998**, *12*, 515–522. [CrossRef]
50. Duncan, R.; Bazar, L.; Michelotti, G.; Tomonaga, T.; Krutzsch, H.; Avigan, M.; Levens, D. A sequence-specific, single-strand binding protein activates the far upstream element of c-myc and defines a new DNA-binding motif. *Genes Dev.* **1994**, *8*, 465–480. [CrossRef]
51. Tomonaga, T.; Levens, D. Activating transcription from single stranded DNA. *Proc. Natl. Acad. Sci. USA* **1996**, *93*, 5830–5835. [CrossRef]
52. Gupta, M.; Sueblinvong, V.; Raman, J.; Jeevanandam, V.; Gupta, M.P. Single-stranded DNA-binding proteins PURalpha and PURbeta bind to a purine-rich negative regulatory element of the alpha-myosin heavy chain gene and control transcriptional and translational regulation of the gene expression. Implications in the repression of alpha-myosin heavy chain during heart failure. *J. Biol. Chem.* **2003**, *278*, 44935–44948.
53. Thakur, S.; Nakamura, T.; Calin, G.; Russo, A.; Tamburrino, J.F.; Shimizu, M.; Baldassarre, G.; Battista, S.; Fusco, A.; Wassell, R.P.; et al. Regulation of BRCA1 transcription by specific single-stranded DNA binding factors. *Mol. Cell Biol.* **2003**, *23*, 3774–3787. [CrossRef]
54. Phillips, B.W.; Sharma, R.; Leco, P.A.; Edwards, D.R. A sequence-selective single-strand DNA-binding protein regulates basal transcription of the murine tissue inhibitor of metalloproteinases-1 (Timp-1) gene. *J. Biol. Chem.* **1999**, *274*, 22197–22207. [CrossRef]
55. Ko, J.L.; Loh, H.H. Single-stranded DNA-binding complex involved in transcriptional regulation of mouse mu-opioid receptor gene. *J. Biol. Chem.* **2001**, *276*, 788–795. [CrossRef]
56. Desveaux, D.; Despres, C.; Joyeux, A.; Subramaniam, R.; Brisson, N. PBF-2 is a novel single-stranded DNA binding factor implicated in PR-10a gene activation in potato. *Plant. Cell* **2000**, *12*, 1477–1489. [CrossRef]
57. Boyle, B.; Brisson, N. Repression of the defense gene PR-10a by the single-stranded DNA binding protein SEBF. *Plant. Cell* **2001**, *13*, 2525–2537. [CrossRef]
58. Desveaux, D.; Allard, J.; Brisson, N.; Sygusch, J. A new family of plant transcription factors displays a novel ssDNA-binding surface. *Nat. Struct. Biol.* **2002**, *9*, 512–517. [CrossRef]
59. Grabowski, E.; Miao, Y.; Mulisch, M.; Krupinska, K. Single-stranded DNA-binding protein Whirly1 in barley leaves is located in plastids and the nucleus of the same cell. *Plant Physiol.* **2008**, *147*, 1800–1804. [CrossRef]
60. Richard, D.J.; Bell, S.D.; White, M.F. Physical and functional interaction of the archaeal single-stranded DNA-binding protein SSB with RNA polymerase. *Nucleic Acids Res.* **2004**, *32*, 1065–1074. [CrossRef]
61. Liu, J.; Kouzine, F.; Nie, Z.; Chung, H.J.; Elisha-Feil, Z.; Weber, A.; Zhao, K.; Levens, D. The FUSE/FBP/FIR/TFIIH system is a molecular machine programming a pulse of c-myc expression. *EMBO J.* **2006**, *25*, 2119–2130. [CrossRef]
62. Michelotti, E.F.; Michelotti, G.A.; Aronsohn, A.I.; Levens, D. Heterogeneous nuclear ribonucleoprotein K is a transcription factor. *Mol. Cell Biol.* **1996**, *16*, 2350–2360. [CrossRef]
63. Yoo, H.H.; Kwon, C.; Lee, M.M.; Chung, I.K. Single-stranded DNA binding factor AtWHY1 modulates telomere length homeostasis in Arabidopsis. *Plant J.* **2007**, *49*, 442–451. [CrossRef]
64. Wang, Y.; Putnam, C.D.; Kane, M.F.; Zhang, W.; Edelmann, L.; Russell, R.; Carrion, D.V.; Chin, L.; Kucherlapati, R.; Kolodner, R.D.; et al. Mutation in Rpa1 results in defective DNA double-strand break repair, chromosomal instability and cancer in mice. *Nat. Genet.* **2005**, *37*, 750–755. [CrossRef]
65. Shi, W.; Bain, A.L.; Schwer, B.; Al-Ejeh, F.; Smith, C.; Wong, L.; Chai, H.; Miranda, M.S.; Ho, U.; Kawaguchi, M.; et al. Essential developmental, genomic stability, and tumour suppressor functions of the mouse orthologue of hSSB1/NABP2. *PLoS Genet.* **2013**, *9*, e1003298. [CrossRef]
66. Burns, M.B.; Lackey, L.; Carpenter, M.A.; Rathore, A.; Land, A.M.; Leonard, B.; Refsland, E.W.; Kotandeniya, D.; Tretyakova, N.; Nikas, J.B.; et al. APOBEC3B is an enzymatic source of mutation in breast cancer. *Nature* **2013**, *494*, 366–370. [CrossRef]

67. Burns, M.B.; Temiz, N.A.; Harris, R.S. Evidence for APOBEC3B mutagenesis in multiple human cancers. *Nat. Genet.* **2013**, *45*, 977–983. [CrossRef]
68. Thorslund, T.; McIlwraith, M.J.; Compton, S.A.; Lekomtsev, S.; Petronczki, M.; Griffith, J.D.; West, S.C. The breast cancer tumor suppressor BRCA2 promotes the specific targeting of RAD51 to single-stranded DNA. *Nat. Struct. Mol. Biol.* **2010**, *17*, 1263–1265. [CrossRef]
69. Venkitaraman, A.R. Tumour suppressor mechanisms in the control of chromosome stability: Insights from BRCA2. *Mol. Cells* **2014**, *37*, 95–99. [CrossRef]
70. Zamborszky, J.; Szikriszt, B.; Gervai, J.Z.; Pipek, O.; Poti, A.; Krzystanek, M.; Ribli, D.; Szalai-Gindl, J.M.; Csabai, I.; Szallasi, Z.; et al. Loss of BRCA1 or BRCA2 markedly increases the rate of base substitution mutagenesis and has distinct effects on genomic deletions. *Oncogene* **2017**, *36*, 746–755. [CrossRef] [PubMed]
71. Shuck, S.C.; Turchi, J.J. Targeted inhibition of Replication Protein A reveals cytotoxic activity, synergy with chemotherapeutic DNA-damaging agents, and insight into cellular function. *Cancer Res.* **2010**, *70*, 3189–3198. [CrossRef] [PubMed]
72. Kustatscher, G.; Collins, T.; Gingras, A.C.; Guo, T.; Hermjakob, H.; Ideker, T.; Lilley, K.S.; Lundberg, E.; Marcotte, E.M.; Ralser, M.; et al. Understudied proteins: Opportunities and challenges for functional proteomics. *Nat. Methods* **2022**, *19*, 774–779. [CrossRef] [PubMed]
73. Kustatscher, G.; Collins, T.; Gingras, A.C.; Guo, T.; Hermjakob, H.; Ideker, T.; Lilley, K.S.; Lundberg, E.; Marcotte, E.M.; Ralser, M.; et al. An open invitation to the Understudied Proteins Initiative. *Nat. Biotechnol.* **2022**, *40*, 815–817. [CrossRef] [PubMed]
74. Levitt, M. Nature of the protein universe. *Proc. Natl. Acad Sci. USA* **2009**, *106*, 11079–11084. [CrossRef]
75. Galperin, M.Y.; Koonin, E.V. 'Conserved hypothetical' proteins: Prioritization of targets for experimental study. *Nucleic Acids Res.* **2004**, *32*, 5452–5463. [CrossRef]
76. Shumilin, I.A.; Cymborowski, M.; Chertihin, O.; Jha, K.N.; Herr, J.C.; Lesley, S.A.; Joachimiak, A.; Minor, W. Identification of unknown protein function using metabolite cocktail screening. *Structure* **2012**, *20*, 1715–1725. [CrossRef]
77. Ellens, K.W.; Christian, N.; Singh, C.; Satagopam, V.P.; May, P.; Linster, C.L. Confronting the catalytic dark matter encoded by sequenced genomes. *Nucleic Acids Res.* **2017**, *45*, 11495–11514. [CrossRef]
78. AlQuraishi, M. Machine learning in protein structure prediction. *Curr. Opin. Chem. Biol.* **2021**, *65*, 1–8. [CrossRef]
79. Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Zidek, A.; Potapenko, A.; et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **2021**, *596*, 583–589. [CrossRef]
80. Wang, S.; Sun, S.; Li, Z.; Zhang, R.; Xu, J. Accurate De Novo Prediction of Protein Contact Map by Ultra-Deep Learning Model. *PLoS Comput. Biol.* **2017**, *13*, e1005324. [CrossRef]
81. Xu, J. Distance-based protein folding powered by deep learning. *Proc. Natl. Acad. Sci. USA* **2019**, *116*, 16856–16865. [CrossRef]
82. Kumar, K.K.; Pugalenthi, G.; Suganthan, P.N. DNA-Prot: Identification of DNA binding proteins from protein sequence information using random forest. *J. Biomol. Struct. Dyn.* **2009**, *26*, 679–686. [CrossRef]
83. Kumar, M.; Gromiha, M.M.; Raghava, G.P. Identification of DNA-binding proteins using support vector machines and evolutionary profiles. *BMC Bioinform.* **2007**, *8*, 463. [CrossRef]
84. Qiu, J.; Bernhofer, M.; Heinzinger, M.; Kemper, S.; Norambuena, T.; Melo, F.; Rost, B. ProNA2020 predicts protein-DNA, protein-RNA, and protein-protein binding proteins and residues from sequence. *J. Mol. Biol.* **2020**, *432*, 2428–2443. [CrossRef]
85. Xu, R.; Zhou, J.; Wang, H.; He, Y.; Wang, X.; Liu, B. Identifying DNA-binding proteins by combining support vector machine and PSSM distance transformation. *BMC Syst. Biol.* **2015**, *9* (Suppl. S1), S10. [CrossRef]
86. Ali, F.; Ahmed, S.; Swati, Z.N.K.; Akbar, S. DP-BINDER: Machine learning model for prediction of DNA-binding proteins by fusing evolutionary and physicochemical information. *J. Comput. Aided. Mol. Des.* **2019**, *33*, 645–658. [CrossRef]
87. Hu, S.; Ma, R.; Wang, H. An improved deep learning method for predicting DNA- binding proteins based on contextual features in amino acid sequences. *PLoS ONE* **2019**, *14*, e0225317. [CrossRef]
88. Lou, W.; Wang, X.; Chen, F.; Chen, Y.; Jiang, B.; Zhang, H. Sequence based prediction of DNA-binding proteins based on hybrid feature selection using random forest and Gaussian naive Bayes. *PLoS ONE* **2014**, *9*, e86703. [CrossRef]
89. Ma, X.; Guo, J.; Sun, X. DNABP: Identification of DNA-Binding Proteins Based on Feature Selection Using a Random Forest and Predicting Binding Residues. *PLoS ONE* **2016**, *11*, e0167345. [CrossRef]
90. Mishra, A.; Pokhrel, P.; Hoque, M.T. StackDPPred: A stacking based prediction of DNA-binding protein from sequence. *Bioinformatics* **2019**, *35*, 433–441. [CrossRef]
91. Motion, G.B.; Howden, A.J.; Huitema, E.; Jones, S. DNA-binding protein prediction using plant specific support vector machines: Validation and application of a new genome annotation tool. *Nucleic Acids Res.* **2015**, *43*, e158. [CrossRef] [PubMed]
92. Qu, Y.H.; Yu, H.; Gong, X.J.; Xu, J.H.; Lee, H.S. On the prediction of DNA-binding proteins only from primary sequences: A deep learning approach. *PLoS ONE* **2017**, *12*, e0188129. [CrossRef] [PubMed]
93. Li, G.; Du, X.; Li, X.; Zou, L.; Zhang, G.; Wu, Z. Prediction of DNA binding proteins using local features and long-term dependencies with primary sequences based on deep learning. *PeerJ* **2021**, *9*, e11262. [CrossRef] [PubMed]
94. Zaitzeff, A.; Leiby, N.; Motta, F.C.; Haase, S.B.; Singer, J.M. Improved data sets and evaluation methods for the automatic prediction of DNA-binding proteins. *Bioinformatics* **2021**, *38*, 44–51. [CrossRef]
95. Wang, W.; Sun, L.; Zhang, S.; Zhang, H.; Shi, J.; Xu, T.; Li, K. Analysis and prediction of single-stranded and double-stranded DNA binding proteins based on protein sequences. *BMC Bioinform.* **2017**, *18*, 300. [CrossRef]

96. Sharma, R.; Kumar, S.; Tsunoda, T.; Kumarevel, T.; Sharma, A. Single-stranded and double-stranded DNA-binding protein prediction using HMM profiles. *Anal. Biochem.* **2021**, *612*, 113954. [CrossRef]
97. Ali, F.; Arif, M.; Khan, Z.U.; Kabir, M.; Ahmed, S.; Yu, D.J. SDBP-Pred: Prediction of single-stranded and double-stranded DNA-binding proteins by extending consensus sequence and K-segmentation strategies into PSSM. *Anal. Biochem.* **2020**, *589*, 113494. [CrossRef]
98. Tan, C.; Wang, T.; Yang, W.; Deng, L. PredPSD: A Gradient Tree Boosting Approach for Single-Stranded and Double-Stranded DNA Binding Protein Prediction. *Molecules* **2019**, *25*, 98. [CrossRef]
99. Selby, C.P.; Sancar, A. A cryptochrome/photolyase class of enzymes with single-stranded DNA-specific photolyase activity. *Proc. Natl. Acad. Sci. USA* **2006**, *103*, 17696–17700. [CrossRef]
100. Pokorny, R.; Klar, T.; Hennecke, U.; Carell, T.; Batschauer, A.; Essen, L.O. Recognition and repair of UV lesions in loop structures of duplex DNA by DASH-type cryptochrome. *Proc. Natl. Acad. Sci. USA* **2008**, *105*, 21023–21027. [CrossRef]
101. Bakalkin, G.; Yakovleva, T.; Selivanova, G.; Magnusson, K.P.; Szekely, L.; Kiseleva, E.; Klein, G.; Terenius, L.; Wiman, K.G. p53 binds single-stranded DNA ends and catalyzes DNA renaturation and strand transfer. *Proc. Natl. Acad. Sci. USA* **1994**, *91*, 413–417. [CrossRef]
102. Bochkareva, E.; Kaustov, L.; Ayed, A.; Yi, G.S.; Lu, Y.; Pineda-Lucena, A.; Liao, J.C.; Okorokov, A.L.; Milner, J.; Arrowsmith, C.H.; et al. Single-stranded DNA mimicry in the p53 transactivation domain interaction with replication protein A. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 15412–15417. [CrossRef]
103. Altschul, S.F.; Madden, T.L.; Schaffer, A.A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D.J. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **1997**, *25*, 3389–3402. [CrossRef]
104. Gribskov, M.; McLachlan, A.D.; Eisenberg, D. Profile analysis: Detection of distantly related proteins. *Proc. Natl. Acad. Sci. USA* **1987**, *84*, 4355–4358. [CrossRef]
105. Feng, Z.P.; Zhang, C.T. Prediction of the subcellular location of prokaryotic proteins based on the hydrophobicity index of amino acids. *Int. J. Biol. Macromol.* **2001**, *28*, 255–261. [CrossRef]
106. Huang, H.L.; Lin, I.C.; Liou, Y.F.; Tsai, C.T.; Hsu, K.T.; Huang, W.L.; Ho, S.J.; Ho, S.Y. Predicting and analyzing DNA-binding domains using a systematic approach to identifying a set of informative physicochemical and biochemical properties. *BMC Bioinform.* **2011**, *12* (Suppl. S1), S47. [CrossRef]
107. Wang, S.; Peng, J.; Ma, J.; Xu, J. Protein Secondary Structure Prediction Using Deep Convolutional Neural Fields. *Sci. Rep.* **2016**, *6*, 18962. [CrossRef]
108. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
109. Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn.* **1995**, *20*, 273–297. [CrossRef]
110. Wang, S.; Sun, S.; Xu, J. AUC-Maximized Deep Convolutional Neural Fields for Protein Sequence Labeling. In *Machine Learning and Knowledge Discovery in Databases*; Springer: Berlin/Heidelberg, Germany, 2016; Volume 9852, pp. 1–16.
111. Hou, Q.; De Geest, P.F.G.; Vranken, W.F.; Heringa, J.; Feenstra, K.A. Seeing the trees through the forest: Sequence-based homo- and heteromeric protein-protein interaction sites prediction using random forest. *Bioinformatics* **2017**, *33*, 1479–1487. [CrossRef]
112. Jo, T.; Cheng, J. Improving protein fold recognition by random forest. *BMC Bioinform.* **2014**, *15* (Suppl. S11), S14. [CrossRef]
113. Noble, W.S. What is a support vector machine? *Nat. Biotechnol.* **2006**, *24*, 1565–1567. [CrossRef]
114. Cheng, J.; Baldi, P. A machine learning information retrieval approach to protein fold recognition. *Bioinformatics* **2006**, *22*, 1456–1463. [CrossRef]
115. Ward, J.J.; McGuffin, L.J.; Buxton, B.F.; Jones, D.T. Secondary structure prediction with support vector machines. *Bioinformatics* **2003**, *19*, 1650–1655. [CrossRef]
116. Ding, C.H.; Dubchak, I. Multi-class protein fold recognition using support vector machines and neural networks. *Bioinformatics* **2001**, *17*, 349–358. [CrossRef]