Fakultät für Informatik
Technische Universität München

Dissertation

Sequence-based prediction reveals effect of protein-, DNA-, RNA-binding residues on sequence variants

Jiajun Qiu

Fakultät für Informatik

Sequence-based prediction reveals effect of protein-, DNA- ,RNA-binding residues on sequence variants

Jiajun Qiu

Vollständiger Abdruck der von der
Fakultät für Informatik
der Technischen Universität München zur Erlangung des akademischen Grades
eines Doktors der Naturwissenschaften (Dr. rer. nat.)
genehmigten Dissertation.

Vorsitzende/-r: Prof. Dr. Julien Gagneur

Prüfende/-r der Dissertation:

1. Prof. Dr. Burkhard Rost
2. Prof. Dr. Stefan Kramer

Die Dissertation wurde am 21.10.2020 bei der Technischen Universität München
eingereicht und durch die Fakultät für Informatik am 27.12.2020 angenommen.

Abstract

Proteins are one of the most important biological macro-molecules and work as parts of complex networks. The biological properties of a protein molecule depend on its physical interaction with other molecules, especially proteins, DNA and RNA. Thus, the intricate details of how proteins bind to them, are crucial for understanding the mechanism of almost all biological processes. Goal of this thesis was to complete a high-throughput analysis of how those binding residues affect genetic variants and vice versa. Toward this end, the first task was the development of a new and comprehensive system (named ProNA2020) that takes only protein sequence as input to predict binding of protein to DNA, RNA and other proteins and the corresponding binding residues. Then it was applied to the analysis of SAVs from 60,706 people. This revealed that SAVs on those macro-molecular binding residues have more effect on protein function than SAVs outside of those binding residues. Overall, this novel research about binding residues might benefit future research in molecular and medical biology (e.g. precision medicine) both in terms of the methodology and in terms of being used as prediction method that is available through an online server and through github.

Zusammenfassung

Proteine Proteine sind eines der wichtigsten biologischen Makromoleküle. Fast jeder Prozess in der Zelle beinhaltet ein oder mehrere Proteine. Anstatt isoliert zu wirken, arbeiten Proteine als Teile komplexer Netzwerke. Die biologischen Eigenschaften eines Proteinmoleküls hängen von seiner physikalischen Wechselwirkung mit anderen Molekülen ab, insbesondere Proteinen, DNA und RNA. Daher sind die komplizierten Details, wie Proteine an Proteine, DNA und RNA binden, entscheidend für das Verständnis des Mechanismus fast aller biologischen Prozesse. Ziel dieser Arbeit war es, eine Hochdurchsatzanalyse durchzuführen, wie diese Bindungsreste genetische Varianten beeinflussen und umgekehrt. Zu diesem Zweck bestand die erste Aufgabe in der Entwicklung eines neuen und umfassenden Systems, das nur die Proteinsequenz als Input verwendet, um die Bindung von Protein an DNA, RNA und andere Proteine und die entsprechenden Bindungsreste vorherzusagen. Das System kombinierte homologiebasierte Inferenz mit maschinellem Lernen und deckte sowohl Vorhersagen pro Protein (Protein bindet / nicht) als auch pro Rest (Bindung wo) ab. Die Vorhersage des Proteinspiegels beim maschinellen Lernen kombinierte motivbasierte Profilkernansätze mit wortbasierten (ProtVec) Lösungen. Nach der Festlegung der Methode wurde sie auf die Analyse von SAVs (auch als SAVs bezeichnet: Single Amino Acid Variants oder Missense SNV) von 60.706 Personen angewendet. Dies zeigte, dass SAVs auf diesen makromolekularen Bindungsresten einen größeren Einfluss auf die Proteinfunktion haben als SAVs außerhalb dieser Bindungsreste. Insgesamt könnte diese neuartige Forschung über Bindungsreste der zukünftigen Forschung in der Molekular- und Medizinbiologie (z. B. Präzisionsmedizin) sowohl hinsichtlich der Methodik (bestimmte Kombination von Werkzeugen zu einem Vorhersagesystem) als auch hinsichtlich der Verwendung als verfügbare Vorhersagemethode zugute kommen über einen Online-Server und über Github.

Acknowledgements

First and foremost, I would like to thank my supervisor Prof. Burkhard Rost for giving me the opportunity to peruse my PhD degree and conduct the relative research in Rostlab. He can always give me great and useful guidance and suggestion. Moreover, he is patient and shows his understanding for my concerns.

In addition, I would like to express my thanks to Michael Bernhofer. He helps me a lot during my 4-years study, especially for helping me to make my method available online. Also I need to thank Michael Heinzinger and Jonas Reeb who give me a lot of helpful advice during my project.

Also special thanks to all other Rostlab members for creating a wonderful research environment. This is really a lovely group! In particular thanks to Inga who helps a lot with bureaucracy.

And I need to show my great thanks to Tim. Besides the technical assistance, you make me feel extraordinary friendship during my study here as a foreigner. Also I need to give my great thanks to my family members: my wife and my mother. Thank you, Xiao. For many years, we go through thick and thin together. Without your support, I will not be able to finish this thesis. You really make me become a better man. I hope that in the days to come we will always believe in love and support each other. Love you, forever!

Publications

This work constitutes a cumulative dissertation based on the following peer-reviewed publications:

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 Mar 27;432(7):2428-2443. doi: 10.1016/j.jmb.2020.02.026. Epub 2020 Mar 4.

Author contribution: Jiajun Qiu(JQ) and Burkhard Rost (BR) conceptualized the work. JQ performed the whole analysis and model training. Tomas Norambuena and Francisco Melo helped creating the training data. Michael Bernhofer helped to make the method available online. Michael Heinzinger and Sofie Kemper provided useful suggestion and idea for the research. BR provided supervision. BR provided funding. JQ wrote the initial manuscript draft with BR. All authors reviewed and approved of the final manuscript.

Qiu J, Nechaev D and Rost B. Protein‐protein and protein‐nucleic acid binding residues important for common and rare sequence variants in human. *BMC Bioinformatics.* 2020 Oct 13;21(1):452. doi: 10.1186/s12859-020-03759-0.

Author contribution: Jiajun Qiu designed and performed the analysis and writing the manuscript; Dmitrii Nechaev prepared part of dataset and helped in manuscript revision; Burkhard Rost designed and guided the analysis and revised the manuscript. All authors have read and approved the final manuscript.

While working on the dissertation I (co)-authored the following publications:

Cai C[#], **Qiu J[#]**, Qiu G, Chen Y, Song Z, Li J, Gong X. Long non-coding RNA MALAT1 protects preterm infants with bronchopulmonary dysplasia by inhibiting cell apoptosis. *BMC Pulm Med*. 2017 Dec 13;17(1):199. doi: 10.1186/s12890-017-0524-1.

Qiu JJ, Liu YN, Ren ZR, Yan JB. Dysfunctions of mitochondria in close association with strong perturbation of long noncoding RNAs expression in down syndrome. *Int J Biochem Cell Biol*. 2017 Nov;92:115-120. doi: 10.1016/j.biocel.2017.09.017. Epub 2017 Sep 29.

Gong X[#], **Qiu J[#]**, Qiu G, Cai C. Adrenomedullin regulated by miRNA-574-3p protects premature infants with bronchopulmonary dysplasia. *Biosci Rep*. 2020 May 29;40(5):BSR20191879. doi: 10.1042/BSR20191879.

Contents

List of Figures

List of Tables

Chapter 1

1 Introduction

Proteins are polymers comprising 20 chemically and structurally different building blocks (amino acids) that fold into a highly specific tertiary structure (Reichmann et al., 2007). It is one of the most important biological macro-molecular. Almost every event that occurs in the cell involves one or more proteins. More importantly, proteins do not act in isolation but instead work as part of complex networks. The biological properties of a protein molecule depend on its physical interaction with other molecules, especially proteins, DNA and RNA. Thus, the researches focusing on the binding sites and binding residues of proteins will lead to a better understanding of how proteins function. And it can further reveal the mechanism of various biological process.

1.1 Interaction between protein and macro-molecules

1.1.1 Protein-protein interaction

Genome sequencing of more than 10,000 plants, animals, and fungi has been done over the past 60 years (van Straalen and Roelofs, 2006). Scientists thought the information about an organism's genome size should be a foundation to understand the

genetic content (complexity) of the organism. However, there is an extraordinary lack of correspondence between organism complexity and their genome size. For example, the genome size of Protopterus aethiopicus (marbled lungfish) is over 40 times larger than that of human. One haploid copy of this fish's genome is composed of 133 billion base pairs, and one copy of a human haploid genome has only 2.9 billion (Table 1.1). This finding suggests that genome size is not an indicator of the genomic or biological complexity of an organism. And it revolutionizes the system biology era, and the postgenomic events takes extra attention toward explaining the phenotypical complexity (Keskin et al., 2016).

Table 1.1: Genome Size and Number of Protein-Coding Genes for a Selected Handful of Species (van Straalen and Roelofs, 2006)

Species and Common Name	Estimated Total Size of Genome (bp)	Estimated Number of Protein-Encoding Genes
Saccharomyces cerevisiae (unicellular budding yeast)	12 million	6,000
Trichomonas vaginalis	160 million	60,000
Protopterus aethiopicus	133 billion	NA
Plasmodium falciparum (unicellular malaria parasite)	23 million	5,000
Caenorhabditis elegans (nematode)	95.5 million	18,000
Drosophila melanogaster (fruit fly)	170 million	14,000
Arabidopsis thaliana (mustard; thale cress)	125 million	25,000
Oryza sativa (rice)	470 million	51,000
Gallus gallus (chicken)	1 billion	20,000-23,000
Canis familiaris (domestic dog)	2.4 billion	19,000
Mus musculus (laboratory mouse)	2.5 billion	30,000
Homo sapiens (human)	2.9 billion	20,000-25,000

One of the mechanisms amplifying the biological complexity is the communication between proteins. Instead of acting in isolation, more than 80% of all proteins in the cell interact with other molecules to become functional (Berggard et al., 2007). Many cellular processes such as transcription, replication, communication between cells, signaling transduction and membrane transport are dependent on protein interactions. Specific protein-protein interactions (PPIs) are essential for maintaining a robust phenotype (Viswanathan et al., 2019). And studies also find the dysfunction or malfunction of signaling pathways and alterations in protein interactions is the cause of diseases, such as neurodegenerative diseases or cancer (del Sol et al., 2010) (Grechkin et al., 2016).

And, interestingly, 20 natural amino acids are not equally important to obtain tight and specific protein-protein binding. In one study, Sidhu and co-workers (Fellouse et al., 2006) obtained an antigen-binding fragment called Fab-YADS2 from a library with chemical diversity restricted to only four amino acids (Tyr, Ser, Ala and Asp). Fab-YADS2 can recognize vascular endothelial growth factor (VEGF). Mutagenesis experiments reveal that the structural paratope is dominated by Tyr side chains, which represent 11 of the 15 functionally important residues. Isothermal titration calorimetry and cell-based assays show that restricted chemical diversity does not limit the affinity or specificity of Fab-YADS2 relative to natural antibodies. Furthermore, the Tyr has been found to be the most common amino acid in binding sites (Nooren and Thornton, 2003).

There was also a study about the extent of exchangeability of amino acids at the binding site (Pal et al., 2006). They used the complex between human growth hormone (hGH) and its receptor (hGHR) as their experimental platform. The hGH site 1 binding to the hGHR contained 35 residues distributing across four regions: helices 1 and 4 of the four-helix bundle (residues 14 –29 and 164 –183) and two connecting loops (residues 41– 48 and 60 – 67). With shotgun approach, they introduced any one of the 20 natural amino acids at all 35 interface positions. This was a rather unusual approach, because mutational analysis was most often restricted to alanine substitution, which didn’t not provide a comprehensive view of the allowed amino acid space at any specific position (Reichmann et al., 2007). And their results was rather interesting. They verified that the interface was highly adaptable to mutations, but the tolerated mutations

were neither chemically nor evolutionarily conserved. Actually, neither chemical nor evolutionary conservation, which seemed to be very context dependent, was a good indicator of allowed mutations. Some of the alanine scanning hotspot positions showed high specificity against substitution, and others did not. However, some highly specific positions were not hotspots at all.

1.1.2 Protein-DNA interaction

Protein–DNA interactions are widely distributed in all living organisms. Previous studies have estimated that 2%–3% of a prokaryotic genome and 6%–7% of a eukaryotic genome encodes DNA-binding proteins (Luscombe et al., 2000). There are many different DNA-binding proteins (DBPs) with different domains, which involve in a variety of important biological processes.

Transcription factors, are proteins that can regulate the transcription of genetic information from DNA to messenger RNA, by binding to a specific DNA sequence.

DNA polymerases, are enzymes that synthesize DNA molecules from deoxyribonucleotides, which are essential for DNA replication. These enzymes usually work in pairs to create two identical DNA strands from a single original DNA molecule.

Nucleases, are enzymes which are essential machinery for many aspects of DNA repairing in living organisms. Nucleases are capable of cleaving the phosphodiester bonds between nucleotides of nucleic acids. Defects in certain nucleases can cause genetic instability or immunodeficiency (Nishino and Morikawa, 2002).

Histones ,which are comprised of lysine and arginine, are very basic proteins found in eukaryotic cell nuclei. Histones can pack and order the DNA into structural units called nucleosomes (Redon et al., 2002).

And those binding residues on DNA binding proteins can form different domains to recognize double- or single-strand DNA, such as: Helix-turn-helix, Zinc finger, Leucine zipper, Winged helix, and Winged helix-turn-helix.

1.1.3 Protein-RNA interaction

RNA-binding proteins (RBPs) are typically thought as proteins that bind RNA through one or multiple globular RNA-binding domains (RBDs) and can change the fate or function of the bound RNAs. RBPs are involved in almost every central process in the cell and often serve essentially functional roles:

Alternative splicing, is a mechanism by which different forms of mature mRNAs (messengers RNAs) are generated from the same gene. Actually, alternative splicing is another mechanism amplifying the genomic/biological complexity besides PPI (Keskin et al., 2016). More than 90% of all human genes are found to generate alternatively spliced mRNA isoforms (Wang et al., 2008).

mRNA localization, is a spatial mechanism for regulating gene activity. mRNA transportation can increase the efficiency and temporal resolution of protein synthesis in response to cellular cues, and facilitate the formation of protein complexes due to higher local concentration of the necessary mRNAs (Re et al., 2014).mRNA translation, can be directly regulated by RBPs. For example, mRNA-specific RBPs can inhibit the interaction between the ribosome 43S complex and the mRNA by physical hindrance in a cap-dependent manner (Muckenthaler et al., 1998).

RNA editing, is a molecular process through which some cells can make changes to some specific nucleotide sequences within an RNA molecule after transcription. The most common type of RNA editing is A-to-I editing by double-stranded RNA-specific adenosine deaminase (ADAR) enzymes which are RBPs binding specific dsRNA structures (Eisenberg and Levanon, 2018).

1.2 Sequence variants on protein binding residues

A genome is the entire set of genetic material (DNA or RNA) for an organism. For human genome, 99.5% of all DNA is shared in human population. Genetic variants are the rest 0.5%, and it's the differences that make each person's genome unique (Mayor, 2007). Those 0.5% really matter. The genetic variants are associated with various phenotypes such as skin color (Sarkar and Nandineni, 2018), vision and health of our

eyes (Singh and Tyagi, 2018) and height (Lango Allen et al., 2010). Single-nucleotide variants (SNVs) are the vast majority of genetic variants in the human population. There are about 3–4 million SNVs apparent in a typical comparison of one human versus the reference, and the dbSNP catalog (build 151) has over 660 million SNVs from diverse sequencing studies (Lappalainen et al., 2019).

On protein level, SNVs would refer to single amino acid variants (SAVs). Since the protein-, DNA- and RNA-protein interactions are so important in a large number of biological processing, the variants or mutations on those binding proteins or residues will lead to serious consequences.

Recently, to investigate the mechanisms by which cancer mutations peturb protein-protein interactions, H. Billur Engin et al have analyzed the distribution of 1,297,414 somatic missense mutations from 138 genes using 3D protein structures. They find an over-representation of missense mutations at PPI interface residues in both tumor suppressors and oncogenes, which indicates that mutations in cancer tend to affect the PPIs.

Ornithine carbamoyltransferase (OCT) catalyzes the conversion of ornithine and carbamoyl phosphate to citrulline during the second step of the urea cycle. OCT is a homotrimer with active sites located at each of the protein-protein interfaces. Nearly 300 mutations have been identified in OCT, with the vast majority leading to either neonatal or late onset OCT deficiency. Over half of the disease mutations (59%) are linked to changes in protomer stability, and approximately 15% are found to disrupt substrate binding (Jubb et al., 2017).

Rett syndrome (RTT) is a severe neurological disorder caused by MECP2 gene mutations. MeCP2 is a protein with high expression level in the brain that participates in the genetic expression and the regulation of RNA splicing. Molecular dynamics simulations find that P152R mutation within MeCP2 can influence the protein binding to DNA. P152R mutation makes MeCP2 Methyl-CpG-binding domain bind more strongly to DNA, while selectively decreases binding affinity to methylated DNA (Franklin, 2019).

And it is same for protein-RNA interaction. It is known that many diseases are caused by mutations on RNA binding proteins. Mutations in PRPF31, PRPF8 and HPRP3,

which result in defect of SnRNP assembly, lead to retinitis pigmentosa (Wang and Cooper, 2007). Mutations in TERC and TERT, which result in defects of RNP telomerase activity, lead to dyskeratosis congenital (Collins and Mitchell, 2002). Mutations in UPF3B, which result in defect in nonsense-mediated mRNA decay surveillance, lead to syndromic mental retardation and nonsyndromic mental retardation (Tarpey et al., 2007)

Overall, mutation or sequence variants on the protein-, DNA- and RNA-protein binding proteins or residues will lead to significantly mutated phenotype which could be serious diseases. So, it is very necessary to do the analysis about the binding residues in human SAVs, which can benefit for both biology and medicine research (e.g. precision medicine). To do so, we firstly need to identify those binding proteins or residues.

1.3 Binding proteins/residues identification

1.3.1 Experimental based binding proteins/residues identification

There are a lot of experimental methods which have been developed to identify those interactions and the binding proteins. For example, fluorescence resonance energy transfer (FRET) can identify PPI. In FRET, bait and prey proteins are fused to donor (don) and acceptor (acc) molecules such as cyan (CFP) and yellow (YFP) variants of GFP. An interaction between the bait and prey proteins brings the donor and acceptor into close proximity, and excitation of the donor fluorophore results in non-radiative energy transfer and acceptor fluorescence emission at a specific wavelength (Petschnigg et al., 2011).

For protein-nucleotide binding, there are methods such as DNA/RNA pull-down assay which can detect the protein-DNA/RNA interaction. A pull-down assay using DNA/RNA-conjugated beads is widely used in various research fields, which is a direct and versatile tool to study DNA/RNA-protein interaction (Sui et al., 2020). First the

biotinylated-DNA/RNA is incubated with streptavidin, then the recombinant or cellular-extract proteins can bind to DNA/RNA. After being washed, the beads are boiled to identify DNA/RNA-bound proteins.

For the residue level identification (binding residues), it needs to determine the 3D structure of the binding proteins. The wildly used experimental methods are X single crystal X-ray diffraction (SC-XRD), nuclear magnetic resonance (NMR) and cryo-electron microscopy (Cryo-EM). According to the statistics of PDB, about 90% protein structures are resolved by SC-XRD (Burley et al., 2017). However, there is no "universal" method since all three of them have their advantages as well as limitations.

The SC-XRD can yield high atomic resolution and is not limited by the molecular weight of the sample. It is suitable for water-soluble proteins, membrane proteins as well as macromolecular complexes. However, SC-XRD also has disadvantages such as the difficulty for crystallization and diffraction. Especially, for membrane proteins, the large size leads to the poor solubilization of the crystallization (Table 1.2).

NMR can measure the three-dimensional structure of macromolecules in the natural state directly with a very high resolution. But NMR cannot be applied in analyzing large biomolecules and it needs relatively large amounts of pure samples (Table 1.2).

Cryo-EM is a much easier method compared with the two methods above. It requires only a small amount of sample, demands less on sample purity, and does not need to crystalize protein. But, as a cost, it has high levels of noise and relatively low resolution (Table 1.2).

So far, it is expensive and time-consuming to identify the binding residues with all above experimental methods. Especially for high-throughput analysis, it is not possible to prepare all the samples. Nowadays, fewer than 0.36% of all proteins with known sequence in UniProt correspond to a known experimental 3D structure in the PDB (Qiu et al., 2020). Thus, it is necessary to apply *in silico* method to binding residues identification.

Table 1.2: The comparison of X-ray crystallography, NMR and Cryo-EM

Methods	Advantages	Disadvantages	Objects	Resolution
X-ray Crystallography	• Well developed • High resolution • Broad molecular weight range • Easy for model building	• Difficult for crystallization • Difficult for diffraction • Solid structure preferred • Static crystalline state structure	• Crystallizable samples • Soluble proteins, membrane proteins, ribosomes, DNA/RNA and protein complexes	High
NMR	• High resolution • 3D structure in solution • Good for dynamic study	• Need for high sample purity • Difficult for sample preparation • Difficult for computational simulation	• MWs below 40–50 kDa • Water soluble samples	High
Cryo-EM	• Easy sample preparation • Structure in native state • Small sample size	• Relatively low resolution • Applicable to samples of high molecular weights only • Highly dependent on EM techniques • Costly EM equipment	• >150 kDa • Virions, membrane proteins, large proteins, ribosomes, complex compounds	Relatively Low (<3.5 Å)

1.3.2 Computational based binding proteins/residues identification

Basically, all the computational methods can be divided into two categories: structure-based methods and sequence-based methods.

1.3.2.1 Structure-based predictors

Structure-based predictors use structural features such as solvent-accessible surface area, crystallographic B-factor and secondary structure. The growing number of available structural complexes assists the accuracy and availability of structure-based methods.

IntPred is a state-of-the-art structure-based RNA-binding residues prediction method (Northey et al., 2018). It uses the structure-based features such as intra-chain disulphide or hydrogen bonds on the certain residue, secondary structure and planarity of the residues which are calculated by finding the root mean squared distance of all atoms of the patch from a plane of best fit. Overall, IntPred achieves a high accuracy 76% with random forest (Northey et al., 2018).

PRISM, a structure-based PPI prediction method, is another example (Baspinar et al., 2014). PRISM first extracts the surface residues of the target proteins using the relatively accessible surface area values. And each interface in the template interface dataset is split into its constituent chains. Then PRISM checks whether complementary sides of a template interface are structurally similar to any region on the surface of target structures (Shatsky et al., 2004). Once similarities are detected, the two target proteins are transformed into the structurally similar template interface constituting a predicted complex structure (Baspinar et al., 2014).

Though structure-based methods achieve good performance in protein binding, there is an obvious limitation: they can only be applied to protein, whose 3D structure are available. And for proteomic and genomic analysis, which is dependent on large amount of predictions, it is necessary to introduce another kind of method which is based on the sequence information of proteins rather than structure.

1.3.2.2 Sequence-based predictors

Sequence-based predictors use only the sequence information of the query proteins as the input to detect the binding residues. Thus, it can be applied to almost any protein and very suitable for high-throughput analysis. Interface residues or binding residues are more conserved than the rest of the protein surface and these conserved positions can be identified by multiple sequence alignments (MSAs) (Esmaielbeiki et al., 2016). Thus, in the past decades, evolutionary information has significantly improved the performance of binding residues prediction (Ofran and Rost, 2003). And now, most of state-of-the-art methods are based on the combination of the evolutionary information with other sequence features.

The first method (Res et al., 2005), which uses the combination of evolutionary information and residue composition, achieves an accuracy of 64%. It increases 6% compared with the previous sequence-based study (Ofran and Rost, 2003). Since then, many studies try to combine evolutionary information with different sequence features. For example, DNA binding residues prediction method DNABR combines evolutionary information with composition of amino acid and physiochemical properties of amino acids (Ma et al., 2012). And some studies try to combine residue spatial sequence profile obtained from the HSSP database with evolutionary information (Wang et al., 2006).

Some sequence-based methods take advantage of predicted structural information such as surface accessibility and secondary structure. For example, InteractionSites improves its accuracy to 68% from a baseline of around 30% (Ofran and Rost, 2007). These results suggest that inclusion of predicted structural information can improve the accuracy of binding residue prediction.

For protein level prediction, there are two possible ways to obtain per protein prediction. The first way is simply to infer from per-residue prediction. Technically, a protein is defined as a binding protein if there is any residue on the protein which is predicted as binding residue by per-residue method. The second way is to use protein level specific methods.

The important and most crucial step during classification of proteins using machine learning techniques is to transform the variable length of protein sequence to fixed

length feature vectors. DNAbinder, which is a DNA binding protein prediction method, transforms position-specific scoring matrix (PSSM) to PSSM-400 vector. PSSM-400 is the composition of occurrences of each type of amino acid corresponding to each type of amino acids in protein sequence, which means for each column there will be 20 values instead of one. Hence, it will be a vector of dimension 20 × 20 for each PSSM matrix (Kumar et al., 2007).

StackDPPred is also a DNA binding protein prediction method (Mishra et al., 2019). To encode protein sequence with a fixed dimensional feature vector, they applied various feature extraction techniques based on the PSSM profile: PSSM-distance transformation (PSSM-DT), Residue probing transformation (RPT) and Evolutionary distance transformation (EDT). PSSM-DT results in two kinds of features: PSSM distance transformation of pairs of same amino acids (PSSM-SDT) and PSSM distance transformation of pairs of different amino acids (PSSM-DDT) (details can be seen in (Mishra et al., 2019)). PSSM-SDT calculates the occurrence probabilities for the pairs of the same amino acids separated by a distance k along the sequence. PSSM-DDT calculates the occurrence probabilities for pairs of different amino acids separated by a distance of k along the sequence (Mishra et al., 2019).

$$PSSM - SDT(j,k) = \sum_{i=1}^{L-k} P_{i,j} * P_{i+k,j}/(L-k)$$

where, j is one type of the amino acid, L is the length of the protein sequence, $P_{i,j}$ is the PSSM score of amino acid j at position i and $P_{i+k,j}$ is the PSSM score of amino acid j at position i + k. Through this approach, 20*K number of PSSM-SDT features are generated, where K is the maximum range of k (k = 1, 2, …, K).

$$PSSM - DDT(i_1, i_2, k) = \sum_{i=1}^{L-k} P_{j,i_1} * P_{j+k,i_2}/(L-k)$$

where, i_1 and $i_2$represent two different types of amino acids.

RPT, proposed by (Jeong et al., 2011),emphasizes domains with similar conservation rates by grouping domain families based on their conservation score in the PSSM profile. And the EDT extracts the information of the non-co-occurrence probability for two amino acids separated by a certain distance in a protein from the PSSM profile (Mishra et al., 2019).

So far, there are some methods which can conduct multiple class prediction. And it can benefit a lot from establishing an all-in-one system. Many methods may not have constant performance due to the different training data they used. For example, the cutoff which is used to define binding residue ranges from 3.5Å to 6Å (Yan et al., 2016). Some use 3.5Å, and the others may use 5Å or 6Å. It has been found that changing the cutoff value will change the performance significantly (Yan et al., 2016).

DRNApred is a method which can predict both DNA and RNA binding residues (Yan and Kurgan, 2017). DRNApred uses a lot of features including a variety of physicochemical and biochemical properties together with hidden Markov model (HMM) based evolutionary profile and predictes intrinsic disorder, secondary structure and solvent accessibility (Yan and Kurgan, 2017).

hybridNAP is the first method which can predict all three classes of binding residues: protein-protein, protein-DNA and protein-RNA (Zhang et al., 2019). And their results suggest that development of the new generation of predictors would benefit from using training data sets that combine all the three protein-, RNA- and DNA-binding proteins and pursuing combined prediction of protein-, DNA- and RNA-binding residues (Yan et al., 2016; Zhang et al., 2019).

DisoRDPbind is another method which can predict all three kinds of binding residues (Peng et al., 2017). DisoRDPbind uses the features such as predicted secondary structure, intrinsic disorder predicted by IUPred (Dosztanyi, 2018), amino acid composition and physiochemical properties of amino acids (Peng et al., 2017). However, there is a limitation for DisoRDPbind. Unlike hybridNAP which can provide general binding residues prediction, DisoRDPbind is designed specifically for the binding prediction on the disorder region. Thus, it has very bad performance on general predictions (Qiu et al., 2020).

As there are already many tools which can predict binding protein or residues, the reasons why it is still necessary to establish the new method in this thesis are as following: 1) Previous review has already found that most binding prediction methods are only available through web servers. However, many of them are either no longer maintained or only transiently online (Yan et al., 2016). Furthermore, it will also negatively affect consensuses that rely on the web server calculations. Thus, unsustainable or short maintenance is one of the challenges for bioinformatics.

PredictProtein server (Yachdav et al., 2014), in which the binding prediction method in this thesis is available, went online as one of the first Internet servers in molecular biology in 1992. Now PredictProtein has already served for almost 30 years. 2) Though methods such as hybridNAP can predict multiple classes of binding residues, so fa, there is no comprehensive system which integrates both the protein level and the residue level prediction. However, a protein level prediction can significantly improve the residue level prediction when the users are not sure whether the input proteins are binding protein or not, for example, in high-throughput analysis. And again, an all-in-one system could have a more constant performance than a combination of many separate ones. 3) Unlike previous studies which heavily depend on evolutionary information, in this thesis, some new techniques such as neutral language processing are applied.

1.4 Conclusion

Protein-, DNA-, RNA-protein binding proteins and residues play important role in many biological processing. And SAVs, the majority of genetic variants, are the genome differences that make each person's genome unique, some of which will lead to serious phenotype such as disease. So it is meaningful to conduct an analysis of those SAVs occurring on the binding proteins and residues. However, experimental and structure-based binding proteins/residues identification methods are not suitable for high-throughput research. Thus, in this thesis, we first develope a sequence-based Protein-, DNA-, RNA-protein binding proteins and residues prediction method which outperforms previous methods. And we further apply our method to analyzing SAVs from 60,706 people.

Chapter 2

2 Sequence-based Protein-, DNA- and RNA-binding prediction system

In this section, we will discuss our new sequence-based comprehensive binding prediction system (ProNA2020). It is a two-level prediction. At first level, the protein level, it can predict whether the input protein is a binding protein or not. If the input protein is predicted as a binding protein, then at the second level, the residue level, it can further predict the binding residues on the input protein.

2.1 Methods

2.1.1 5-fold cross validation

In this thesis, we use a 5-fold cross validation approach (Figure 2.1). Basically, the training data is divided into 5 parts (the details of data preparation are shown in the journal article at the end of this section). Every time, three parts serve as training set which are used to train the model, and one part serves as cross-training set which is used to select features and optimize the hyperparameters such as number of

hidden nodes and learning rate, and the rest one part is the test part which is used to evaluate the final performance of the model.

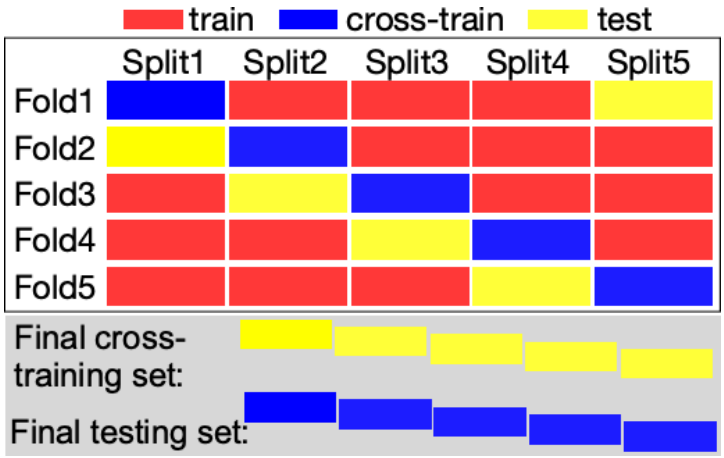


Figure 2.1: Cross-validation procedure. The original non-redundant training data is split into five splits (Split1-Split5). Three splits are used for training, one for cross-training, one for testing. This process is repeated five times (5-fold cross-validation).

2.1.2 Profile kernel

Profile kernel is a kind of kernel function of support vector machine (SVM). An SVM is a supervised machine learning model that uses classification algorithms for two-group classification problems. The target of SVM is to find a decision boundary (also known as the hyperplane), which can separate two groups of samples from one or more feature vectors. And this hyperplane is a straight line and the distance from it to the nearest data point on each side (red nodes and blue nodes in Figure 2.2) is maximized (maximum-margin).

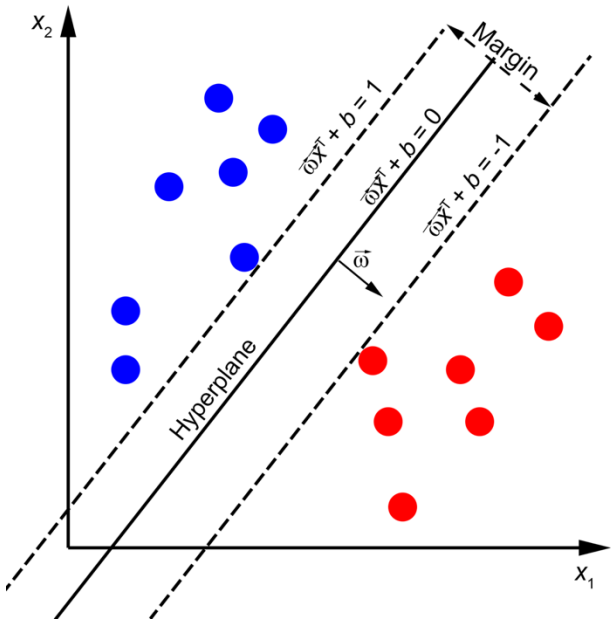


Figure 2.2 Linear SVM model. Classification between blue and red samples. To separate two groups of samples, SVM will find a hyperplane with the maximum margin.

Given a labeled training dataset:

$$(\vec{x}_1, y_1), \cdots, (\vec{x}_n, y_n), \vec{x}_i \in R^d \; and \; y_i \in (-1, +1)$$

where $\vec{x}_i$ is a feature vector representation and y_i is the class label (either 1 or -1) of a training sample i. Any hyperplane can be defined as:

$$\vec{\omega}\vec{x}^T + b = 0$$

where $\vec{\omega}$ is the weight vector, $\vec{x}$ is the input feature vector, and b is the bias.

For the linearly separable data, there are two parallel hyperplanes (two dashed lines in Figure 2.2) which can separate the two groups of data, so that the distance between them is as large as possible. The “margin” is the region bounded by these two parallel hyperplanes, and the maximum-margin hyperplane is the hyperplane that lies halfway between them. The above two hyperplanes can be described by:

$$\vec{\omega}\vec{x}^T + b = 1$$

anything on or above this hyperplane belongs to one class (blue nodes). And

$$\vec{\omega}\vec{x}^T + b = -1$$

anything on or below this hyperplane belongs to one class (red nodes).

And the $\vec{\omega}$ and b would satisfy the following inequalities for all samples in the training data:

$$\vec{\omega}\vec{x_i}^T + b \geq 1 \; if \; y_i = 1$$

$$\vec{\omega}\vec{x_i}^T + b \leq 1 \; if \; y_i = -1$$

The distance between these two hyperplanes is $\frac{2}{\|\vec{\omega}\|}$. Thus, the objective of SVM is to maximize the distance between two hyperplanes which means minimizing $\|\vec{\omega}\|$.

The SVM is originally designed for linear classifier. For non-linear problem, there is an alternative use for SVM called kernel method. A kernel function can make it easier to calculate the inner product of two feature vectors in higher dimensional space, so as to transform a non-linear problem to a linear problem (Figure 2.3).

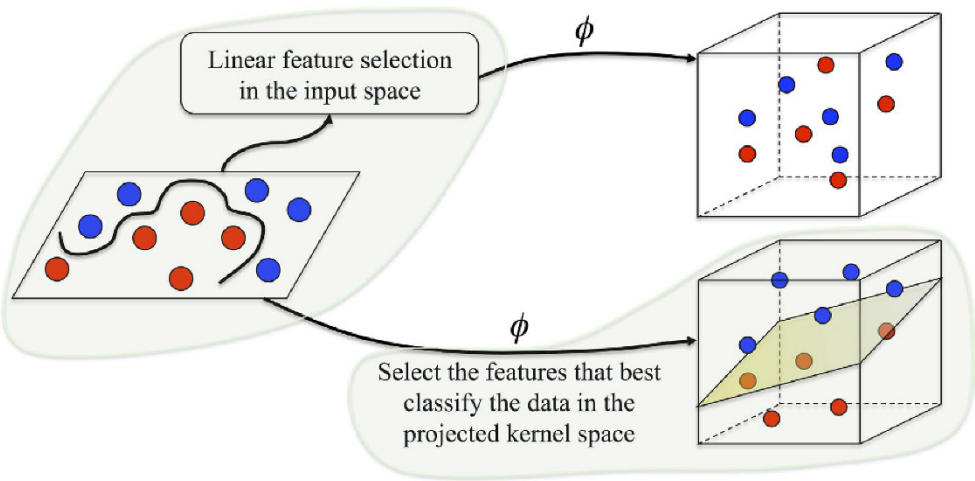


Figure 2.3: Introduction of kernel function (Adeli et al., 2017). Classification is between blue and red sample. It is not possible to find a hyperplane in linear feature

space. Then with a suitable kernel function φ, a hyperplane can be found in higher dimensional space.

Given K as the kernel function:

$$K(x,y) =< f(x), f(y) >$$

Where x, y are n dimensional inputs. f is a function used to map the input from n dimensional to m dimensional space. With the kernel functions, it is possible to compute the scalar product between two sample points in a higher dimensional space without explicitly mapping the data point into higher dimensional space.

Profile kernel is a kind of kernel function for SVM. The original profile kernel has been introduced in (Kuang et al., 2005) and, in this thesis, an accelerated version of profile kernel from our lab is used (Hamp et al., 2013).

Technically, the profile kernel uses probabilistic profiles, such as PSSM matrix produced by the PSI-BLAST algorithm, to define position-dependent mutation neighborhoods along protein sequences for inexact matching of k-length subsequences ("k-mers") (Kuang et al., 2005).

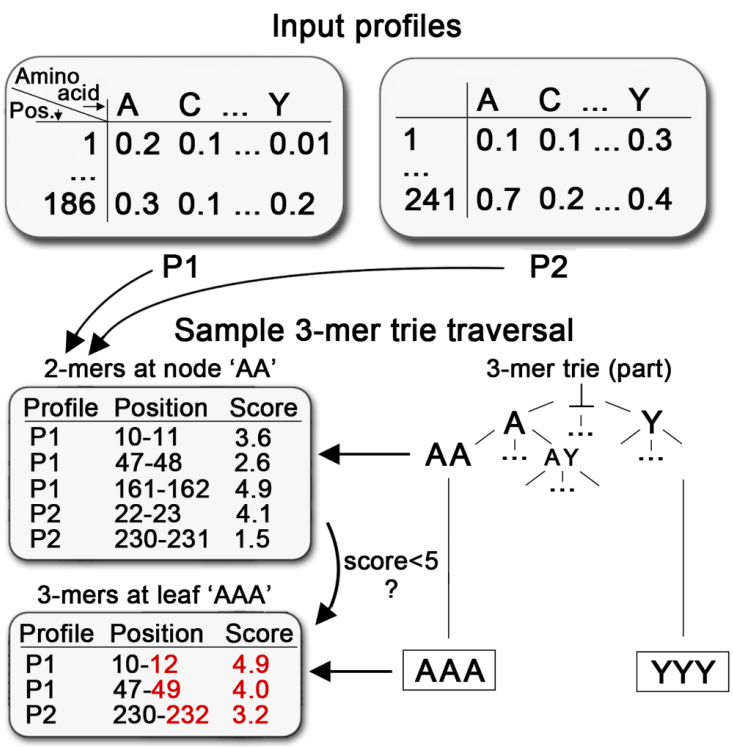


Figure 2.4 Introduction of profile kernel (Hamp et al., 2013). This shows how profile kernel is calculated with two input profiles: P1 and P2. These two profiles are generated from proteins that are 186 (P1) and 241 residues long (P2; tables on the top). In profile calculation, it counts the number of conserved multi-mers at each node that fall below the substitution score threshold σ. Here is an example of 3-mer with a threshold σ=5. At each node, profile-kernel counts the number of 3-mer motif (such as "AAA") on the protein with a score below 5. And technically, using 3-mer means mapping protein onto a 20*20*20 (8000) dimensional vector.

Here is an example which explains the process of profile kernel calculation (Figure 2.4) (Hamp et al., 2013). At first, two blast profiles (such as PSSM matrix) are generated (two tables on the top of Figure 2.4). Then, there are two important parameters in profile kernel: *k-mer* and σ. *k-mer* indicates how many consecutive residues are taken into consideration in profile kernel, and σ is the threshold for conservation score. Figure 2.4 is an example for *3-mer* and σ is set to be 5. Instead of using the conservation score of single residues in original profile, the conservation is now calculated as the sum of the scores for 3 consecutive residues. Thus, *3-mer* means that it maps the profile to a 20^*k*-dimensional vector of integers. Each dimension represents one *combination* of *k* consecutive residues and a value gives the number of times this *k-mer* combination is conserved (conservation score below σ) in a profile of the corresponding proteins (Hamp et al., 2013).

2.1.3 Word2Vec

Artificial neural networks (ANN), which are inspired by the biology neural networks, are widely used in machine learning. The basic component of neural network is neurons which is also referred to as perceptron. The simplest neural network consists of just one perceptron, which receives and sums up the input signal and evaluates this sum using a threshold function (activation function), which produces the output value. The following formula describes how the input signals are summed up with their weights:

$$Sum(s) = \sum_{j=1}^{n} i_j * w_j$$

And for activation functions, there are a lot of functions available, such as the widely used sigmoid function which can normalize the input value to be between 0 and 1:

$$Sigmoid(s) = \frac{1}{1 + e^{-1}}$$

And a schematic of basic ANN is depicted in Figure 2.5.

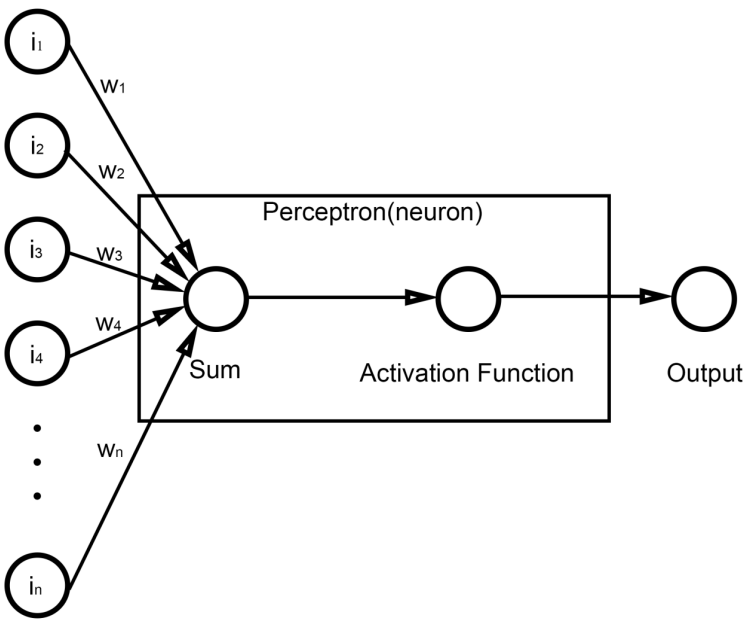


Figure2.5: A schematic of basic ANN component (neuron). The perceptron (neuron) is represented by rectangle. It receives inputs i from different input perceptrons and then sums up the signal. The activation function uses the sum as input and calculates the output of the perceptron.

The basic version of ANN is able to solve simple linear classification problem. However, in complicate non-linear classification such as the binding prediction in this thesis, because of much bigger feature vectors and overlapping data points, it is necessary to use more complex ANN which contains multiple neurons. In most application, an ANN consists of three layers (Figure 2.6). The first layer is called input layer which contains as many nodes as the length of the input feature vector is. There is no calculation at this layer, and it just passes the information to the second layer which is called hidden layer. The hidden layer consists of hidden nodes, all of which are perceptrons. The final layer is the output layer which presents the quantity of output classes.

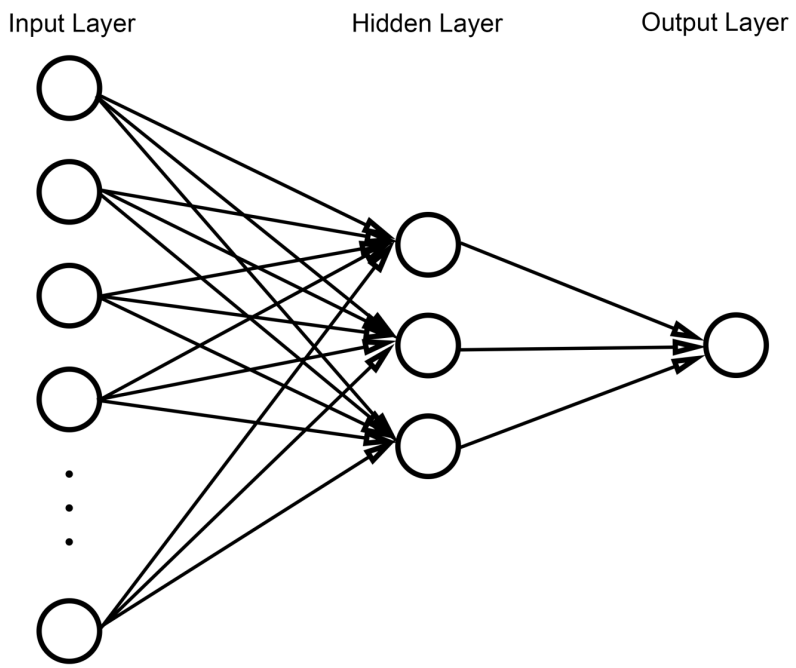


Figure 2.6: Fully connected feed forward network. There are connections between every node in input layer and that in hidden layer, and also between the nodes in hidden layer and that in output layer. This kind of network topology is called a fully connected feed forward network.

Word2Vec is a group of ANNs, which are used to produce word embeddings. It was developed by Tomas Mikolov in 2013 at Google (Mikolov et al., 2013). Word embedding, which can represent words by vectors, is one of the most popular representation of document vocabulary.

There are two different kinds of ANNs in Word2Vec which are trained for certain tasks: CBOW and Skip-gram (Figure 2.7). Assuming a window approach with size 5 (2 on each side), CROW uses the surrounding words to predict the probability for every word in the vocabulary of being the “central word” in the window approach. However, Skip-gram type uses the word in the middle to predict the probability for every word in the vocabulary of being the neighbors in the window approach.

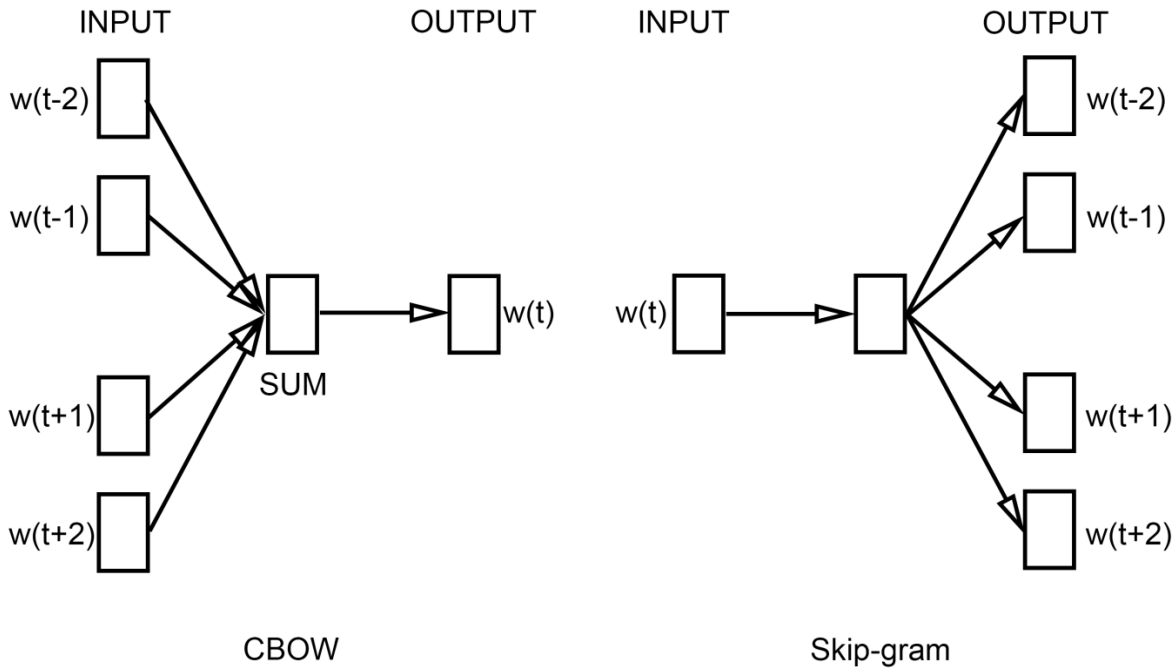


Figure 2.7: Two different kinds of Word2Vec neural network: CBOW and Skip-gram. The difference between CBOW and Skip-gram neural network is: the CBOW model uses the distributed representations of neighbor words to predict the word in the middle. While the Skip-gram model uses the distributed representation of the input word to predict the surrounding words.

In the thesis, we used a skip-gram neural network of Word2Vec. To train the Word2Vec model, the first step is to collect the samples. Here, we assume the source text is “The quick brown fox jumps over the lazy dog”. Then, a window approach with a certain size (size=5 in Figure 2.8) goes through the context sentence and picks up the pairs of

central work and its neighbors in the window (Figure 2.8). The central words will serve as inputs for the network and the neighbors will be the targets.

Then, we can set the neural network (Figure 2.9). It will have three layers:1) input layer. The input is the one-hot vector for the input word;2) hidden layer. There is no activation function on the hidden layer neurons, and, as an example, here we set the number of hidden nodes to be 4 (Figure 2.9); 3) output layer. It has nine neurons with softmax activation function which represent the probability distribution of words (Figure 2.9). The basic idea of skip-gram network is to learn the statistics from the number of times each sample pair shows up. Thus, the softmax output layer shows which words in the vocabulary have the higher possibility to be the neighbors of the input word.

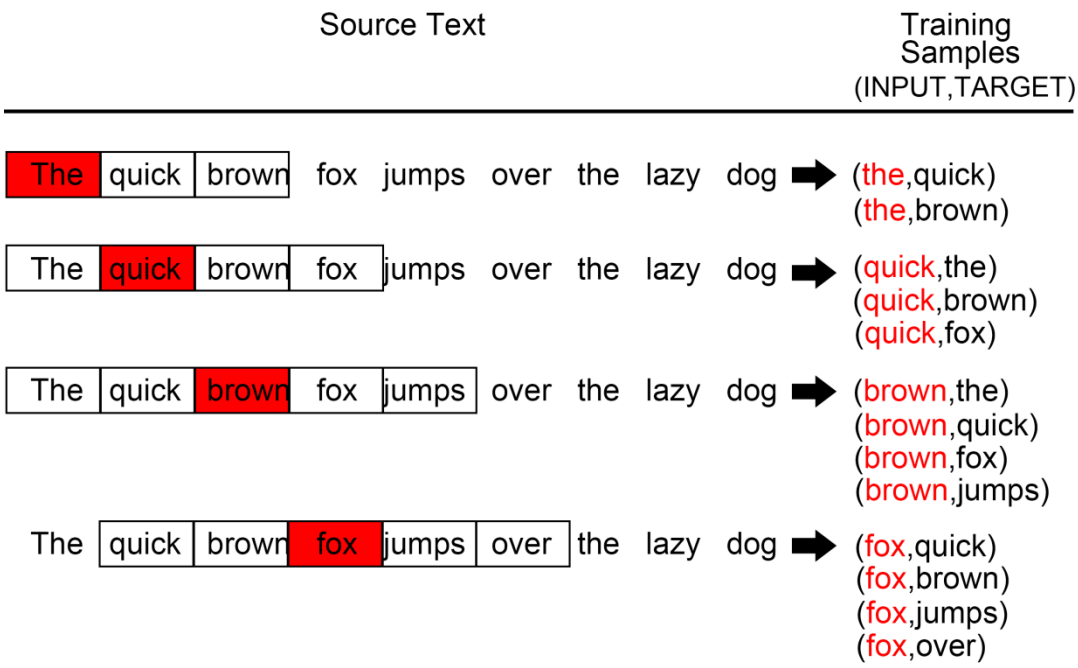


Figure 2.8: Sample preparation of Word2Vec. Assuming the source text is “The quick brown fox jumps over the lazy dog”, a window approach with size 5 goes through the sentence and picks up the pairs of samples: central word and its neighbor words.

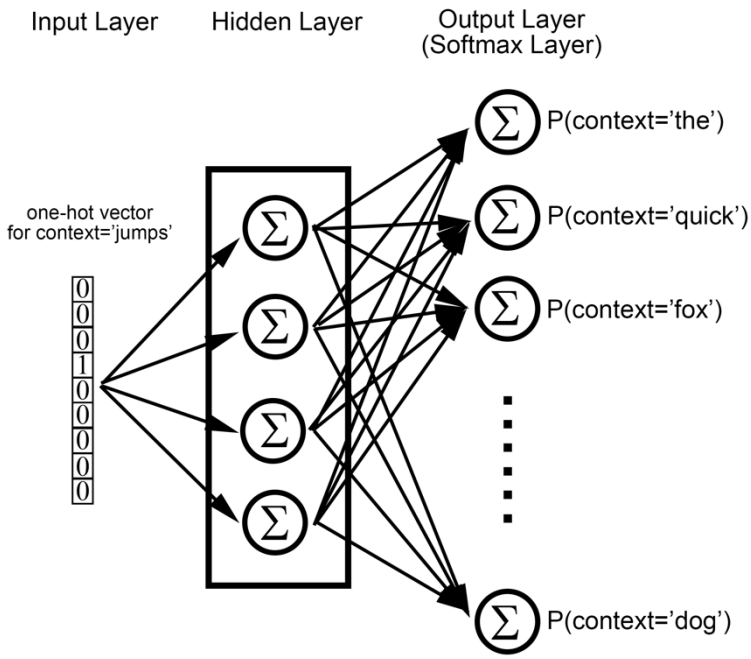


Figure 2.9: Architecture for skip-gram model. The output of the neural network is a softmax layer which shows the the probabilities of each words in the corpus to be the neighbor words of the input word. And weight matrix of the hidden layer is what we need for next step of Word2Vec (here we uses 4 neurons as an example).

After training the network, instead of the network itself, what we need is only the weight matrix in the hidden layer. In this example, since there are 4 neurons in the hidden layer and 9 words in the vocabulary, it is a 9x4 matrix (Figure 2.10). The final word vector can be produced through multiplying the one-hot vector for the input word by the weight matrix (Figure 2.10). And the length of word vector will simply equal to the number of neurons in the network.

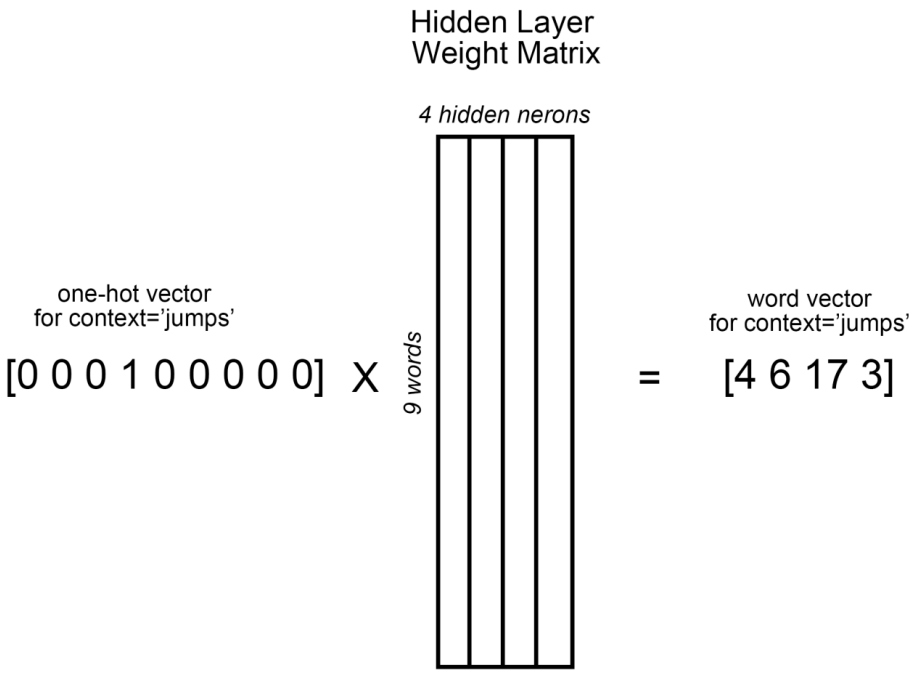


Figure 2.10: Producing word vector by Word2Vec. Using the hidden layer weight matrix with 4 hidden nodes from the network in Figure 2.8, Word2Vec is able to calculate the final representation of the input word.

In our study, the resource contexts are all the protein sequences from UniProt database (The UniProt, 2017). To train the representations for proteins, we need to break the protein sequences into sub sequences so that we can define the "biological words". N-grams is the widely used technique in bioinformatics to study protein sequences. Normally, an overlapping window approach is applied in n-gram modeling of protein research. In this thesis, instead of the window approach, we generate n lists of shifted non-overlapping words (Figure 2.11 shows an example of 3-grams) (Asgari and Mofrad, 2015). So in Figure 2.11, 3 consecutive residues are considered to be a 'biological word'. For a certain protein sequence, all the possible "biological words" and their neighbors are used to train the word2vec skip-gram neural network which we talk about above. And parameter n is determined through cross-validation. The final representation of each protein sequence in our training set is produced by concatenating the vector representation of every possible "word" (n consecutive residues) on the protein sequence.

Orignal sequence

$^{(1)}$M$^{(2)}$F$^{(3)}$RTKRSALVRRLWRSRAPG...

Splittings

(1)MFR,TKR,SAL,VRR,LWR,SRA,...
(2)FRT,KRS,ALV,RRL,WRS,RAP,...
(3)RTK,RSA,LVR,RLW,RSR,APG,...

Figure 2.11: Protein sequence splitting with 3-grams. To prepare the training sample for the word2vec skip-gram neural network, each protein sequence is represented as three sequences (1, 2, 3) of 3-grams and 3 consecutive amino acids is a “biological word”.

2.1.4 ANN for residue level prediction

For residue level prediction, we used ANN with the features from PredictProtein (Yachdav et al., 2014). The PredictProtein (PP) server is an automatic service that searches up-to-date public sequence databases, creates alignments, and predicts aspects of protein structure and function (Yachdav et al., 2014).The features include：

PSSM, which is calculated out of a multiple sequence alignment against big_80 database. Big_80 is a redundancy-reduced (at 80% threshold) database which concatenates UniProt and PDB together (Burley et al., 2017; The UniProt, 2017).

Predicted secondary structure and solvent accessibility. Secondary structure is predicted by a system of neural networks with three states helix, strand and loop rating at an expected average accuracy of 72% (Rost and Sander, 1993). The solvent accessibility is another important feature for binding residue prediction. Those residues on the surface of a protein which have better accessibility are more likely to be the binding residues. And solvent accessibility is predicted by a neural network method rating at a correlation coefficient (correlation between experimentally observed and predicted relative solvent accessibility) of 0.54 (Rost and Sander, 1994).

B-value, which describes the mobility of residues. Functional residues such as binding residues usually show a larger mobility than non-functional (non-binding) residues. In PredictProtein, B-value is predicted by PROFbval (Schlessinger et al., 2006).

Other features: protein length, amino acid composition and physical properties of amino acids. Table 2.1 and Table 2.2 show the details of the features we used.

For the architecture of the neural network, we used a classic three-layer network: one input layer, one hidden layer and one output layer (Figure 2.12). Specially, there are two nodes with sigmoid function at the output layer: one for binding prediction and one for non-binding prediction. So, the raw output score of the neural network will be:

$$score_{raw} = node_{binding} - node_{non-binding}$$

Besides, a second level filter is applied. Instead of the raw prediction of single residue, we use a window approach which takes neighbor residues into consideration:

$$score_{final} = \frac{1}{w} \sum_{i=-\frac{w-1}{2}}^{\frac{w-1}{2}} score_{raw\ i}\ , (score_{raw\ i} > 0)$$

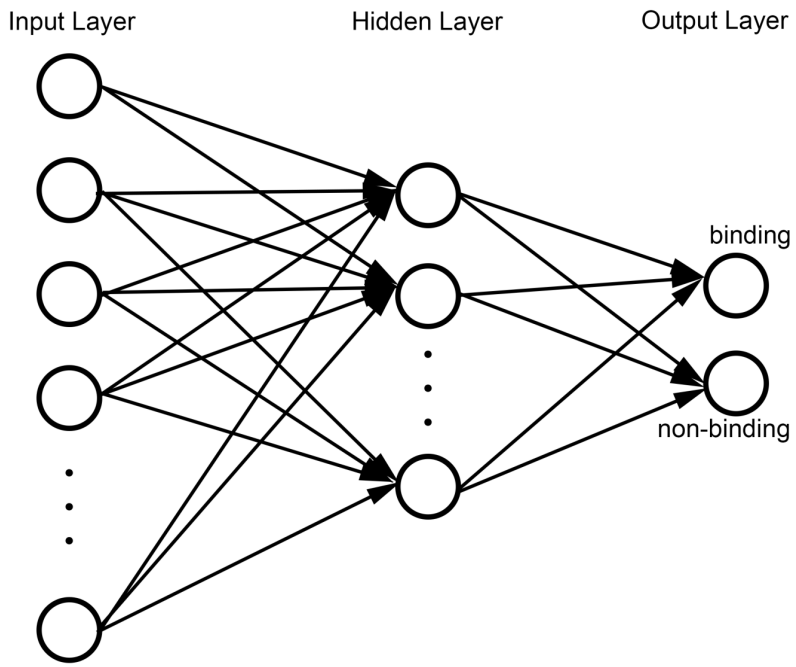


Figure 2.12: Architecture of ANN used in residue level prediction. It is a three-layer network. Specially, we set two nodes with sigmoid function at output layer: one for binding prediction and the other for non-binding prediction.

2.1.5 Performance evaluation

We applied the standard metrics with the acronyms (TP: true positives: observed and predicted in class C; TN: true negative: observed and predicted in non-C; FP: false positives: predicted in C, observed in non-C; FN: false negatives: predicted in non-C, observed in C):

PRE(C)=PrecisionC=TP/(TP+FP)

REC(C)=RecallC=TP/(TP+FN)

Q2=(TP+TN)/(TP+TN+FP+FN)

F1(C)=2*PRE(C)*REC(C)/(PRE(C)+REC(C))

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

Table 2.1: Input features for protein binding per-residue binding predictions

Name	*Window size (number of residues)*	*Description*
pssm	11	evolutionary Profile: normalized absolute conservation of aa at specific positions
infPP	11	information per position: information content of specific position in PSSM and PERC
helix	11	helix predicted
loop	11	loop predicted
strand	11	strand predicted
md_raw	11	raw disorder prediction score
md_minus	11	no disordered region predicted
md_plus	11	intrinsically disordered region predicted
profbval_raw1	11	raw residue flexibility score
profbval_raw2	11	raw residue non flexibility score
b	11	buried predicted
e	11	exposed predicted
i	11	intermediate predicted
composition	1	relative occurrence of an AA in the entire sequence
length_category1[1]	3	length category 1-60 aa
length_category2	3	length category 61-120 aa
length_category3	3	length category 121-180 aa
length_category4	3	length category 181- aa
chemprop_hbreaker[2]	3	aa is a helix breaker
chemprop_mass[2]	3	mass of the amino acid
chemprop_vol[2]	3	volume of the amino acid (size)
chemprop_cbeta[2]	3	aa is a c-beta branching aa
chemprop_charge[2]	3	charge in 3 states
chemprop_hyd[2]	3	hydrophobicity of the amino acid
position	3	position of aa in protein sequence

[1] For the protein with a length smaller than 60, length_category1 is 0.5. Otherwise, length_category1 is 1.

[2] chemprop_mass and chemprop_vol were taken from http://prowl.rockefeller.edu/aainfo/contents.htm; chemprop_hyd was from Kyte-Doolittle (e.g. http://en.wikipedia.org/wiki/Hydropathy_index); chemprop_cbeta was according to http://www.russell.embl-heidelberg.de/aas/cbb.html; chemprop_hbreaker (helix breaker) was proline; chemprop_charge was according to side chain charge

Table 2.2 Input features for DNA/RNA binding per-residue binding predictions

Name	*window size*	*Description*
pssm	11	evolutionary profile: normalized absolute conservation of aa at specific positions
infPP	9	information per position: information content of specific position in PSSM and PERC
relW	5	relative weight: information content of specific positions on PSSM and PERC
md_raw	11	raw disorder prediction score
md_ri	9	disorder prediction reliability score
profbval_raw1	5	raw residue flexibility score
profbval_raw2	5	raw residue non flexibility score
helix	11	helix predicted
loop	11	loop predicted
strand	7	strand predicted
OtE	9	raw prediction output of Sheet
OtL	9	raw prediction output of Loop
OtH	9	raw prediction output of Helix
ri_sec	11	reliability index of secondary structure prediction, applies to helix, sheet, loop and OtE, OtH, OtL
b	7	buried predicted
e	7	exposed predicted
i	7	intermediate predicted
Rel_acc	11	predicted relative solvent accessibility in %
Ri_acc	9	reliability index of solvent accessibility prediction: applies to e,i,b and rel_acc
chemprop_hyd[1]	7	hydrophobicity of the amino acid
chemprop_charge[1]	3	charge in 3 states
chemprop_mass[1]	9	mass of the amino acid
Exposed_composition3	1	for each buried, intermediate, exposed the relative occurrence is given in 3 categories with each 3 states
buried_composition3	3	for each buried, intermediate, exposed the relative occurrence is given in 3 categories with each 3 states
intermediate_composition3	1	for each buried, intermediate, exposed the relative occurrence is given in 3 categories with each 3 states
Helix_composition2	1	the relative occurrence of helix is given in 3 categories
composition	1	relative occurrence of an AA in the entire sequence

[1] chemprop_mass and chemprop_vol were taken from http://prowl.rockefeller.edu/aainfo/contents.htm; chemprop_hyd was from Kyte-Doolittle (e.g. http://en.wikipedia.org/wiki/Hydropathy_index); chemprop_cbeta was according to http://www.russell.embl-heidelberg.de/aas/cbb.html; chemprop_hbreaker (helix breaker) was proline; chemprop_charge was according to side chain charge

2.2 Results and discussion

For protein level prediction, we use a combination of two distinct algorithms: 1) the sequence alignment-based profile-kernel and 2) neutral language based Word2Vec. In our study, we find that profile-kernel are better at predicting the proteins from large protein families that have more alignment from blast, while Word2Vec has a higher performance for the proteins from small families. Thus, the combination can make them benefit from each other.

After establishing the protein level mode in training data, we compare the performance of our method with other state of art algorithms. First, our method (ProNA2020) outperforms all other methods in predicting binding proteins of all three classes: protein-binding, DNA-binding and RNA binding (Table 2.3).

Besides the specific protein-level methods, residue level prediction method can also be used in protein level prediction. Basically, we just define the proteins holding at least one predicted binding residue as the binding proteins. We find residue level methods tend to predict almost all input proteins as binding proteins (Table 2.3). This makes them have very high recall, but low precision. These results approve that it is necessary to develop the specific protein level method since the residue level methods are not suitable to predict protein level binding.

For residue level prediction, we use the classic ANN with a lot of features from PreidctProtein server such as predicted secondary structure and solvent accessibility. We compare our method in two different ways: unknown mode (Table 2.4) and known mode (Table 2.5). Unknown mode means, for a query protein Q, it is not known whether it binds DNA/RNA/Protein. And known mode means only binding proteins are included in the performance comparison. For example, when assessing the performance of DNA binding residues prediction, we only use DNA binding proteins for known mode. But, for unknown mode, non-binding proteins are also included together with the binding proteins. In known mode comparison which is based on only binding proteins , our method (ProNA2020) has the higher MCC and F1 than others (Table 2.5).However, in high-throughput analysis, the input proteins are not limited to the binding proteins, and actually most of the inputs will be non-binding proteins. Thus, the results in unknown mode which mixes the binding and non-binding proteins should be

more close to the performance in real practice (Table 2.4). In unknown mode comparison, besides MCC and F1, our method (ProNA2020) also has the highest Q2 accuracy in all three tasks: 83±1% for DNA binding residues prediction; 88±2% for RNA binding residues prediction and 75±3% for protein binding residues prediction (Table 2.4). All these results indicate ProNA2020 should so far be the best binding residues prediction method, especially for high-throughput analysis. And for the availability, besides the source code on github, ProNA2020 can also be used through PredictProtein server (Figure 2.13).

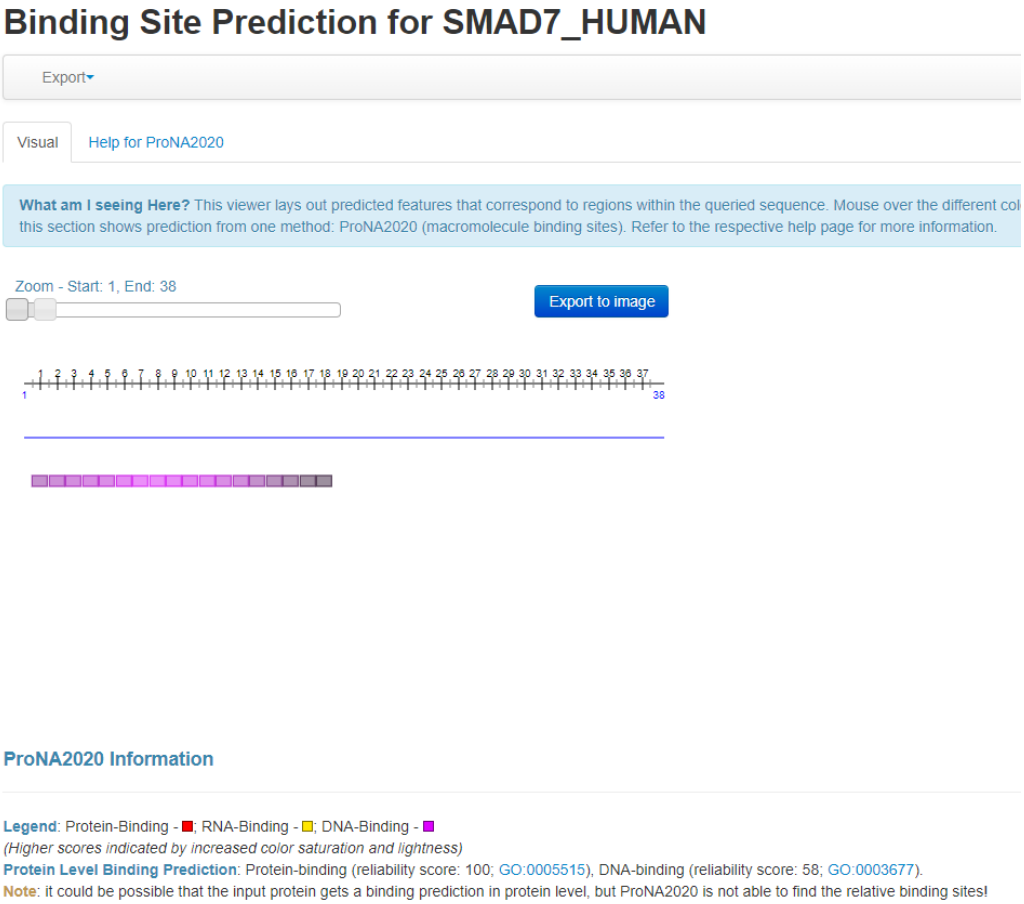


Figure 2.13: ProNA2020 on PredictProtein server. The protein level predictions are given with GO annotations and reliability score. And the predicted binding residues are assigned with colored rectangle and the color saturation and lightness correspond to the reliability of the predictions (the higher the saturation, the reliable the prediction).

Table 2.3: Per-protein performance for independent test set

Method	*Binding*	*Q2(%)*	*PRE(%)*	*REC(%)*	*F1(%)*	*MCC*
DisoRDPbind(Peng and Kurgan, 2015)[1]	DNA	54±3	47±4	78±4	59±3	0.17±0.06
DRNApred(Yan and Kurgan, 2017)[1]		49±3	44±4	83±4	57±3	0.08±0.06
hybridNAP(Zhang et al., 2017)[1]		42±3	42±3	100	59±3	0
NucBind(Su et al., 2019)[1]		49±3	45±3	**99±1**	62±3	0.21±0.04
DNAbinder(Kumar et al., 2007)		62±3	53±4	81±3	64±3	0.31±0.06
DNABIND(Szilagyi and Skolnick, 2006)		59±3	50±4	61±5	55±4	0.17±0.06
SomeNA(Hönigschmid, 2012)[1]		42±3	42±3	**99±1**	59±3	0.02±0.06
StackDPPred(Mishra et al., 2019)		67±3	57±3	90±3	70±3	0.42±0.05
ProNA2020		**77±3**	**67±4**	77±3	**76±3**	**0.56±0.05**
DisoRDPbind(Peng and Kurgan, 2015)[1]	RNA	36±3	22±3	77±6	35±4	0.02±0.06
DRNApred(Yan and Kurgan, 2017)[1]		45±3	25±3	60±6	32±3	0.007±0.06
hybridNAP(Zhang et al., 2017)[1]		22±3	22±3	**100**	36±3	0
NucBind(Su et al., 2019)[1]		34±3	24±3	91±4	38±4	0.11±0.05
RBPPred(Zhang and Liu, 2017)		59±3	29±4	61±6	39±5	0.16±0.07
RNABindRPlus(Walia et al., 2014)[1]		25±3	23±3	**100**	37±3	0.10±0.02
SomeNA(Hönigschmid, 2012)[1]		34±3	24±3	98±1	38±4	0.15±0.03
SPOT-RNA(Yang et al., 2014)		**79±3**	**54±5**	33±6	41±5	0.31±0.06
TriPepSVM(Bressin et al., 2019)		77±3	49±6	61±6	54±5	0.40±0.06
ProNA2020		72±3	43±5	82±5	**57±5**	**0.44±0.05**
DisoRDPbind(Peng and Kurgan, 2015)[1]	Protein	50±3	91±3	41±3	57±3	0.21±0.05
hybridNAP(Zhang et al., 2017)[1]		**80±3**	80±3	100	**89±2**	0
BSpred(Mukherjee and Zhang, 2011)[1]		**80±3**	80±3	100	**89±2**	0
CRF-PPI(Wei et al., 2015)[1]		**80±3**	80±3	100	**89±2**	0
InteractionSites(Ofran and Rost, 2007)[1]		**80±3**	80±3	100	**89±2**	-0.04±0.01
iPPBS-PseAAC(Jia et al., 2016)[1]		**80±3**	80±3	100	**89±2**	0
LORIS(Dhole et al., 2014)[1]		**80±3**	80±3	100	**89±2**	0
PPIS (Liu et al., 2016)[1]		**80±3**	80±3	100	**89±2**	0
SPRINGS (Gurdeep Singh, 2014)[1]		**80±3**	80±3	100	**89±2**	0
SSWRF-PPI(Zhi-Sen Wei, 2016)[1]		**80±3**	80±3	100	**89±2**	0
ProNA2020		**80±3**	**82±3**	96±1	**89±2**	**0.22±0.08**

[1] per-residue methods “mis-used” for per-protein prediction

Table 2.4: Per-residue performance for independent test set - mode unknown °

Method	*Binding*	*Q2(%)*	*PRE(%)*	*REC(%)*	*F1(%)*	*MCC*
DisoRDPbind(Peng and Kurgan, 2015)	*DNA*	75±3	34±3	13±2	19±3	0.09±0.02
DRNApred(Yan and Kurgan, 2017)		74±2	36±4	24±3	28±3	0.13±0.03
hybridNAP(Zhang et al., 2017)		64±2	29±3	45±2	35±2	0.12±0.02
NucBind(Su et al., 2019)		70±3	34±9	36±3	35±5	0.16±0.07
SomeNA(Hönigschmid, 2012)		78±2	51±4	39±2	44±2	0.31±0.03
ProNA2020		**83±1**	**60±3**	**59±3**	**60±2**	**0.49±0.02**
DisoRDPbind(Peng and Kurgan, 2015)	*RNA*	80±2	17±5	16±4	15±4	0.05±0.03
DRNApred(Yan and Kurgan, 2017)		78±5	19±5	22±6	21±5	0.08±0.06
hybridNAP(Zhang et al., 2017)		68±3	18±3	**45±4**	26±3	0.11±0.02
NucBind(Su et al., 2019)		67±4	14±4	32±4	20±6	0.03±0.06
RNABindRPlus(Walia et al., 2014)		**88±2**	**56±6**	37±4	45±4	0.40±0.04
SomeNA(Hönigschmid, 2012)		86±3	40±6	16±2	23±2	0.19±0.04
ProNA2020		**88±2**	53±4	40±4	**46±3**	**0.40±0.03**
DisoRDPbind(Peng and Kurgan, 2015)	*Protein*	73±3	23±8	3±1	5±2	-0.03±0.03
hybridNAP(Zhang et al., 2017)		67±2	35±3	38±2	37±2	0.14±0.02
BSpred(Mukherjee and Zhang, 2011)		65±1	22±3	16±1	18±2	-0.04±0.02
CRF-PPI(Wei et al., 2015)		56±1	26±3	40±1	31±2	0.02±0.01
InteractionSites(Ofran and Rost, 2007)		73±3	33±3	9±1	14±1	0.05±0.02
iPPBS-PseAAC (Jia et al., 2016)		70±3	30±2	15±1	20±1	0.04±0.02
LORIS(Dhole et al., 2014)		56±1	25±3	39±1	31±2	0.001±0.007
PPIS (Liu et al., 2016)		55±1	26±3	**42±1**	32±2	0.01±0.01
SPRINGS (Gurdeep Singh, 2014)		56±1	25±3	36±1	32±2	0.004±0.007
SSWRF-PPI(Zhi-Sen Wei, 2016)		57±1	27±3	**42±1**	33±2	0.02±0.01
ProNA2020		**75±3**	**52±3**	36±3	**42±3**	**0.28±0.03**

° Mode-unknown: for a query protein Q, it is **not** known whether it binds DNA/RNA/Protein. Instead, this binding also has to be predicted.

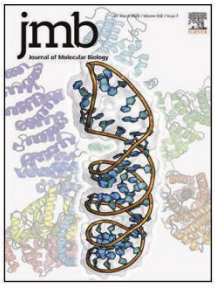
Table 2.5: Per-residue performance for independent test set – mode known °

Method	*Binding*	*Q2(%)*	*PRE(%)*	*REC(%)*	*F1(%)*	*MCC*
DisoRDPbind(Peng and Kurgan, 2015)	DNA	66±2	36±4	13±3	19±3	0.04±0.02
DRNApred(Yan and Kurgan, 2017)		66±2	42±4	24±3	30±3	0.10±0.03
hybridNAP(Zhang et al., 2017)		57±2	36±4	46±2	40±1	0.08±0.02
NucBind(Su et al., 2019)		**78±1**	**86±2**	37±3	52±2	0.47±0.02
SomeNA(Hönigschmid, 2012)		71±2	55±5	39±2	45±2	0.27±±0.04
ProNA2020		**78±1**	65±2	**67±2**	**66±1**	**0.50±0.02**
DisoRDPbind(Peng and Kurgan, 2015)	RNA	71±3	27±4	16±5	20±4	0.04±0.03
DRNApred(Yan and Kurgan, 2017)		69±3	29±3	24±6	26±5	0.07±0.04
hybridNAP(Zhang et al., 2017)		60±3	27±3	45±3	34±2	0.08±0.03
NucBind(Su et al., 2019)		**81±1**	**67±8**	32±4	43±5	0.37±0.05
RNABindRPlus(Walia et al., 2014)		78±1	51±4	**50±3**	**50±3**	0.36±0.03
SomeNA(Hönigschmid, 2012)		77±2	49±1	16±2	25±3	0.17±0.06
ProNA2020		79±2	55±3	45±3	**50±2**	**0.37±0.03**
DisoRDPbind(Peng and Kurgan, 2015)	Protein	66±1	31±10	3±1	5±2	-0.001±0.008
hybridNAP(Zhang et al., 2017)		61±2	41±3	37±2	39±2	0.11±0.02
BSpred(Mukherjee and Zhang, 2011)		60±1	30±2	16±1	20±1	-0.036±0.009
CRF-PPI(Wei et al., 2015)		55±1	34±2	41±1	38±2	0.03±0.01
InteractionSites(Ofran and Rost, 2007)		65±2	42±3	9±1	15±1	0.05±0.02
iPPBS-PseAAC (Jia et al., 2016)		63±1	36±2	15±1	22±1	0.027±0.008
LORIS(Dhole et al., 2014)		54±1	36±2	39±1	36±1	0.005±0.008
PPIS (Liu et al., 2016)		54±2	34±3	**42±1**	38±2	0.02±0.01
SPRINGS (Gurdeep Singh, 2014)		54±1	33±2	37±1	35±2	-0.01±0.008
SSWRF-PPI(Zhi-Sen Wei, 2016)		54±1	34±3	41±1	38±2	0.02±0.01
ProNA2020		**70±2**	**58±3**	39±4	**47±3**	**0.28±0.03**

° Mode-known: for a query protein Q, it is known that it binds DNA/RNA/Protein. For instance, when assessing methods for the DNA per-residue prediction, only DNA-binding proteins are presented.

2.3 Journal article

Jiajun Qiu(JQ) and Burkhard Rost (BR) conceptualized the work. JQ performed the whole analysis and model training. Tomas Norambuena and Francisco Melo helped creating the training data. Michael Bernhofer helped to make the method available online. Michael Heinzinger and Sofie Kemper provided useful suggestion and idea for the research. BR provided supervision. BR provided funding. JQ wrote the initial manuscript draft with BR. All authors reviewed and approved of the final manuscript.



ProNA2020 predicts protein–DNA, protein–RNA, and protein–protein binding proteins and residues from sequence

Jiajun Qiu[1,2], Michael Bernhofer[1,2], Michael Heinzinger[1,2], Sofie Kemper[1], Tomas Norambuena[3], Francisco Melo[3,4] and Burkhard Rost[1,5,6,7]

1 - Department of Informatics, *I12-Chair of Bioinformatics and Computational Biology, Technical University of Munich (TUM), Boltzmannstrasse 3, 85748, Garching, Munich, Germany*
2 - TUM Graduate School, *Center of Doctoral Studies in Informatics and Its Applications (CeDoSIA), Garching, 85748, Germany*
3 - Molecular Bioinformatics Laboratory, *Facultad de Ciencias Biológicas, Pontificia Universidad Católica de Chile, Santiago, Chile*
4 - Institute of Biological and Medical Engineering, *Pontificia Universidad Católica de Chile, Santiago, Chile*
5 - Columbia University, *Department of Biochemistry and Molecular Biophysics, 701 West, 168th Street, New York, NY, 10032, USA*
6 - Institute of Advanced Study (TUM-IAS), *Lichtenbergstr. 2a, 85748, Garching/Munich, Germany*
7 - Germany & Institute for Food and Plant Sciences (WZW) Weihenstephan, *Alte Akademie 8, 85354 Freising, Germany*

Correspondence to Jiajun Qiu: *Fax: +49 (89) 289 19414. jiajunqiu@hotmail.com*
https://doi.org/10.1016/j.jmb.2020.02.026
Edited by Rita Casadio

Abstract

The intricate details of how proteins bind to proteins, DNA, and RNA are crucial for the understanding of almost all biological processes. Disease-causing sequence variants often affect binding residues. Here, we described a new, comprehensive system of *in silico* methods that take only protein sequence as input to predict binding of protein to DNA, RNA, and other proteins. Firstly, we needed to develop several new methods to predict whether or not proteins bind (per-protein prediction). Secondly, we developed independent methods that predict which residues bind (per-residue). Not requiring three-dimensional information, the system can predict the actual binding residue. The system combined homology-based inference with machine learning and motif-based profile-kernel approaches with word-based (ProtVec) solutions to machine learning protein level predictions. This achieved an overall non-exclusive three-state accuracy of 77% ± 1% (±one standard error) corresponding to a 1.8 fold improvement over random (best classification for protein–protein with F1 = 91 ± 0.8%). Standard neural networks for per-residue binding residue predictions appeared best for DNA-binding (Q2 = 81 ± 0.9%) followed by RNA-binding (Q2 = 80 ± 1%) and worst for protein–protein binding (Q2 = 69 ± 0.8%). The new method, dubbed ProNA2020, is available as code through *github* (https://github.com/Rostlab/ProNA2020.git) and through PredictProtein (www.predictprotein.org).

Introduction

Physical interactions between proteins and large DNA, RNA, and proteins crucially determine all essential biological processes, including mechanisms relevant for health and disease [1,2]. The development of new drugs requires detailed molecular understanding of the binding residues [3]. Typically, binding residues are only available through the detailed three-dimensional (3D) structure of a protein. UniProt now (Dec. 2019) contains 179 million protein sequences [4], of which, fewer than 0.36% contain the experimental protein structure data from X-ray crystallography and NMR spectroscopy in the Protein Database, PDB [5], whereas good 3D models of structures are available for fewer than 20% of all the residues of all known proteins [6]. For all of those, binding residues remain largely unknown. However, even knowing which residues are involved in binding without knowing the binding pocket or any details of the 3D structure might already help in designing experiments. Often, it might already help to know that a protein binds to DNA|RNA or other proteins. Despite the pivotal importance of transient physical protein–-protein interactions (PPIs), some important proteins appear not to bind *in vivo* to any other protein [1]. Possibly 6–8% of all proteins in a eukaryote might

bind RNA (**RBPs**: RNA-binding proteins) [7]. For eukaryotes, the fraction of DNA-binding proteins (**DBPs**) appears similar to that of RBPs (6–7%) [8]; for prokaryotes, typically 2–3% of a genome encodes DBPs [8].

Typically, proteins binding other proteins, DNA, or RNA form the targets of structure-based drug design [9]. Understanding protein binding residues becomes a basis for structure-based drug design. Drug molecules usually affect the interaction between the target protein and its ligand [10]. However, fewer than 0.36% of all proteins of known sequence in UniProt correspond to a known experimental 3D structure in the PDB [4,5]. Therefore, it is essential to build computational tools to reliably and rapidly identify protein-, DNA- and RNA-binding proteins or residues.

Given that structure annotations remain missing for most proteins (for >120 million in June 2019), there continues to be a high demand even for low-resolution predictions of aspects pertaining to proteins binding protein, DNA, and RNA from sequence alone. Not surprisingly, many *in silico* methods cater to this need and predict binding proteins (protein binds or not) or binding residues (which residues bind) from sequence. These include (sorted by date) methods optimized for per-protein predictions (protein binds or not) DNABIND [11], SomeNA [12], and StackDPPred [13] for DNA binding, and RBPPred [14], SPOT-RNA [15] and TriPepSVM [16] for RNA binding. Other aspects are provided by tools optimized for per-residue predictions (predicting which residues bind), including some that predict binding for DNA and RNA (sorted by date): DRNApred [17] and NucBind [18], and others capturing all three targets: hybridNAP [19] and DisoRDPbind [20]. The later predicts binding in intrinsically unstructured proteins. However, we are not aware of any existing method combining machine learning prediction and homology-based inference of per-protein and per-residue binding for the three most important large macromolecules (PPI, DNA, or RNA) into one comprehensive system.

Here we present a novel sequence-based system for the comprehensive identification of proteins that bind to protein, DNA, and RNA and the prediction of the residues involved in binding. One crucial novelty of this work is the demonstration that per-protein predictions are performed only very poorly by methods optimized on per-residue predictions, i.e. users need different tools to predict which protein binds a protein, DNA or RNA (per-protein) and where it binds (per-residue) if it does. Toward this end, we also demonstrate how very different machine learning methods can be combined best and how predictions without using evolutionary information may contribute to performance. Another methodological novelty was the embedding of natural language processing (NLP) concepts [21]. Our new system has three major advantages over some existing approaches. Firstly, it combines and assesses per-protein and per-residue prediction in the same framework. All prediction methods are grafted into a common framework although they require very different individual solutions. Secondly, it combines homology-based inference with machine learning (also done by: DisoRDPbind [20]). Thirdly, all the three major macromolecules (protein, DNA, and RNA) are integrated into one hierarchical prediction with sustained performance estimates for the entire system (also done by hybridNAP [19] and DisoRDPbind [20]).

Materials and Methods

Data sets

Reducing sequence redundancy in data sets

For all data sets, UniqueProt reduced redundancy such that no protein pair in the set had sequence similarity of HVAL>0 [22] (e.g. corresponding to 20% pairwise sequence identity for alignments longer than 250 residues) or PSI-BLAST E-value>10^{-3} with the minimum alignment length of 45 residues [22]. Redundancy was reduced to avoid overestimating performance [23].

Data sets for per-residue information (PPI, DNA, and RNA)

DNA-protein binding data was extracted from the Protein–DNA Interface Database (PDIdb, version April 2010 [24]). PDIdb contained 992 entries of proteins with high-resolution 3D structure from the Protein Data Bank (PDB [5] with 1317 different protein chains binding DNA. **RNA-protein** binding data was extracted from the Protein–RNA Interface Database (PRIDB, version RB1179 [25]). PRIDB contained 1179 non-redundant PDB protein chains binding RNA. All PDB entries were mapped to UniProtKB sequences using SIFTS [4,26]. Only 3D structures from X-ray crystallography with resolutions <2.5 Å (0.25 nm) were included; DNA or RNA (in the following NA) interactions were considered only when the closest pair of atoms (between protein and NA) was within 6Å (0.6 nm). **Protein-Protein binding** data was provided by Tobias Hamp [27]. Structures were obtained from PDB (2015) with a resolution of <2.5 Å. After removing all structures from the PPI set mapping to fewer than two different UniProtKB IDs and the proteins with fewer than five residues within 6 Å (0.6 nm) of any atom of the other protein, the protein–protein binding data sets contained 3957 PPIs from 2914 unique proteins representing the species diversity of the PDB. Although reducing redundancy, we maintained alternative binding residues. Assume, A–B (A binds B), A–B′, and EVAL(B,B′)>T, EVAL(A,B)<T, EVAL(A,B′)<T (where T is the threshold for redundancy reduction; EVAL(A,B) the PSI-BLAST Expectation-value, or E-value, for the alignment between A and B). We removed B′ from the data set, but kept the labels of "interacting residues" on A marked by the interaction A–B′. We deliberately did not consider homo-dimers

assuming that they bind in a biophysically different manner from the type of transient physical PPIs that the prediction method targeted [28]. All data sets are available through github (https://github.com/Rostlab/ProNA2020.git); statistics are provided in Tables 1 and 2.

Data sets for per-protein information

Besides the proteins used in per-residue data set, proteins with the experimental annotations were also collected in positive data set for per-protein (described in the next section). Total numbers of non-redundant proteins: protein binding/not binding: 524/282, DNA-binding/not DNA|RNA-binding: 199/555, RNA-binding/not DNA|RNA-binding: 263/555 (Table 2).

GO annotations for negatives (only per-protein)

Due to a variety of reasons, experimentally characterized negatives are rare. To compensate for that, we used GO annotations [29] with experimental evidence codes as proxies for negatives and those used for homology-based inference. We collected proteins with the experimental annotations of protein binding (GO:0005515), DNA-binding (GO:0003677), and RNA-binding (GO:0003723). All proteins with neither of those three, nor with any indirect annotations (keywords: *DNA*, *RNA*, *nucleotide*) served as negatives. This procedure was only applied for per-protein predictions (e.g. protein binds DNA or not). For all per-residue predictions (e.g. which residues bind DNA), all residues NOT annotated to bind in a particular PDB chain (e.g. DNA) served as negatives.

Independent data sets for comparisons to existing methods

In order to compare our new method to others, we built new sets without sequence redundancy (HVAL < 0 [22]) to the proteins used for developing our method. We also applied another HVAL < 5 filter to rule out possible overlap between any protein used for testing ProNA2020 components and those proteins used to develop the prediction methods used as input through the PredictProtein [30] server; this applied in particular to predicted secondary structure and solvent accessibility. The advantage of this solution was that we could compare tools based on the same data sets for proteins not similar to those used for development. The problem was that these rigorous

Table 1. Non-redundant[a] cross-validation[b] set for per-residue predictions.[c]

	No. of binding residues	No. of non-binding residues	No. of all residues	Percentage binding
Protein-binding residues	29,438	78,608	108,046	27.2%
DNA-binding residues	6644	19,227	25,871	25.7%
RNA-binding residues	8588	21,538	30,126	28.5%

[a] For all data sets, UniqueProt reduced redundancy such that no protein pair in the set had sequence similarity of HVAL > 0 (corresponding to 20% pairwise sequence identity for alignments longer than 250 residues).
[b] Cross-validation: We separated the whole development/cross-validation set into five parts. Training used three of five (training set); one of five (cross-training set) was used to optimize hyper-parameters (incl. different input feature combinations, window sizes, combinations of methods). For all decisions, optimal was defined as the highest F1 score. The last of the five was used to evaluate the performance of the final model (testing set). The sets were rotated five times such that each protein in the data set had been used for testing (and cross-training) exactly once.
[c] Per-residue prediction: prediction of which residue in a protein binds DNA|RNA|protein (or combinations thereof). All residues NOT observed to bind were considered NOT binding.

Table 2. Non-redundant[a] cross-validation[b] set for per-protein predictions.[c,d]

Data set	Number of binding proteins
Protein-binding proteins	524
Negative for protein-binding proteins	282
DNA-binding proteins	199
RNA-binding proteins	263
Negative for DNA and RNA-binding proteins	555
Overlap between protein-binding negative and DNA/RNA-binding negative[a]	108

[a] For all data sets, UniqueProt reduced redundancy such that no protein pair in the set had sequence similarity of HVAL > 0 (corresponding to 20% pairwise sequence identity for alignments longer than 250 residues).
[b] Cross-validation: We separated the whole development/cross-validation set into five parts. Training used three of five (training set); one of five (cross-training set) was used to optimize hyper-parameters (incl. different input feature combinations, window sizes, combinations of methods). For all decisions, optimal was defined as the highest F1 score. The last of the five was used to evaluate the performance of the final model (testing set). The sets were rotated five times such that each protein in the data set had been used for testing (and cross-training) exactly once.
[c] Per-protein prediction: prediction that a protein binds DNA|RNA|protein (or combinations thereof) as opposed to where it binds, i.e. the binding residues. Toward this task, we need to consider a representative data set of proteins NOT binding.
[d] When testing the performance of the whole system, the overlap between neither protein-binding nor DNA/RNA-binding served as the data set for non-binding.

constraints resulted in relatively small sets. PDB sequences from 2010 were selected to assess DNA- and RNA-binding; PDB sequences from 2016 for PPI. All data sets were processed (resolution, distance threshold, and redundancy reduction) in the same way as the development data sets (Tables 1 and 2), namely: PDB resolutions <2.5 Å; binding residues within 6 Å of molecule (statistics in Table 3). PISA server is used to define the biological interface [31].

Prediction methods

Homology-based inference

Homology-based inference refers to the following process. Assume that a particular phenotype (e.g. protein binds DNA) is known for protein X, and that protein U has a sequence similarity to X exceeding some threshold (EVAL(U,X)>T), above which the phenotype is typically conserved between evolutionarily related proteins. Then we will infer that U has the same phenotype as X (e.g. U also binds DNA). The alignments for homology-based inference were generated by PSI-BLAST using the following standard protocol implemented, e.g. in the PredictProtein Server [30]. For each protein, build the PSI-BLAST profile using an 80% non-redundant database combining UniProt and PDB (two iterations, inclusion threshold E-value $\leq 10^{-3}$). These profiles were then aligned against all proteins with experimental annotation of binding (proteins have experimental annotations of protein binding (GO:0005515), DNA-binding (GO:0003677), and RNA-binding (GO:0003723))(inclusion E-value $\leq 10^{-3}$). PSI-BLAST hits to the protein in the test set were excluded to avoid over-estimate [32].

Cross-training and testing

All hyper-parameter optimizations were done on the cross-training sets. This included the choice of alternative machine learning methods (e.g. between profile-kernel SVM and ProtVec Local). All results for the final estimates of performance were compiled either on the test set or on the independent test set. No parameter was optimized on these. For instance, the decision to combine SVM and ProtVec Local on each node of the per-protein level prediction rather than to use the single best at each node (Fig. 1) appeared optimal for the cross-training set, not for the independent test set (we did provide the estimate for the combination, i.e. not the one performing best in comparison to other methods). Overall, different parts from the identical data set served as training, cross-training, and testing sets; all were rotated through so that every protein in the redundancy-reduced set was used for testing exactly once and for cross-training exactly once, implying that the cross-training and testing sets were identical (Fig. S1): five-fold cross-validation was accomplished by using three splits of the data for training, one for cross-training (optimize hyper-parameters, including number of hidden units in NN, early stop) and one for testing. Overall, we optimized the parameters (such as the number of node, learning rate for NN; *k*-mer, σ for profile-kernel) and features for residue-level prediction in the cross-training set and tested the final performance on the testing set. This implied that we actually trained five different machine learning models for each task, and that each protein from the main development data set was used for testing/cross-training exactly once. We picked the optimal hyper-parameters with best average performance in cross-training splits. This along with avoiding feature-selection decreased the likelihood of over-fitting. In fact, the choice of input units essentially followed what had been best for earlier methods developed in our lab.

Random prediction

All performance values were compared to random predictions. A random prediction was created by choosing a random number between 0 and 1, if >0.5, the residue was predicted as binding. The random per-protein predictions used the same tree-like hierarchical prediction system as the machine learning method (Fig. 1).

Prediction methods

When training the various machine learning models, protein binding and nucleotide binding were considered as separate tasks solved by two different systems of decision trees (Fig. 1, Table S1; each node represented one binary machine learning model typically trained on different data sets with different inputs and outputs).

(1) **Per-protein: profile-kernel SVM.** Support Vector Machines (SVMs) were implemented through WEKA [33]. The profile-kernel function

Table 3. Non-redundant[a] independent[b] test data set.

	For per-protein predictions		For per-residue predictions	
	No. of binding proteins	No. of non-binding proteins	No. of binding residues	No. of non-binding residues
Protein-binding	209	52	5174	10,447
DNA-binding	109	152	3645	8345
RNA-binding	57	204	1444	4711

[a] For all data sets, UniqueProt reduced redundancy such that no protein pair in the set had sequence similarity of HVAL > 0 (corresponding to 20% pairwise sequence identity for alignments longer than 250 residues). In addition, none of those proteins had HVAL > 0 to any protein used for development of any of the methods compared.

[b] Independent test set refers to the fact that those experimental measurements have become available AFTER the data sets used for the development of ProNA2020. Again not only were those proteins new, they also differed significantly in terms of sequence similarity (HVAL < 0) to any that had been available before.

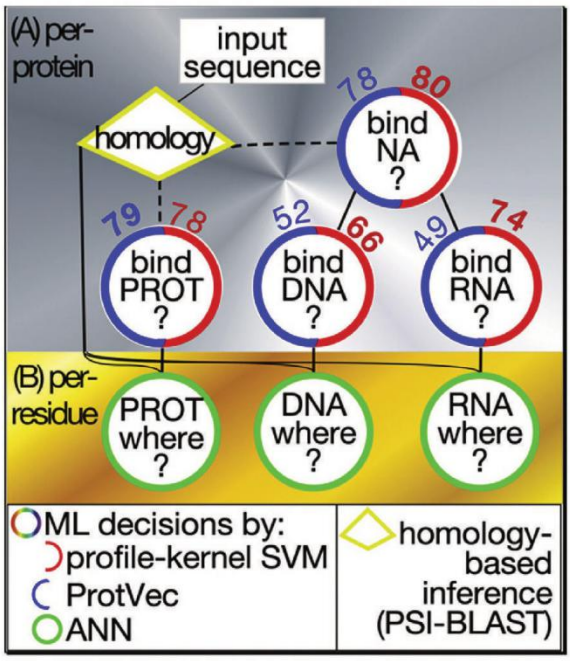


Fig. 1. Hierarchical prediction system. The branches represent the paths for the protein sorting, the nodes mark particular prediction methods (circles: machine learning (ML) models, rhombus: homology-based inference. Full lines mark part of the hierarchy the system will follow (higher in the image: earlier in the processing hierarchy). In contrast, dashed lines (from the homology-based inference) are those that might lead to bypass full lines. (A) Per-protein: The top silver gray panel is the major novelty of this contribution, namely the integration of modules specialized for per-protein level prediction. These are four ML modules predicting whether a query binds any: nucleotide (NA), proteins (PROT), DNA, or RNA. The values above the red/blue ML nodes give the F1 score of profile-kernel SVMs (red) and ProtVec (blue) based on the cross-training set (best method in bold numbers). (B) Per-residue: The lower gold panel marks per-residue predictions that have been integrated into servers before. The green circles mark three separate prediction methods predicting which residues bind PROT, DNA, and RNA. Proteins are filtered through the per-protein prediction on top and passed only to the module found appropriate by the previous step. Upon request, the sorting can be bypassed if users know the binding mode (PROT|DNA|RNA) of the query protein.

mapped the PSSM profile of each protein family to a vector indexed by all possible subsequences of length k from the alphabet of amino acids. Another parameter σ in the profile-kernel SVM was the threshold to decide when a particular k-mer was considered to be conserved in the multiple sequence alignment (family) or not. So each element in the final vector represented one particular k-mer and its score gave the number of occurrences of this k-mer that was below a certain user-defined threshold σ. The dot product between two k-mer vectors reflected the similarity of two protein sequence profiles. The best combinations of profile kernel parameters (k, σ) and of SVMs were found through 5-fold cross-validation [32–34].

(2) **Per-protein: protein vectors (ProtVec).** Continuous vector representation, as a distributed representation for words, has been recently established in NLP as an efficient way to capture semantic/syntactic units [21,35]. The basic underlying idea is to elucidate the meaning of a word through its context, i.e. neighboring words. Words with similar vectors show multiple degrees of similarity. For instance, *vector(king)* − *vector(man)* + *vector(woman)* is closest to *vector(queen)* [21,35].

The method ProtVec [21,35] applies this concept of so-called skip-gram natural language models to protein sequences. In this way, consecutive amino acids are grouped into words and the whole protein sequence becomes a sentence described by an n-dimensional vector by considering contexts of different size (i.e. word lengths). These n-dimensional vectors were input into the downstream machine learning.

We used the Word2Vec [21,35] to re-implement our own version of *ProtVec* (referred to as *ProtVec Local*). Parameters optimized included the dimensionality of the feature vectors (size), the maximum distance between words within a sentence (window), and the minimum number of the words (min_count). We also tested different word lengths k of consecutive residues (k-mer, e.g. the enzyme lactase begins with the 3-mer MEL), and whether or not to use the feature "phrase". Using "phrase" implied to automatically detect common phrases (multiword expressions) from a stream of sentences. The best combination was found by five-fold cross-validation [21,35]. For the subsequent machine learning algorithm, we compared SVM, Random Forests (RF), and Neural Networks (NN).

(3) **Per-residue: neural networks and smoothing filter.** Following earlier publications [2,36], we applied a two-step process to predict per-residue binding residues. First level: We trained standard feed-forward neural networks with back-propagation and momentum term using the sliding-window approach as input (for a window size of w, when predicting for residue j, all residues from j − INT(w/2) to j + INT(w/2) were included). All input features were taken from PredictProtein [30] including, but not limited to, predicted secondary structure, predicted relative solvent accessibility, and biophysical properties of amino acids. The combinations of features and other hyper-parameters

(e.g. window sizes and hidden units) were optimized on the cross-training set using the F1 score (complete list of features: SOM Tables S2 snd S3). Second level: The final prediction score for a residue was calculated by the average of the positive values in the certain window as follows:

$$score = \frac{1}{\omega} \sum_{i=-(\omega-1)/2}^{(\omega-1)/2} raw\_score_i, (raw\_score_i > 0) \quad (1)$$

$$NPV(C) = TN/(TN+FP); TNR(C) = TN/(TN+FN)$$
$$F1(C) = 2*PRE(C)*REC(C)/(PRE(C)+REC(C)) \quad (2)$$

We also provided the confusion matrix containing the raw values for TP, TN, FP, and FN for the test set of each of our methods separately. Toward this end, we only provided results for the cross-validation test set due to the larger data set size. These raw numbers are particularly relevant to correct for overall estimates [39]; for that correction, estimates based on larger data sets appear most helpful. In addition, we monitored the overall two-state accuracy (Q_2) and the Matthews correlation coefficient (MCC):

$$Q2 = (TP+TN)/(TP+TN+FP+FN)$$
$$MCC(C) = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \quad (3)$$

Reliability index (prediction strength)

The reliability (or strength) of a prediction was described through a *reliability index* (RI) ranging from 0 (weak prediction) to 100 (confident prediction). For per-protein predictions, the RIs were computed directly from the machine learning output. For per-residue predictions, the RIs were computed from the second-level scores (Eq. (2)). For homology-based inferences from PSI-BLAST, RIs were compiled from the percentage pairwise sequence identity (PIDE). As in our settings PSI-BLAST did not find any relations at PIDE < 10%, prediction performance did not change for PIDE $\leq$ 10 (Fig. S4). Thus, RIs were re-normalized accordingly [32].

Performance evaluation

Many publications fall short of comprehensively assessing performance through a diversity of measures [37,38]. While we tried to avoid this pitfall, we also tried to confine additional analyses that only confirmed previous results to the Supporting Online Material (SOM) wherever possible to eschew obfuscation.

Proteins might bind more than one target. Thus, we intrinsically had to assess a multi-class problem. For several aspects of the evaluation, we simplified by calculating the per-protein performance for each class, by only considering that class. With the standard acronyms (TP: true positives, observed and predicted in class C; TN: true negatives, observed and predicted in non-C; FP: false positives: predicted in C, observed in non-C; FN: false negatives: predicted in non-C, observed in C), we applied the standard definitions:

$$\begin{aligned} PRE(C) &= PrecisionC \\ &= TP/(TP + FP); REC(C) \\ &= RecallC = TP/(TP + FN); \end{aligned}$$

The overall non-exclusive three-class accuracy on the protein level was defined as:

$$Accuracy(A) = \frac{1}{n} \sum_{i=1}^{n} \frac{|prd_i \cap obs_i|}{|prd_i \cap obs_i|} \quad (4)$$

where $prd_i|obs_i$ are the numbers of classes predicted|observed for protein *i*. For instance, if protein A binds DNA and other proteins, and the prediction is *RNA&Protein* binding, the Accuracy(A) would be 1/3; the random prediction would reach $A_{random} = 43 \pm 1\%$.

Family size comparison

The number of sequences in each protein family was obtained from https://pfam.xfam.org/. For a protein with multiple families, the largest family was assigned.

Error estimates

Error rates for the evaluation measures were estimated by bootstrapping [40] (without replacement to render more conservative estimates), i.e. by re-sampling the set of proteins/residues used for the evaluation 1000 times and calculating the standard deviation over those 1000 different results. Each of these sample sets contained 50% of the original proteins/residues (picked randomly, again: without replacement).

Method comparison

We did compare performance with other methods task by task using the following publicly available methods. For DNA binding, these were DNAbinder [41], DNABIND [11], NucBind [18], SomeNa [12], and StackDPPred [13]. For RNA binding, these were RNABindRPlus [42], RBPPred [14], SomeNa [12],SPOT-RNA [15], and TriPepSVM [16]. For protein binding, these were BSpred [43], iPPBS-

PseAAC [44], InteractionSites [36], LORIS [45], PPIS [46], and SPRINGS [47]. The following multi-class binding prediction methods were included: DisoRDPbind [20], DRNApred [17], hybridNAP [19], and NucBind [18]. One important novelty of this work is the finding that different machine-learning methods are needed to predict where a protein binds (per-residue level), and whether a protein binds (per-protein level). Toward this end, we can turn a method optimized for the per-residue level into a per-protein prediction by simply considering that the method predicted the protein not to bind if no residue was predicted as binding (modes of assessment summarized in Table 4).

Results

Tree-like hierarchy for prediction system complicates assessment

We implemented an intuitive tree-like hierarchy for the entire per-protein prediction system (Fig. 1). While the system was not optimized for performance, at each node in the hierarchy (Fig. 1), we tried different solutions for the machine learning and for the combination of machine learning and homology-based inference (Methods). Methods were assessed on their specific tasks and on how they performed embedded into the hierarchy (Table 4). For instance, assume the *DNA-binding* ML module correctly predicts protein P to bind DNA. Assume further that the first module *nucleotide-binding* made a mistake (Fig. 1: top right circle, Table 4: unknown binding mode). Then the *DNA-binding* module would never be activated, i.e. the system would classify incorrectly although the isolated module was indeed correct. Both aspects needed assessment because users might over-ride some components of the system. All decisions (hyper-parameter optimizations) were done on the cross-training set (Methods), NOT on the test set.

Per-protein: profile-kernel SVM and ProtVec best together

We created two versions of machine-learning classifications for each node in our protein level prediction tree-like hierarchy (Fig. 1, Tables S4 and S5): one used a profile kernel SVM and the other the skip-gram like *ProtVec* approach. For each node, the better solution was identified on the cross-training set (Fig. 1: values above circles valid for cross-training). Thus, the performance values were relevant only to set up the final system. For some tasks, ProtVec performed better (Fig. 1: blue values, numerically higher for protein binding); however, for most, the profile kernel SVM did (Fig. 1: red values, significantly better for DNA- and RNA-binding). The best result originated from running both methods for a protein and then choosing the one with the higher score. Overall, the profile-kernel performed better on proteins from larger families (Fig. 2, $P = 0.05$).

Homology-based inference embedded into the prediction system

Merging machine learning directly with homology-based inference might improve both [32]. We measured sequence similarity through PSI-BLAST at a threshold of $T = 10^{-15}$, i.e. the annotation was inferred for a query protein Q if its sequence similarity to a protein of known binding K was below T (PSI-BLAST expectation E-value(Q,K) $< 10^{-15}$; Fig. S2). For combination, we used homology-based inference (PSI-BLAST) where available (below threshold $T < 10^{-15}$), and machine learning prediction, otherwise. This combination outperformed the machine learning method, reaching an overall performance of 77 ± 1% (Eq. (4)). For all three classes, the combined predictions improved over machine-learning (Fig. S3, Table S6) and significantly over random (Fig. 3A, Table S7).

Per-residue predictions

All per-residue prediction methods were standard two-layer feed-forward neural networks, trained exclusively on a subset of protein from each class (e.g. to learn the prediction of DNA-binding residues, only proteins observed to bind DNA were used). There are two ways to assess the final system. Firstly, we measured performance for proteins known to e.g. bind DNA. Toward this end, each prediction task was tested separately, e.g. when

Table 4. Summary of three prediction modes.

	Performance measures	Description
Protein sorting mode	Accuracy, Q_2, PRE, REC, NPV, TNR, F1, MCC	Per-protein level prediction
Residue known binding mode	Q_2, PRE, REC, NPV, TNR, F1, MCC	Per-residue level prediction for proteins for which it is known THAT they bind protein/DNA/RNA for which the residue is predicted (no sorting needed)
Residue unknown binding mode	Q_2, PRE, REC, NPV, TNR, F1, MCC	Per-residue level prediction for proteins for which it is NOT known what they bind and for which the residue is predicted (mistakes in protein sorting are added to mistakes in per-residue prediction)

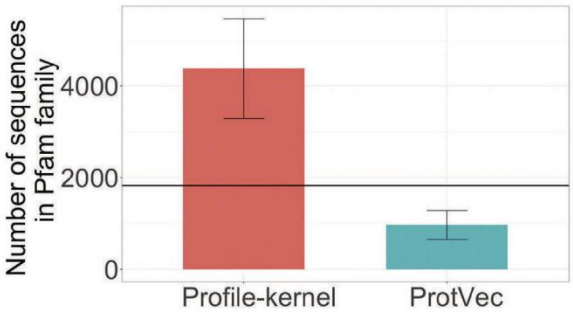


Fig. 2. Correct predictions exclusive to profile-kernel SVM vs. ProtVec. Bases for this plot are all proteins correctly predicted by only one of the two per-protein prediction algorithms, namely either by the profile-kernel SVM or by the ProtVec. The y-axis shows the average number of family members in each of the families. The horizontal black line gives the average over all families. Clearly, the profile-kernel SVMs do better for unusually large families, while the ProtVec tends to win for unusually small families.

testing DNA-binding, all DNA-binding proteins were assessed with respect to per-residue performance and all proteins experimentally known to bind DNA and those known not to bind for per-protein performance. This constitutes the standard way in which all other methods have been tested (Fig. 3A, B, D). The 2nd level filter smoothened spikes (Eq. (1) averaging over adjacent residues); it increased precision (Eq. (2)) to PRE(protein) = 46 ± 0.3% (from 35 ± 0.2% without filter), to PRE(DNA) = 57 ± 0.6% (from 48 ± 0.4%), and to PRE(RNA) = 54 ± 1% (from 46 ± 1%; Tables S8 and S5). DNA residue-binding reached the highest MCC (0.42 ± 0.006), followed by RNA residue-binding (MCC = 0.36 ± 0.006) and protein residue-binding (MCC = 0.25 ± 0.003 Fig. 3D, Tables S8 and S5). The MCC improvement was similar (Eq. (2); Fig. 3B). The improvement over random was again highest for DNA-binding (Fig. 3B, Tables S8 and S5).

Secondly, we assessed the entire sorting system, i.e. per-protein mistakes reduced per-residue performance (Fig. 3C). Overall, DNA-, RNA-binding reached similar performance; protein-binding was slightly below (Fig. 3C, Table S9). All per-residue prediction methods performed better on non-binding than on binding residues, e.g. reflected by very high levels of the overall two-state per-residue accuracy Q_2 (Eq. (3)) which was dominated by non-binding (Table 1). The test-set results were Q_2 68–70%, 80–82%, and 79–81% for protein, DNA, RNA, respectively (ranges encapsulated ± one standard error rounded to closest integer; details about error estimates are provided in Table S9). With respect to DNA/RNA confusion, 24% of the DNA binding residues were mis-predicted as RNA binding residues (Table S10).

The detailed inspection of particular examples for typical predictions (Fig. 4) suggested that ProNA2020 identified some core of a binding residue (yellow in Fig. 4). This was impressive because the method "sees" only sequence, i.e. has no notion of "binding residue", instead it only predicts "binding residues".

Predictions strength measured by reliability index (RI) correlated with performance

The confidence of each prediction was measured through a reliability index (RI) that scaled from −100 (high confidence for non-binding) to 100 (high confidence for binding). Technically, RI reflected the strength of a prediction. For homology-based

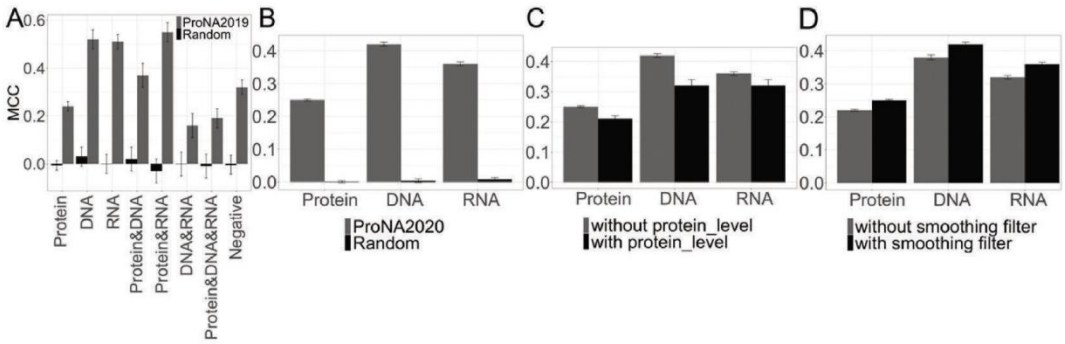


Fig. 3. Test set performance of ProNA2020. All plots show performance for the test set used to assess our new system. The first two panels give the MCC (Eq. (2)) for the per-protein (panel **A**) and per-residue predictions (panel **B**). Our new method, ProNA2020, improved over random (black vs. gray bars) by many standard deviations (±σ shown at each bar). The second two panels both give per-residue performance. Panel **C** compares values with or without errors of the protein sorting system: dark bars: with sorting (i.e. with system errors); gray without sorting (i.e. without system errors). The dark bars provide estimates for predicting binding residues without any prior knowledge; the gray bars estimate performance for users who know that their protein was a binding protein and want to find the residues involved in binding. Panel **D** compares performance between the raw ML solution (gray bars) and the smoothing filter (dark bars) that improved for all classes.

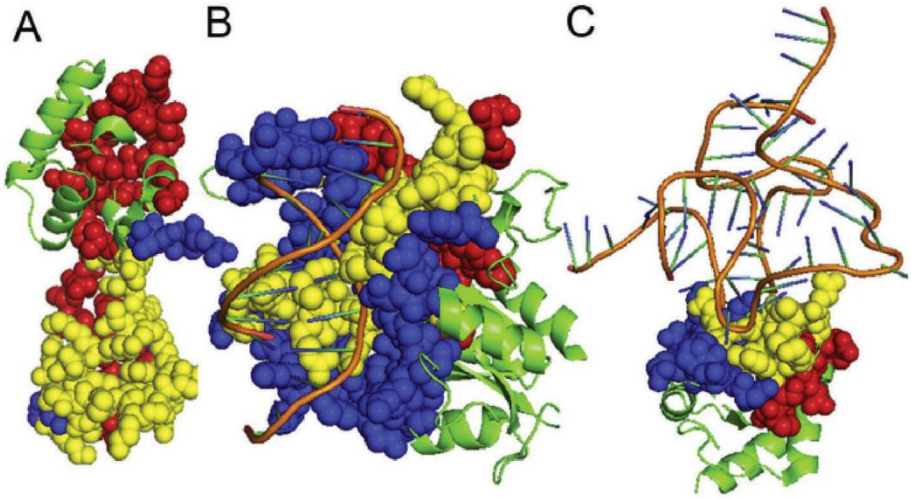


Fig. 4. Representative per-residue predictions. We picked three proteins of known 3D structure to visualize correct and incorrect predictions of binding residues for protein, DNA, and RNA. Coordinates were taken from the PDB [5]. Although each prediction was an average case for its task (complete distribution of predictions in Fig. S6), all three happened to be examples of relatively small "chains" (i.e. protein domain-like regions) that almost entirely bind. Yellow marks correctly predicted residues, blue residues observed in the binding but not predicted (under-predicted false negatives) and magenta residues predicted but not observed (over-predicted false positives). Panel **A** shows the protein binding prediction (6HA7 [57], Q2 = 71%), panel **B** gives a DNA binding prediction (5DWA [58], Q2(this protein) = 78%), and panel **C** samples an RNA binding prediction (5XTM [59], Q2(this protein) = 76%). Note that none of the 3D information was used for the prediction.

inference, the RIs were normalized values for percentage pairwise sequence identities read of the PSI-BLAST alignments (Fig. S4). For the per-protein machine learning predictions, the RIs were taken directly from the ML method output (Method). For the per-residue level, the RIs were taken from the smoothened values (Methods). The binding prediction, higher RIs corresponded to more precise (high PRE, Eq. (2)) but fewer (lower REC, Eq. (2)) predictions (Fig. 5). For instance, for the per-protein sorting, the subset of predictions stronger than 0 (RI $\geq$ 0) reached levels of >60% precision for DNA and RNA (Fig. 5A: full blue and red lines at x = 0). This level was reached for about 70% of all predictions (Fig. 5A: dashed blue and red lines at x = 0). Prediction strength correlated also with performance for the per-residue predictions of binding proteins, e.g. for RI > 0 about 50% of all protein–protein binding residues were correctly predicted (Fig. 5B: full green line), and these constituted over 40% of all the PP-binding predictions (Fig. 5B: dashed green line). For the prediction of non-binding, reversely, lower RIs implied better predictions (Fig. S5).

ProNA2020 performed best in independent comparison

To compare our new method, ProNA2020, with others, we added another independent test set without significant sequence similarity (HVAL<0) to sets used for development. For the per-protein sorting (protein sorting mode, Table 4), ProNA2020 reached the highest F1 score and MCC in protein-binding, RNA-binding, and DNA-binding prediction (Fig. 6, Table S11). Values for precision and recall never are directly comparable because some methods find different balance points, i.e. perform very well on one of the two at the price of performing poorly on another. For instance, hybridNAP reached a recall of 100% on DNA binding and RNA binding at the cost of levels of precision below 42% for DNA and below 22% for RNA. On the other extreme end, SPOT-RNA reached high precision for RNA and DisoRDPbind for protein–protein, but both achieved this at rather low recall (DisoRDPbind 41% for protein–protein, SPOT-RNA 33% for RNA). DisoRDPbind even achieved a second highest MCC in protein binding prediction by the high precision (MCC: 0.21, Fig, 6), because most other methods predicted all proteins as protein binding (NPV = 0 Table S11). Overall, for per-protein prediction, ProNA2020 numerically outperformed all state-of-the-art sequence-based binding protein prediction methods tested (in terms of F1 and MCC; in terms of Q2 for RNA binding, SPOT-RNA and TriPepSVM did better due to under-prediction, Table S11).

Methods developed to predict which residues bind e.g. DNA (per-residue level) could be employed to predict which proteins bind DNA (per-protein level). Our results highlighted the problems originating from such an approach: for all prediction tasks, all per-residue methods clearly over-predicted binding on the per-protein level. This led to very high levels of *Recall*

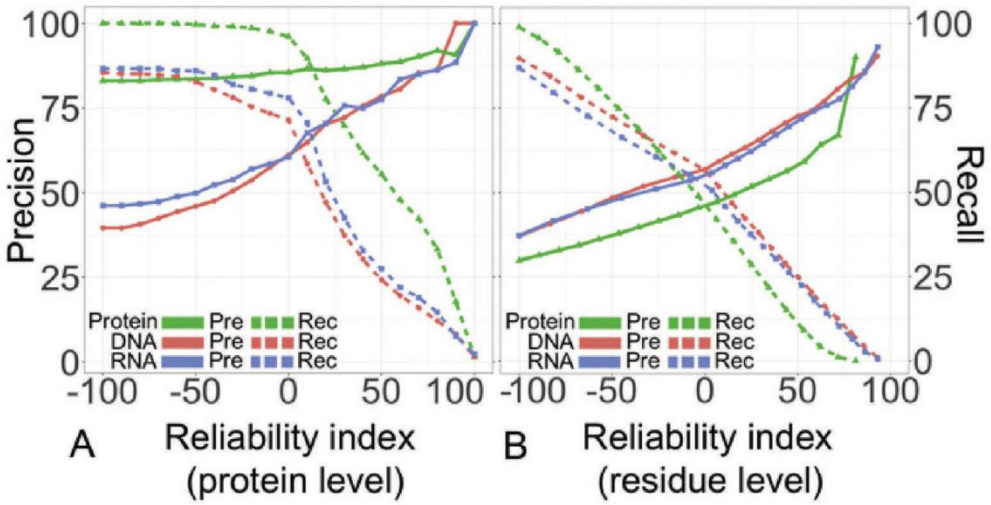


Fig. 5. Reliability index (RI) to focus on best predictions. All machine learning solutions reflect the strength of a prediction even for binary classifications (binding/not). These graphs relate prediction strength to performance. The x-axes give prediction strength as the reliability index (from −100: very non-binding to 100: very binding). The y-axes reflect the percentage precision (full lines, Eq. (2)) and recall (dashed lines, Eq. (2)) for proteins binding to DNA (red), RNA (blue), and other proteins (green). The left panel (A) shows the per-protein methods and the right one (B) the per-residue predictions. For all models, precision is proportional to prediction strengths, i.e. predictions with higher RI are, on average, better. All plots are cumulative, e.g. answering the question: if you looked at all per-residue predictions for DNA (panel B red full line) or RNA (panel B blue full line) with RI > 50 about 75% of all residues you looked at are expected to be correct predictions. Above that threshold, the methods have found slightly over 12.5% of all residues observed to bind DNA (B: dashed red) and RNA (B: dashed blue).

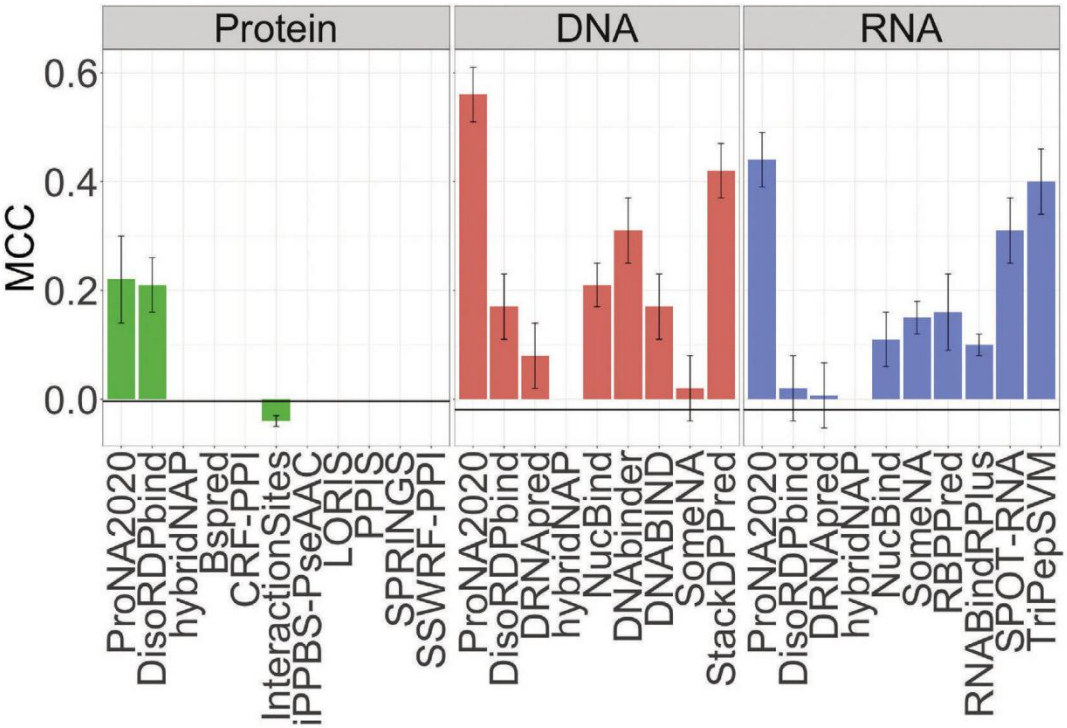


Fig. 6. Per-protein prediction of ProNA2020 in comparison for independent data set. All values are based on three new independent data sets (protein, DNA, and RNA, Table 1) without significant level of sequence similarity to those proteins used for development of all methods. The y-axis gives the MCC (Eq. (2)). Error bars define ±one standard error. All numbers were compiled on exactly the same data set. The horizontal black lines mark random predictions. Note that most data sets were imbalanced, most extreme that for protein–protein binding, as a result all but two methods (DisoRDPbind and ProNA2020) reached the same MCC (Table S11) by simply always predicting protein–protein binding, i.e. by never correctly rejecting any protein. Consequently, the MCC (Eq. (2)) was exactly 0 for all methods (Table S11) other than DisoRDPbind (MCC = 0.21 ± 0.05, Table S11) and ProNA2020 (MCC = 0.22 ± 0.08, Table S11).

at low levels of *Precision* (Table S11) and relatively low F1 scores. This problem was less severe for the identification of proteins that bind other proteins: all methods reached relatively high levels for the independent test set which contained few non-binding proteins, i.e. over-prediction of binding was rewarded, in the most extreme: always predicting binding resulted in F1 = 89%, Q2 = 80% (Precision = 80%, Recall = 100%). Consequently, the negative predictive value (NPV, Eq. (3)) for those methods might be as low as 0% (on a scale of 0–100, Table S11); the MCCs were also all 0 (Fig. 6, Table S11).

Comparing the per-residue level performance, we had to, again, distinguish the two different scenarios. First, users do not know whether or not their query Q binds (residue unknown binding mode, Table 4). Second, they do know that it binds and want to find out where it binds (residue known binding mode, Table 4). For the first scenario (unknown binding mode), no method reached higher F1 or MCC (Table 5 and Table S11, F1: unknown mode) for any task than ProNA2020. For per-residue RNA binding predictions, RNABindRPlus reached a highest MCC together with ProNA2020 (MCC = 0.40), but a slightly lower F1 than ProNA2020 ($F1_{ProNA2020} = 46$ vs. $F1_{RNABindRPlus} = 45$).

Overall, our new method, ProNA2020, appeared to be the best among all state-of-the-art per-residue prediction methods we tested with these new independent data sets. ProNA2020 clearly significantly outperformed other multi-task predictions: DRNApred, NucBind, hybridNAP, and DisoRDPbind (Table 5).

For the second scenario (known binding mode, Table 4), we e.g. only used RNA binding proteins for the per-residue RNA-binding comparison (Table 5 rightmost column, Table S13). ProNA2020 reached the highest F1 score and MCC in the DNA and protein binding per-residue prediction. The higher values were statistically significant (difference more than two standard errors, i.e. $p < 0.1$; Table 5). For RNA binding, ProNA2020 numerically reached the top MCC, followed by NucBind and RNABindRPlus; however, those two were within a single standard error of the top value, i.e. the differences were statistically not significant (Table 5). Statistically significantly lower was rank four with the other multi-task methods, namely hybridNAP with F1 = 34%, albeit at an MCC of 0.08 (Table 5). For protein binding, ProNA2020 came consistently on top highest F1 and MCC (Table S13). Performance was almost same between overall independent test

Table 5. Overall per-residue performance for independent test set[a].

Method	Binding	Unknown binding mode		Known binding mode	
		F1	MCC	F1	MCC
DisoRDPbind [20][3]	*DNA*	19 ± 3	0.09 ± 0.02	19 ± 3	0.04 ± 0.02
DRNApred [17][2]		28 ± 3	0.13 ± 0.03	30 ± 3	0.10 ± 0.03
hybridNAP [19][3]		35 ± 2	0.12 ± 0.02	40 ± 1	0.08 ± 0.02
NucBind [18][2]		35 ± 5	0.16 ± 0.07	52 ± 2	0.47 ± 0.02*
SomeNA [12][3]		44 ± 2	0.31 ± 0.03	45 ± 2	0.27±±0.04
ProNA2020[3]		**60 ± 2**	**0.49 ± 0.02**	**66 ± 1**	**0.50 ± 0.02**
DisoRDPbind [20][3]	*RNA*	15 ± 4	0.05 ± 0.03	20 ± 4	0.04 ± 0.03
DRNApred [17][2]		21 ± 5	0.08 ± 0.06	26 ± 5	0.07 ± 0.04
hybridNAP [19][3]		26 ± 3	0.11 ± 0.02	34 ± 2	0.08 ± 0.03
NucBind [18][2]		20 ± 6	0.03 ± 0.06	43 ± 5*	**0.37 ± 0.05***
RNABindRPlus [42]		45 ± 4*	0.40 ± 0.04*	**50 ± 3***	0.36 ± 0.03*
SomeNA [12][2]		23 ± 2	0.19 ± 0.04	25 ± 3	0.17 ± 0.06
ProNA2020[3]		**46 ± 3**	**0.40 ± 0.03**	**50 ± 2**	**0.37 ± 0.03**
DisoRDPbind [20][3]	*Protein*	5 ± 2	−0.03 ± 0.03	5 ± 2	−0.001 ± 0.008
hybridNAP [19][3]		37 ± 2*	0.14 ± 0.02	39 ± 2	0.11 ± 0.02
BSpred [43]		18 ± 2	−0.04 ± 0.02	20 ± 1	−0.036 ± 0.009
CRF-PPI [60]		31 ± 2	0.02 ± 0.01	38 ± 2	0.03 ± 0.01
InteractionSites [36]		14 ± 1	0.05 ± 0.02	15 ± 1	0.05 ± 0.02
iPPBS-PseAAC [44]		20 ± 1	0.04 ± 0.02	22 ± 1	0.027 ± 0.008
LORIS [45]		31 ± 2	0.001 ± 0.007	36 ± 1	0.005 ± 0.008
PPIS [46]		32 ± 2	0.01 ± 0.01	38 ± 2	0.02 ± 0.01
SPRINGS [47]		32 ± 2	0.004 ± 0.007	35 ± 2	−0.01 ± 0.008
SSWRF-PPI [61]		33 ± 2	0.02 ± 0.01	38 ± 2	0.02 ± 0.01
ProNA2020[3]		**42 ± 3**	**0.28 ± 0.03**	**47 ± 3**	**0.28 ± 0.03**

[a] **Methods**: superscript numbers give number of tasks for methods that address more than one (maximum is three: DNA, RNA, protein). **Mode-unknown**: for a query protein Q it is **not** known whether it binds DNA/RNA/Protein, instead, this binding has to also be predicted. Methods incorrectly predicting that Q binds DNA will likely mis-predict more residues than those correctly rejecting such a binding mode. Thus, values on right are mostly higher than on left. **Mode-known:** for a query protein Q it is known that it binds DNA/RNA/protein. For instance, when assessing methods for the DNA per-residue prediction, only DNA-binding proteins are presented. **Percentages** for F1 and MCC (Eq. (2)). **BOLD values and * marks**: the numerically top method in each mode is bolded; methods within two standard errors of the numerical top (p-value of difference >0.1).

set and PISA reduced independent test set (biology interface only) (Table S14).

Predictions different for prokaryotes and eukaryotes and similar for unknown data

Separately analyzing the performance for prokaryotic and eukaryotic proteins, we first observed that our training data had more residues annotated as binding RNA in prokaryotes than in eukaryotes (5351 vs. 2308, Table S16); the percentage of RNA-binding residues was also almost twice as high in prokaryotes than in eukaryotes (38% vs. 20%, Table S16); the corresponding percentages were slightly higher in prokaryotes than in eukaryotes for protein-binding (31% vs. 26%, Table S16) and this ratio was inversed for DNA-binding (24% vs. 29%, Table S16). Protein- and RNA-binding residues were predicted substantially better for prokaryotes than for eukaryotes (F1(protein) = 48 ± 0.4 vs. 45 ± 0.4; F1(RNA) = 63 ± 0.2 vs. 49 ± 0.3; Table S15). In contrast, DNA-binding residues were predicted better in eukaryotes (F1(DNA) = 54 ± 0.9 vs. 60 ± 0.8; Table S15). The differences in the amount of binding data used for training correlated but did not explain the differences in performance: protein: observed ratio binding residue (prokaryote/eukaryote) = 1.2 vs. performance (F1) of 1.05; DNA: observed ratio: 0.8, performance 0.9; RNA: observed ratio 1.9, performance 1.3.

Often experimental data sets are biased and machine learning methods inherit the training bias. For instance, all methods predicting the effects of single amino acid variants (SAVs) upon protein function perform very similar for the tiny data sets with experimental annotations, although they perform very differently for proteins without annotations [48]. The independent test sets helped to assess whether or not methods behave the same way for annotated proteins used for development and those not used. Obviously, we cannot "assess" performance for proteins without annotations. However, what we can do is to at least analyze whether the score distributions from a prediction method look similar for proteins of known and unknown function. Toward this end, we applied ProNA2020 to all human proteins and found the distribution of prediction scores to resemble that for the data sets with experimental annotations (Fig. S7).

Discussion

New system works overall better than previous tools

The major objective of this work was the combination of several prediction tasks into one comprehensive prediction system for the prediction of protein–protein, protein–DNA, and protein–RNA binding. The system included the per-protein level to automatically handle predictions for entirely sequenced organisms or metagenomes for which many proteins remained without annotations for these binding modes. The system also combined homology-based inference and machine learning to help users to the best possible prediction for each case. Many of these ideas had been realized before, e.g. the multi-task predictions (for nucleotides: SomeNA [12], DRNApred [17], and NucBind [18]; for nucleotides and proteins: DisoRDPbind [20] and hybridNAP [19]), or per-protein and per-residue level predictions (SomeNA [12]), or the combination of homology-based and machine learning (DisoRDPbind [20]). However, no system had really simultaneously addressed all aspects.

All data sets were too small for out-of-the-box Deep Learning. *Word2vec,* used so successfully by Google [33] and others, including for proteins [35,49] and in *ProtVec* [21], did provide interesting new angles (Fig. 1: blue numbers from *ProtVec*). However, profile-kernel SVMs tailored to protein prediction [12,27,34] performed better overall (Fig. 1: red mostly higher than blue numbers). Similar trends have been observed for other applications in biology [27,32,50–53]. The profile-kernel SVM mines evolutionary information as contained in multiple sequence alignments of protein families, while *ProtVec* aspires at understanding the protein sequence in a different way through NLP. It seems that the machine learning model underlying ProtVec might be too simplistic to achieve this objective. Less simplistic models reach further [54,55]. One problem for profile-kernel SVMs are un-informative (lack of diversity) and incorrect alignments. In such cases, ProtVec can perform better.

The *ProtVec*-like solution performed particularly well for the top-level protein–protein and protein–NA (nucleic acid) sorting (Fig. 1). For these, it outperformed or was *on par* with the profile-kernel SVM (Fig. 1: middle top and left top circle). Conversely, the profile-kernel SVMs clearly performed better for DNA and RNA (Fig. 1: middle circles on right and in center). One common trend was that the larger the data set, the relatively better the *ProtVec*. The finding that the best combination used whichever prediction had the highest score (reliability) suggested that methods had learned independent aspects.

One task often implicitly left to the user is the combination of homology-based inference with machine learning. Building such a combination into a system can improve and simplify predictions [32]. For ProNA2020, performance also improved through in-built combination of machine learning with homology-based inference (Fig. S3). For example, protein-binding protein Q9Y3Y4 cannot be predicted by

machine learning, while Q9Y3Y4 hits another protein-binding protein Q9T0K5 through homology-based inference.

The non-redundant independent data set was composed of proteins for which experimental data became available after the proteins used for development (cross-validation). Thus, this set was completely "novel" with respect to independently testing our method. However, several of the other methods compared had access in their development to some (older methods) or most (newer methods) of those proteins, i.e. our independent comparison was conservative in that it likely under-estimated the performance of our methods with respect to that of others. Nevertheless, in this test, no other method statistically significantly outperformed our method and no method combined as many crucially relevant components into a system as ours. Some performance measures cannot be directly compared between methods, e.g. precision and recall: each method finds a different balance. Is method M1 with Precision = 60% and Recall = 30% better than M2 with P = 40%, R = 50%? The only way to answer is through composite scores such as the F1 or MCC. When scanning such composite scores, our new method ProNA2020 reached numerically the highest value for all three per-protein predictions (Table S11, Fig. 6) and for all per-residue assessments (Table 5).

Another important feature of our prediction system that is not assessed through the independent test set is the integration of homology-based inference. By design, the independent test set could not be subjected to homology-based inference, i.e. the method comparison was confined to assessing the machine learning part of ProNA2020. Other methods use homology-based inference (e.g. SBI). In fact, for some or all of the proteins in the independent data set, those methods might have used SBI instead of *de novo* prediction.

Overall, we accomplished our goals: we developed the most comprehensive and most automated system for the prediction of binding of proteins to DNA, RNA, and other proteins. The only limitation of the system are specific predictions: it cannot predict which proteins, DNA, or RNA in particular will bind, only that they will bind and where in the protein that will happen. In absence of knowing 3D structure, the system can also not identify entire binding residues: although when mapping it onto 3D structures (Fig. 4), we observed that parts of binding residues non-consecutive in sequence and close in space had been predicted; however, without the knowledge of 3D structure, this information would not have been available. Thus, the prediction of many non-consecutive protein binding residues might indicate two separate binding pockets, or one very large one. The comprehensive system, ProNA2020, consists of parts, none of which appeared worse than any state-of-the-art prediction method, and while the system will be available to users as a whole, the separate components are also available for expert users through github.

Estimates for sustained performance challenging

When assessing machine learning, proper cross-validation is essential. This includes to have non-redundant data sets and to separate all hyper-parameter optimization and model choice (based on the cross-training set, Fig. S1) from the performance estimates for the final method, for which we used two test sets—the first from our original data set (Fig. S1) and the other independent test set, which most likely had not been used for the development of other methods and clearly not used by us (Methods). We applied the final test sets only to the system that was found best using the cross-training set. This implied that some of the results shown had to be taken from this "development phase" (Fig. 1, Fig. S3), while others were taken from the test set (Fig. 3) or the independent test set (Fig. 6, Table 5). Only these results reflected the final performance estimates for the method. Values for cross-training and testing results might differ more than the estimates of standard errors suggest; this is just an aspect of development. In contrast, if values differed between test and independent test sets, this would suggest some mistake in performance estimates. Indeed, all differences (F1) between the independent and the cross-validation test set remained within less than a single standard error (Table 5, Table S11). Thus, these differences did not challenge the technical correctness of our estimates. Consistent performance of ProNA2020 in cross-validation and the independent test sets suggested that there was rather limited bias from the development set, in particular, in comparison to other methods, some of which tended to perform below the levels published when faced with new proteins between independent test set and publication (Table 5: rightmost two column, Table S13).

Many of our performance comparisons were complicated by the small sets of proteins with experimental annotations that are neither sequence similar to any protein used by any of the methods compared, nor sequence similar to each other. This double constraint has complicated comparisons in many fields of protein prediction, in particular when high-resolution data continues to be impossible for high-throughput experiments. When each novel structure continues to cost over $100,000 [56], data sets with "only" 108 novel protein binding proteins (independent test set, Table 3) carry very high value. Some methods (alphabetically: NucBind [18] and RNABindRPlus [42]) reached a similar value on the independent data set as published. Others remained below the expectations. For one of

those, namely for DisoRDPbind [20], the difference was easily explained by that it only focused on the binding residues on the disorder region. Unfortunately, we could not analyze this separately, because for none of the proteins in our independent data set did we find experimental annotations about disorder.

Another particular problem often arising from proper cross-validation is that some alternative way of solving a problem might turn out to be best according to the cross-training set (Fig. 1, e.g. numbers in blue vs. those in red), but not best for the test or the independent test set. We encountered this for the final solution for the protein sorting system: whichever prediction method (profile-kernel SVM or ProtVec Local) had the highest score at each node of the per-protein sorting (Fig. 1) was best for the cross-training but was not best for the independent test set. Proper procedure, in cases such as this, is to trust the procedure and stick with the cross-training results, at the expense of reducing the values in the direct face-to-face comparison to other methods.

Conclusion

Each component of ProNA2020 essentially outperformed the state-of-the-art methods in per-protein sorting (Table S11, Fig. 6). With respect to most criteria, ProNA2020 also outperformed most per-residue prediction methods. When it did not outperform, it was *on par*, or at least not worse by a statistically significant margin (Table 5, Tables S12 and S13). Our method ProNA2020 is available through *github* (below), so that users could combine different components of our system with their solutions. One important novelty is the combination of per-protein sorting and per-residue prediction. We did not use existing annotations, such as Pfam domains, or Swiss-Prot annotations explicitly as input. Therefore, our system is available to be applied to high-throughput analyses, such as comparisons on the level of entire proteomes between organisms. Toward that end, ProNA2020 is available through https://github.com/Rostlab/ProNA2020.git and PredictProtein (http://www.predictprotein.org).

Acknowledgement

We thank Tim Karl for technical and Inga Weise (both TUM) for administrative assistance. We also are happy to very much thank the anonymous reviewers who pointed out many shortcomings in our initial submission and thereby crucially contributed to improving this work. Financial support was obtained from the program of China Scholarship Council. This work was supported by a grant from the Alexander von Humboldt foundation through the German Ministry for Research and Education (BMBF: Bundesministerium fuer Bildung und Forschung), as well as by the Bavarian Ministry for Education. Particular thanks to all who make databases available and all those who contribute their experimental data to such public resources.

Conflict of Interest

None declared.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jmb.2020.02.026.

Received 28 November 2019;
Received in revised form 17 February 2020;
Accepted 23 February 2020
Available online 04 March 2020

Keywords:
binding protein prediction;
binding residue prediction;
profile kernel SVM;
ProtVec;
machine learning

Abbreviations used:
3D structure, three-dimensional coordinates of protein structure; **AUC**, area under the ROC curve; **DBP**, DNA-binding protein; **FPR**, false positive rate; **PPI**, protein–-protein interaction; **ProtVec**, protein vector; **RBP**, RNA-binding protein; **RI**, reliability index; **SVM**, support vector machine; **TPR**, true positive rate; **NA**, used to describe either DNA or RNA.

References

[1] P. Liu, L. Yang, D. Shi, X. Tang, Prediction of protein-protein interactions related to protein complexes based on protein interaction networks, BioMed Res. Int. 2015 (2015), 259157.
[2] Y. Ofran, V. Mysore, B. Rost, Prediction of DNA-binding residues from sequence, Bioinformatics 23 (2007) i347–i353.
[3] C. Sacca, S. Teso, M. Diligenti, A. Passerini, Improved multi-level protein-protein interaction prediction with semantic-based regularization, BMC Bioinf. 15 (2014) 103.
[4] L. Breuza, S. Poux, A. Estreicher, M.L. Famiglietti, M. Magrane, M. Tognolli, et al., The UniProtKB guide to

the human proteome, Database : Off. J. Bio. Databases Curation 2016 (2016).
[5] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, et al., The protein Data Bank, Nucleic Acids Res. 28 (2000) 235–242.
[6] S. Bienert, A. Waterhouse, T.A. de Beer, G. Tauriello, G. Studer, L. Bordoli, et al., The SWISS-MODEL Repository-new features and functionality, Nucleic Acids Res. 45 (2017) D313–D319.
[7] J. Si, J. Cui, J. Cheng, R. Wu, Computational prediction of RNA-binding proteins and binding sites, Int. J. Mol. Sci. 16 (2015) 26303–26317.
[8] J. Si, R. Zhao, R. Wu, An overview of the prediction of protein DNA-binding sites, Int. J. Mol. Sci. 16 (2015) 5194–5215.
[9] A.C. Anderson, The process of structure-based drug design, Chem. Biol. 10 (2003) 787–797.
[10] J.L. Ludington, Protein binding site analysis for drug discovery using a computational fragment-based method, Methods Mol. Biol. 1289 (2015) 145–154.
[11] A. Szilagyi, J. Skolnick, Efficient prediction of nucleic acid binding function from low-resolution protein structures, J. Mol. Biol. 358 (2006) 922–933.
[12] P. Hönigschmid, Improvement of DNA- and RNA- Protein Binding Prediction, Technical University Munich, Munich, 2012.
[13] A. Mishra, P. Pokhrel, M.T. Hoque, StackDPPred: a stacking based prediction of DNA-binding protein from sequence, Bioinformatics 35 (2019) 433–441.
[14] X. Zhang, S. Liu, RBPPred: predicting RNA-binding proteins from sequence using SVM, Bioinformatics 33 (2017) 854–862.
[15] Y. Yang, H. Zhao, J. Wang, Y. Zhou, SPOT-Seq-RNA: predicting protein-RNA complex structure and RNA-binding function by fold recognition and binding affinity prediction, Methods Mol. Biol. 1137 (2014) 119–130.
[16] A. Bressin, R. Schulte-Sasse, D. Figini, E.C. Urdaneta, B.M. Beckmann, A. Marsico, TriPepSVM: de novo prediction of RNA-binding proteins based on short amino acid motifs, Nucleic Acids Res. 47 (2019) 4406–4417.
[17] J. Yan, L. Kurgan, DRNApred, fast sequence-based method that accurately predicts and discriminates DNA- and RNA-binding residues, Nucleic Acids Res. 45 (2017) e84.
[18] H. Su, M. Liu, S. Sun, Z. Peng, J. Yang, Improving the prediction of protein-nucleic acids binding residues via multiple sequence profiles and the consensus of complementary methods, Bioinformatics 35 (2019) 930–936.
[19] J. Zhang, Z. Ma, L. Kurgan, Comprehensive review and empirical analysis of hallmarks of DNA-, RNA- and protein-binding residues in protein chains, Briefings Bioinf. 20 (4) (2019) 1250–1268.
[20] Z. Peng, L. Kurgan, High-throughput prediction of RNA, DNA and protein binding regions mediated by intrinsic disorder, Nucleic Acids Res. 43 (2015) e121.
[21] E. Asgari, M.R. Mofrad, Continuous distributed representation of biological sequences for Deep proteomics and genomics, PloS One 10 (2015), e0141287.
[22] S. Mika, B. Rost, UniqueProt: creating representative protein sequence sets, Nucleic Acids Res. 31 (2003) 3789–3791.
[23] B. Rost, Enzyme function less conserved than anticipated, J. Mol. Biol. 318 (2002) 595–608.
[24] T. Norambuena, F. Melo, The protein-DNA interface database, BMC Bioinf. 11 (2010) 262.
[25] B.A. Lewis, R.R. Walia, M. Terribilini, J. Ferguson, C. Zheng, V. Honavar, et al., PRIDB: a Protein-RNA interface database, Nucleic Acids Res. 39 (2011) D277–D282.
[26] S. Velankar, J.M. Dana, J. Jacobsen, G. van Ginkel, P.J. Gane, J. Luo, et al., SIFTS: structure integration with function, taxonomy and sequences resource, Nucleic Acids Res. 41 (2013) D483–D489.
[27] T. Hamp, B. Rost, Evolutionary profiles improve protein-protein interaction prediction from sequence, Bioinformatics 31 (2015) 1945–1950.
[28] Y. Ofran, B. Rost, Analysing six types of protein-protein interfaces, J. Mol. Biol. 325 (2003) 377–387.
[29] C. Gene Ontology, J.A. Blake, M. Dolan, H. Drabkin, D.P. Hill, N. Li, et al., Gene Ontology annotations and resources, Nucleic Acids Res. 41 (2013) D530–D535.
[30] G. Yachdav, E. Kloppmann, L. Kajan, M. Hecht, T. Goldberg, T. Hamp, et al., PredictProtein–an open resource for online prediction of protein structural and functional features, Nucleic Acids Res. 42 (2014) W337–W343.
[31] E. Krissinel, K. Henrick, Inference of macromolecular assemblies from crystalline state, J. Mol. Biol. 372 (2007) 774–797.
[32] T. Goldberg, M. Hecht, T. Hamp, T. Karl, G. Yachdav, N. Ahmed, et al., LocTree3 prediction of localization, Nucleic Acids Res. 42 (2014) W350–W355.
[33] E. Frank, M. Hall, L. Trigg, G. Holmes, I.H. Witten, Data mining in bioinformatics using Weka, Bioinformatics 20 (2004) 2479–2481.
[34] T. Hamp, T. Goldberg, B. Rost, Accelerating the original profile kernel, PloS One 8 (2013), e68459.
[35] Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. Distributed representations of words and phrases and their compositionality. Advances in Neural Information Processing Systems2013. p. 3111-3119.
[36] Y. Ofran, B. Rost, ISIS: interaction sites identified from sequence, Bioinformatics 23 (2007) e13–e16.
[37] M. Littmann, K. Selig, L. Cohen, Y. Frank, P. Hönigschmid, E. Kataka, et al., Validity of machine learning in biology and medicine increased through collaborations across fields of expertise, Nat. Mach. Intell. 2 (2020) 18–24.
[38] M. Vihinen, How to evaluate performance of prediction methods? Measures and their interpretation in variation effect analysis, BMC Genom. 13 (Suppl 4) (2012) S2.
[39] V. Marot-Lassauzaie, M. Bernhofer, B. Rost, Correcting mistakes in predicting distributions, Bioinformatics 34 (2018) 3385–3386.
[40] B. Efron, R. Tibshirani, Statistical data analysis in the computer age, Science 353 (1991) 390–395.
[41] M. Kumar, M.M. Gromiha, G.P. Raghava, Identification of DNA-binding proteins using support vector machines and evolutionary profiles, BMC Bioinf. 8 (2007) 463.
[42] R.R. Walia, L.C. Xue, K. Wilkins, Y. El-Manzalawy, D. Dobbs, V. Honavar, RNABindRPlus: a predictor that combines machine learning and sequence homology-based methods to improve the reliability of predicted RNA-binding residues in proteins, PloS One 9 (2014), e97725.
[43] S. Mukherjee, Y. Zhang, Protein-protein complex structure predictions by multimeric threading and template recombination, Structure 19 (2011) 955–966.
[44] J. Jia, Z. Liu, X. Xiao, B. Liu, K.C. Chou, Identification of protein-protein binding sites by incorporating the physicochemical properties and stationary wavelet transforms into pseudo amino acid composition, J. Biomol. Struct. Dyn. 34 (2016) 1946–1961.

[45] K. Dhole, G. Singh, P.P. Pai, S. Mondal, Sequence-based prediction of protein-protein interaction sites with L1-logreg classifier, J. Theor. Biol. 348 (2014) 47–54.
[46] G.H. Liu, H.B. Shen, D.J. Yu, Prediction of protein-protein interaction sites with machine-learning-based data-cleaning and post-filtering procedures, J. Membr. Biol. 249 (2016) 141–153.
[47] K.D. Gurdeep Singh, Priyadarshini P. Pai, Sukanta Mondal, SPRINGS: Prediction of Protein-Protein Interaction Sites Using Artificial Neural Networks PeerJ PrePrints, 2014.
[48] J. Reeb, M. Hecht, Y. Mahlich, Y. Bromberg, B. Rost, Predicted molecular effects of sequence variants link to system level of disease, PLoS Comput. Biol. 12 (2016), e1005047, https://doi.org/10.1371/journal.pcbi.
[49] J.M. Cejuela, A. Bojchevski, C. Uhlig, R. Bekmukhametov, S. Kumar Karn, S. Mahmuti, et al., nala: text mining natural language mutation mentions, Bioinformatics 33 (2017) 1852–1858.
[50] R. Kuang, C.S. Leslie, A.S. Yang, Protein backbone angle prediction with machine learning approaches, Bioinformatics 20 (2004) 1612–1621.
[51] R. Kuang, E. Ie, K. Wang, K. Wang, M. Siddiqi, Y. Freund, et al., Profile-based string kernels for remote homology detection and motif extraction, J. Bioinf. Comput. Biol. 3 (2005) 527–550.
[52] W.S. Noble, R. Kuang, C. Leslie, J. Weston, Identifying remote protein homologs by network propagation, FEBS J. 272 (2005) 5119–5128.
[53] I. Melvin, E. Ie, R. Kuang, J. Weston, W.N. Stafford, C. Leslie, SVM-Fold: a tool for discriminative multi-class protein fold and superfamily recognition, BMC Bioinf. 8 (Suppl 4) (2007) S2.
[54] M. Heinzinger, A. Elnaggar, Y. Wang, C. Dallago, D. Nechaeev, F. Matthes, et al., Modeling the Language of Life – Deep Learning Protein Sequences, bioRxiv, 2019, https://doi.org/10.1101/614313.
[55] M. Heinzinger, A. Elnaggar, Y. Wang, C. Dallago, D. Nechaev, F. Matthes, et al., Modeling aspects of the language of life through transfer-learning protein sequences, BMC Bioinf. 20 (2019) 723.
[56] J. Liu, G.T. Montelione, B. Rost, Novel leverage of structural genomics, Nat. Biotechnol. 25 (2007) 849–851.
[57] Y. Yan, C. Rato, L. Rohland, S. Preissler, D. Ron, MANF antagonizes nucleotide exchange by the endoplasmic reticulum chaperone BiP, Nat. Commun. 10 (2019) 541.
[58] G. Tamulaitiene, V. Jovaisaite, G. Tamulaitis, I. Songailiene, E. Manakova, M. Zaremba, et al., Restriction endonuclease AgeI is a monomer which dimerizes to cleave DNA, Nucleic Acids Res. 45 (2017) 3547–3558.
[59] K. Oshima, X. Gao, S. Hayashi, T. Ueda, T. Nakashima, M. Kimura, Crystal structures of the archaeal RNase P protein Rpp38 in complex with RNA fragments containing a K-turn motif, Acta Crystallogr. F Struct. Biol. Commun. 74 (2018) 57–64.
[60] Z.S. Wei, J.Y. Yang, H.B. Shen, D.J. Yu, A cascade random forests algorithm for predicting protein-protein interaction sites, IEEE Trans. NanoBioscience 14 (2015) 746–760.
[61] K.H. Zhi-Sen Wei, Jing-Yu Yang, Hong-Bin Shen, Dong-Jun Yu, Protein-protein interaction sites prediction by ensembling SVM and sample-weighted random forests, Neurocomputing 193 (2016) 201–212.

Chapter 3

3 Effect of Protein-, DNA- and RNA-binding residues on common and rare sequence variants in human

3.1 Genetic variants in human

There are no two human holding identical genome. Human genetic variation is the genetic difference among the population which makes everyone unique. It determines almost every biological phenotype of human being, such as height, skin color and even behavior. More importantly, genetic variations are related to most of human diseases. Thus, researches about genetic variation can not only make us have a better understanding of ourselves, but also bring benefit to the medicine progress, especially personalized medicine.

3.2 High-throughput sequencing

Unlike the first reference version of human genome released in 2001 which heavily depend on Sanger Sequencing (Schlessinger et al., 2006), nowads more and more genome researches utilize the high-throughput sequencing (HTS) methods, also

referred to as next-generation sequencing (NGS). Since 2006, a lot of next-generation sequencing companies and technologies have been created, and the corresponding field of bioinformatics has exploded as a major scientific and training discipline (Levy and Myers, 2016). These brought us from the first draft of the human reference genome to the ability to routinely sequence human genomes at a cost decreasing from billions of dollars to thousands of dollars (Levy and Myers, 2016).

The first aim of whole gene sequence (WGS), which is one of the most widely application in NGS, is to create a high-quality map of genome variation. And variant calling is a key step which lays the foundation for all downstream analyses about genome interpretation and genetic discovery. So far, there are three general WGS strategies (Lappalainen et al., 2019) (Figure 3.1):

Short-read WGS, can yield paired-end 150 bp reads with low error rates (0.1%-0.5%) (Lappalainen et al., 2019). Short-read approaches fall into two major categories: sequencing by ligation (SBL) and sequencing by synthesis (SBS) (Goodwin et al., 2016). The most evident difference between SBS and SBL is that SBS uses DNA polymerase to incorporate complementary nucleotides to the elongating strand, while SBL uses ligase to seal the junction between the elongating strand and the newly incorporated complementary oligonucleotides. Due to the fact that DNA polymerase is an essential enzyme in the cell, SBS is a more natural approach compared with SBL (Huang et al., 2012).

Long-read WGS, using single molecule technologies, can yield 10–100 kb reads with high error rates in the range of 10%–15% (Lappalainen et al., 2019). Genomes are found highly complex with many long repetitive elements, copy number alterations and structural variations that are related to evolution, adaptation and disease. These complex elements are so long that short-read sequencing is insufficient to resolve them. Long reads WGS, however, can span complex or repetitive regions with a single continuous read (Goodwin et al., 2016).

Linked-read WGS, using the technology from 10X Genomics, can provide the long range information missing from standard approaches. By adding a unique barcode to every short read generated from a longer molecule (e.g.50 kb), we can link the short reads together (Lappalainen et al., 2019).

Figure 3.1 shows the approach of genetic variation detection by WGS.

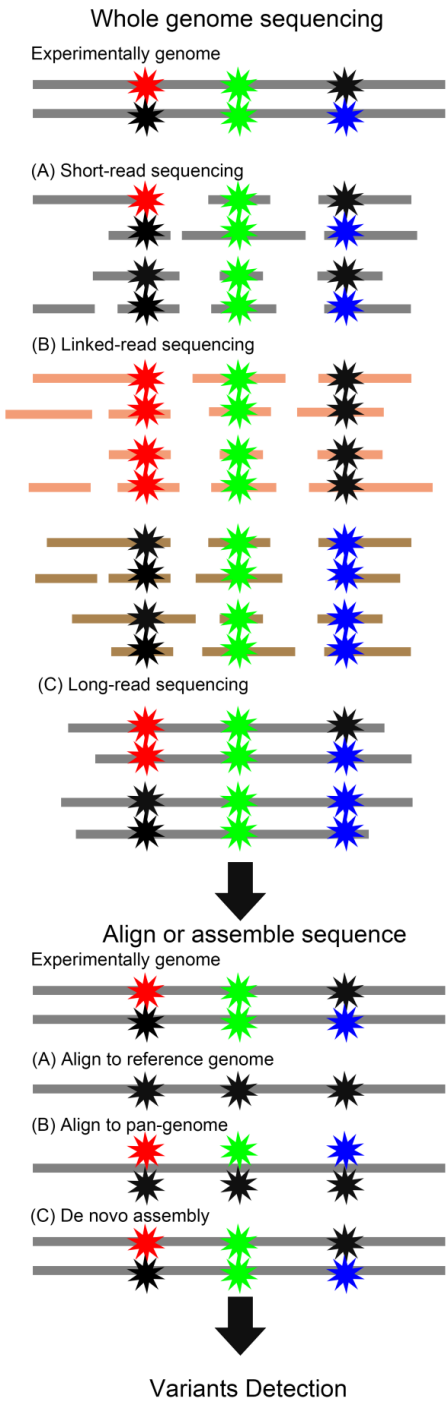


Figure 3.1: Variant detection approaches with WGS. The experimentally genome has two heterozygous variants, each of which is located on a different chromosome (blue and red stars) and one homozygous variant (green stars). Reference alleles are represented by solid lines and black stars.

3.3 Types of genetic variation

With the help of the WGS technologies, a large number of genetic variations are identified. Overall, there are four major kinds of genetic variants: SNV, Small Insertion/Deletion Variation (indel), Structure Variation (SV) and Tandem Repeat Variation. SNVs and indels comprise the majority of the genetic variants in the human genome (Table 3.1) (Lappalainen et al., 2019). On average, the genome of an individual human has 3-4 million SNVs and 0.4-0.5 million indels when compared with the reference genome. Structure variation (SV) is a diverse kind of variation that includes copy number variants (CNVs), rearrangements, and mobile element insertions (MEIs) (Table 3.1). And Tandem Repeat Variation is the variant involving high-copy repeat (Table 3.1) (Lappalainen et al., 2019).

Table 3.1: Human genetic variants (Lappalainen et al., 2019).

Variant class	Sub-class	Size	Num. / genome
Single Nucleotide Variation (SNV)		1bp	$3.5x10^6$
Small Insertion/Deletion Variation (indel)		1-49bp	$4.5x10^5$
Structural Variation (SV)	copy number variation	>50 bp	5,000
	insertion		1,500
	balanced rearrangement		40
	complex genomic rearrangement	>1 mb	0.01
	extremely large copy number variant	>1 mb	0.01
	retrogene insertion	gene coding length	10
	mobile element insertion (MEI)	0.3-7 kb	2,000
Tandem Repeat Variation	short tandem repeat (STR)	1-6 bp (repeat unit)	$1x10^5$
	variable number tandem repeat (VNTR)	7-49 bp (repeat unit)	unknown
	centromeric & heterochromatic repeats	various	unknown

In this thesis, we focus on the SNVs which are the easiest type of variants to be identified by short-read WGS. There are two sub-types of SVNs in coding regions:

synonymous or non-synonymous SNVs. Synonymous SNVs change the DNA sequence, but do not change the encoded amino acids, which is the result of the redundancy of genetic code (multiple codons code for the same amino acid). Unlike the synonymous, non-synonymous SNVs are nucleotide variations that alter the amino acids on the protein sequence, which result in biological changes and are subject to natural selection. Nonsense variants, which is a special case of non-synonymous, change a tri-nucleotide encoding for an amino acid to be a STOP-codon which leads to the premature termination of translation.

3.4 Common and rare variants

So far, the vast amount (99%) of known SAVs are found as rare variants, i.e. they are observed in fewer than 1% of the population; only about 0.5% of the SAVs are common variants, i.e. they are observed in over 5% of the population (Mahlich et al., 2017).

According to the evolutionary theory, those disease-causing variants should most likely be rare variants. Many researches based on WGS have studied properties of rare variants and their relevance for complex traits and diseases (Bomba et al., 2017). For example, Styrkarsdottir (Styrkarsdottir et al., 2013) found that gene LGR4 holds a nonsense variant associated with bone mineral density (BMD). The study has 4931 individuals with BMD and 69,034 individuals as control group. Steinthorsdottir (Steinthorsdottir et al., 2014) also discovered four rare variants in CCND2, PAM and PDX1 genes which affect the risk of Type 2 diabetes. Helgason (Helgason et al., 2013) found C3 gene holds a rare variant associated with age-related macular degeneration (AMD). Also, rare variants in TREM2 and APP genes were found associated with Alzheimer's disease (AD) (Jonsson et al., 2012; Jonsson et al., 2013).

In contrast, very few of common variants have been functionally validated to associate with diseases. However, model organism researches find common variant contributions to complex phenotypes (Gibson, 2012). And, in our previous study, we found common SAVs are predicted with more effects than rare SAVs, which means common SAVs affect molecular function more than rare SAVs (Mahlich et al., 2017).

In this thesis, we will focus on the parts of SAVs occurring at protein-protein, -DNA and –RNA binding interfaces.

3.5 Prediction of functional effects of sequence variants

The early methods for predicting effects of sequence variants utilize position-specific profiles as well as the the evolutionary conservation, which is the probabilities specifically for each position in an alignment, such as SIFT and PANTHER-subPSEC. The hypothesis behind it is that some sites are more conserved than others and do not change in order to maintain the protein functions. Thus, changes at well-conserved positions tend to be predicted as deleterious. To predict whether a sequence variant will affect protein function, SIFT takes both the position where the changes occur and the type of amino acid change into consideration (Ng and Henikoff, 2003). Given an input protein sequence, SIFT will construct the MSA through a homology search with PSI-blast. Based on the amino acid appearing at each position in the alignment, SIFT calculates the occurrence probability of every amino acid at every position which is normalized by the frequency of the most common amino acid. If this normalized value is less than an empirically defined threshold, the variant is predicted to have an effect (Ng and Henikoff, 2003).

Instead of PSI-blast, PANTHER-subPSEC (Thomas et al., 2003), which is also an early method, uses hidden Markov models in the construction of alignments. Another difference between PANTHER-subPSEC and SIFT is how the amino acid probabilities are used to determine a quantitative variant effect score. SIFT (Ng and Henikoff, 2003) uses the ratio between probability of the substituted amino acid and that of the most common amino acid at the position in the MSA. PANTHER-subPSEC (Thomas et al., 2003) uses the absolute value of the ratio between the probabilities of the wild-type and substituted variants. PANTHER-subPSEC (Thomas et al., 2003) focuses on the magnitude of the change, which means a variant could be predicted as effect if it dramatically decreases or increases the probability compared to the wild type.

PolyPhen (Ramensky et al., 2002) is the first widely used algorithm to combine sequence conservation information with structural features. In PolyPhen (Ramensky et al., 2002), TMHMM algorithm (Krogh et al., 2001) is used to predict transmembrane

regions, and the Coils2 algorithm (Lupas et al., 1991) is applied to predict coiled coil regions and the SignalP method (Nielsen et al., 1997) is for the prediction of signal peptide regions of the protein sequences. If the input variant is in a transmembrane region, PolyPhen uses the PHAT transmembrane-specific matrix score (Ng et al., 2000) to evaluate possible functional effect of a nsSNP on the transmembrane region. After these steps, PolyPhen empirically derives rules to predict whether a variant is damaging (affecting protein function) or neutral (no prototypical effect) (Ramensky et al., 2002).

Nowadays, machine learning approach is widely applied in variant effect prediction based on the above conservation concept and structure features.

One typical example is PolyPhen2，which is a successor of PolyPhen (Adzhubei et al., 2010). PolyPhen-2 uses 11 predictive features such as secondary structure, change in electrostatic charge, change in accessible surface area propensity and PHAT transmembrane-specific matrix score which is also used in PolyPhen These features were selected by an iterative greedy algorithm. (Adzhubei et al., 2010). For the classification method, PolyPhen2 uses Naïve Bayes which is a probability classifier (i.e, for a mutant allele, it assigns a probability of being damaging or neutral) (Adzhubei et al., 2010).

PhD-SNP is a method based on SVM (Capriotti et al., 2006). PhD-SNP is a system consisting of different SVMs with RBF kernel function which classifies mutations into disease-related and neutral polymorphism. 1) The first SVM is called “SVM-Sequence” whose input vector consists of 40 values: the first 20 (the 20 residue types) explicitly define the mutation situation (wild-type or mutation); the last 20 input provide the mutation sequence environment (the number of the residue type in a window approach) (Capriotti et al., 2006); 2) The second SVM is called “SVM-Profile” whose two inputs are based on MSA: one of the input elements is the ratio between the frequencies of the mutated residues and that of wild-type; the other one is the number of aligned sequences regarding to the variant (Capriotti et al., 2006).

Comparing to SVM, neural network works are found to have a better performance in the research of SNAP (Bromberg and Rost, 2007). The features SNAP used include but are not limited to: PSSM vectors from PSI-BLAST output, bio-chemical properties of the

mutated residue, the residue type, predicted accessibility and secondary structure and flexibility (Bromberg and Rost, 2007). Since the immediate local sequence environment can determine the effect of a variant, SNAP uses a window approach to capture the sequence environment information (Bromberg and Rost, 2007).

In our thesis, we use SNAP2, the successor of SNAP, to predict the effect of sequence variants (Hecht et al., 2015). SNAP2 is also a neural network based method like SNAP but include some new features such as statistical contact potentials, predicted binding residues, predicted disordered regions, co-evolving positions and residue annotations from Pfam (Hecht et al., 2015). Figure 3.2 shows an example of the SNAP2 output. The output scores of SNAP2 range from -100:very neutral to 100:very effective (Hecht et al., 2015).

Figure 3.2: Example of SNAP2 output. The output scores range from -100 (blue:neutral) to 100 (red:effective). The x-axis shows the residues in the protein sequence and y-axis represents 20 different variants for each position (black is the wild-type residue).

3.5 Results

Overall, we found both common and rare variants are less likely to be on the binding residues which agrees with the hypothesis that most SAVs are benign. However, we found that binding SAVs are over-represented for those very effective SAVs (SNAP2-scores⩾50) in both common and rare variants.

We further analyzed the distribution of SAVs according to the strength of the effect prediction (SNAP2-score). The binding SAVs are found to be more effective than non-binding SAVs. In our previous study (Mahlich et al., 2017), we found common variants seem to be more effective than rare variants. In this study, we not only confirmed this phenomenon, but also found common binding variants are the most effective SAVs. Especially, those SAVs occurring on multiple binding residues (binding all three classes of macro-molecules: DNA, RNA and protein) are found more effective than those on single binding residue (only binding DNA or RNA or protein).

3.6 Journal article

Jiajun Qiu designed and performed the analysis and writing the manuscript; Dmitrii Nechaev prepared part of dataset and helped in manuscript revision; Burkhard Rost designed and guided the analysis and revised the manuscript. All authors have read and approved the final manuscript.

BMC Bioinformatics

RESEARCH ARTICLE Open Access

Protein–protein and protein-nucleic acid binding residues important for common and rare sequence variants in human

Jiajun Qiu[1,2,5*], Dmitrii Nechaev[1,2] and Burkhard Rost[1,3,4]

*Correspondence: jiajunqiu@hotmail.com
[5] Biobank of Ninth People's Hospital, Shanghai Ninth People's Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai 200125, China
Full list of author information is available at the end of the article

Abstract

Background: Any two unrelated people differ by about 20,000 missense mutations (also referred to as SAVs: Single Amino acid Variants or missense SNV). Many SAVs have been predicted to strongly affect molecular protein function. Common SAVs (> 5% of population) were predicted to have, on average, more effect on molecular protein function than rare SAVs (< 1% of population). We hypothesized that the prevalence of effect in common over rare SAVs might partially be caused by common SAVs more often occurring at interfaces of proteins with other proteins, DNA, or RNA, thereby creating subgroup-specific phenotypes. We analyzed SAVs from 60,706 people through the lens of two prediction methods, one (SNAP2) predicting the effects of SAVs on molecular protein function, the other (ProNA2020) predicting residues in DNA-, RNA- and protein-binding interfaces.

Results: Three results stood out. Firstly, SAVs predicted to occur at binding interfaces were predicted to more likely affect molecular function than those predicted as not binding (p value $< 2.2 \times 10^{-16}$). Secondly, for SAVs predicted to occur at binding interfaces, common SAVs were predicted more strongly with effect on protein function than rare SAVs (p value $< 2.2 \times 10^{-16}$). Restriction to SAVs with experimental annotations confirmed all results, although the resulting subsets were too small to establish statistical significance for any result. Thirdly, the fraction of SAVs predicted at binding interfaces differed significantly between tissues, e.g. urinary bladder tissue was found abundant in SAVs predicted at protein-binding interfaces, and reproductive tissues (ovary, testis, vagina, seminal vesicle and endometrium) in SAVs predicted at DNA-binding interfaces.

Conclusions: Overall, the results suggested that residues at protein-, DNA-, and RNA-binding interfaces contributed toward predicting that common SAVs more likely affect molecular function than rare SAVs.

Keywords: Genome sequence analysis, Single amino acid variants (SAVs), Macromolecular binding residues, DNA-binding, RNA-binding, Protein–protein binding, Common versus rare sequence variants, Effect of sequence diversity

Journal : BMCOne 12859	Dispatch : 7-10-2020	Pages : 17
Article No : 3759	☐ LE	☐ TYPESET
MS Code :	☑ CP	☑ DISK

Background

Focus on SAVs, binding proteins/DNA/RNA, and predictions

Single nucleotide variants (SNVs; prior to modern sequencing referred to as SNPs) constitute the most frequent form of human genetic variation [1]. Non-synonymous or missense SNVs (also referred to as missense SNVs, nsSNVs, nsSNPs, or SAAVs) are one of the best-studied groups of variants in human diseases. These are SNVs altering the amino acid sequence of the encoded protein, now often termed Single Amino acid Variant (SAV) or missense variant [2]. The vast amount of known unique SAVs are rare, i.e. observed in fewer than 1% of the population; only about 0.5% of the unique SAVs are common, i.e. observed in over 5% of the population [1]. For simplicity, we referred to the subset of the residues in a protein interface that bind to either DNA, RNA, or other proteins as to *ProNA-binding residues*.

Experimental ProNA-binding annotations exist for few human proteins (Table 1). For instance, only about 1% of all SAVs considered in this study had PDB-based annotations (Method [3]) about ProNA-binding (Table 1). Although this number has increased substantially since our original analysis [1], 1% was still too small for a representative analysis, in particular given that only 18 residue positions were observed at ProNA-binding

Table 1 Data sets with experimental annotations

Type of annotation	Database	Common SAVs (LDAF > 5%)	Rare SAVs (LDAV < 1%)
Protein–protein binding			
Interface	PDB	16	7710
Other	PDB	219	56,312
Protein-DNA binding			
Interface	PDB	0	1182
Other	PDB	22	5706
Protein-RNA binding			
Interface	PDB	2	420
Other	PDB	9	2488
SUM ProNA binding			
Interface	PDB	18	9194
Other	PDB	247	62,983
Effect	OMIM|HumVar|PMD	149	7198
SUM experimental	PDB| OMIM|HumVar|PMD	404	78,993
Variant (SAV)	ExAC	34,309	6,639,624

Map of the 6,698,149 SAVs from the ExAC representing ~60 k individuals [5] onto high resolution (≤ 2.5 Å) structures from the PDB [3] to check how many SAVs are experimentally annotated at binding interfaces (labelled as *interface* in the 2nd column: closest residue atom within < 6 Å to substrate atom), with the three substrates being other proteins, DNA and RNA. *PDB* indicated usage of additional experimental data (Methods; all residues NOT explicitly annotated in a particular protein as *binding* were considered as "other"; in contrast to the ProNA2020 prediction method, this does not imply non-binding). The row labelled *SUM ProNA binding* summed over all annotations in each protein (due to possible double-binding, e.g. to DNA and RNA, the sum can be smaller than the parts). Overall 9212 SAVs (0.14%; 18 + 9194) had at least one positive ProNA-binding annotation in the PDB, and for another 63,230 SAVs (0.94%) there was some negative ProNA-binding annotation (the macro-molecule binding was in that experiment not found to bind at that position; note the total over all positive and negative ProNA-binding summed to 72,442 SAVs). The last row "*Effect annotation*" mapped variants from three databases annotating variant effects, namely OMIM [19], HumVar [20], and PMD [21] onto ExAC SAVs. For instance, 149 *common* SAVs and 7198 *rare* occurred at a residue position with an experimental effect (sum 0.11% of all SAVs). The total over both types of experimental annotations (binding/effect) provided the upper limit for SAVs with an experimental annotation about either binding or effect or both, namely 79,397 SAVs (1.2%): 404 of these for common SAVs and 78,993 for rare SAVs (2nd to last row labelled *SUM experimental*)

Journal : BMCOne 12859 | Dispatch : 7-10-2020 | Pages : 17
Article No : 3759 | ☐ LE | ☐ TYPESET
MS Code : | ☑ CP | ☑ DISK

interfaces with common SAVs (18 of 34,309, i.e. 0.05%).Therefore, results had to be based on a prediction method, namely ProNA2020, predicting DNA- RNA- and protein–protein binding interface residues [4]. The same rationale held with respect to the prediction of effects upon molecular protein function (Table 1) [5].

Common SAVs more likely than rare SAVs to affect molecular function

SAVs can impact protein function in many ways. Molecular mechanisms altering function include direct changes of binding sites [6, 7], or indirect impacts upon protein stability [7–10]. Genes and their products, the proteins, function as components of complex networks of macromolecules through biochemical or physical interactions [11]. Binding residues are important for disease pathology, e.g. 20% of the mutations on the surface of known cancer genes affect the protein–protein interaction (PPI) interface, for both tumor suppressors and oncogenes [12]. For a small subset of SAVs in regions for which some experimental annotations about protein function exist, it has been shown that SAVs are less often observed in residues important for function than expected by chance [7]. Most residues important for function considered in that study [7] related to the binding of large molecules (DNA, RNA, and protein). This suggested a selection against observing SAVs in *ProNA-binding* residues.

Predicting the effect of SAVs on molecular protein function for the ExAC data set of 60,706 exosomes [5], it has been shown that a higher fraction of all common than of all rare SAVs affect molecular protein function [1]. One possible explanation is that proteins function differently in sub-populations; an example for this are G-coupled receptors (GPCR) [13] (in fact, all proteins with seven transmembrane helices such as GPCRs stand out in the difference of effect between common and rare SAVs [14]).

Here we hypothesized that the higher fraction of common than rare SAVs with effect on molecular protein functions might be explained by residues at the interfaces that bind DNA, RNA, or proteins (collectively referred to as *ProNA-binding residues*). The rationale is the follow-up assumption that differences in binding might lead to different phenotypes in sub-populations, i.e. all those who have the variant have specifically different binding. We tried to falsify our hypothesis using SAVs with experimental annotations but had too little data to even distinguish between common and rare SAVs (Table 1). Therefore, we included all known 6,699,150 SAVs from 60,706 people [5]. For all SAVs two prediction methods were applied: SNAP2 [15, 16] predicted the effect of each SAV on molecular protein function, and ProNA2020 [4] predicted whether or not that SAV is in a ProNA-binding interface.

For each SAV, SNAP2 predicts a score scaled between −100 (strongly predicted as neutral) and +100 (strongly predicted as effect). The higher the absolute value of the score, the more reliable the prediction, i.e. the more likely to be correct. Positive values also partially correlate with the magnitude of an effect [17, 18], i.e. stronger effects are predicted more reliably. Typically, we observed differences in the distributions of common versus rare, binding versus non-binding, and strongly predicted with effect/neutral (and all combinations of those three alternatives). However, for simplicity, we frequently shortened the results to statements such as "common binding SAVs were predicted with higher effect than rare binding SAVs", to summarize the more technically correct but more complex observation that "the fraction of all common SAVs observed at residue

Journal : BMCOne 12859 | Dispatch : 7-10-2020 | Pages : 17
Article No : 3759 | ☐ LE | ☐ TYPESET
MS Code : | ☑ CP | ☑ DISK

positions that were predicted by ProNA2020 as binding, for which the SNAP2-score exceeded a certain threshold over all common SAVs was higher than the fraction of all rare SAVs observed at residue positions that were predicted by ProNA2020 as binding, for which the SNAP2-score exceeded a certain threshold over all rare SAVs". Although such shortcuts were essential for the readability of the manuscript, we tried to remain more verbose wherever deemed possible.

Results

ProNA-binding ratios similar for residues with and without known SAVs

ProNA2020 predicted residues in the binding interface of the query protein to DNA, RNA, or other proteins for all 6,698,149 SAVs (Single Amino acid Variants; or missense SNVs) from 60,706 individuals [5] with SNAP2 predictions available for their impact upon molecular function [1]. For simplicity, we referred to all those residues as to *ProNA-binding* residues. The 6.7 M SAVs hit 5,561,332 different residues in 64,301 human proteins; 75% of the residues in the same proteins were not covered by any observed SAV. All SAVs observed in fewer than one percent of the 60.7 K people were considered as rare (< 1%); common SAVs were observed in over five percent of the population (> 5%); all SAVs in between these two extremes were ignored to avoid problems with choosing a particular threshold in the distinction of common/rare. Overall, about $22.5 \pm 0.1\%$ of the SAVs hit *ProNA2020* predicted binding interface residues ($\pm$one standard error; protein-binding: $9.6 \pm 0.1\%$, DNA-binding: $12.4 \pm 0.1\%$, RNA-binding: $8.0 \pm 0.1\%$). This low standard error resulted from bootstrapping on a data set with over one million points suggesting that any sufficiently large subset would give the same result (at 95% confidence interval: between 22.3% and 22.7%). In the same set of proteins, overall 75% of the residues were not covered by observed SAVs. For these residues without observed SAVs, the fraction predicted as ProNA-binding was similar, namely $22.6 \pm 0.1\%$.

Mapping ExAC SAVs to proteins of known experimental 3D structure from the PDB (Table 1) revealed that 72,442 common or rare SAVs could be mapped to structures with ProNA-binding. Of these, 9212 SAVs had positive evidence for binding, while for 63,230 the particular PDB structure suggested no binding to the molecule (protein, DNA, or RNA) tested. Since the absence of evidence for binding under particular conditions (optimal for binding the molecule shown bound in the structure) is not evidence for the absence of binding to any molecular under any condition, we could only consider the 9212 SAVs as explicit experimental evidence. These constituted 0.14% of all SAVs (0.05% for common, and 0.14% for rare SAVs). For 7198 (0.11%) SAVs experimental effect annotations were available from OMIM [19], HumVar [20], or PMD [21] (Table 1; common: 0.43%; rare: 0.11%).

SAVs binding residues under-represented

SAVs predicted to be at ProNA-binding interfaces differed from randomly chosen positions (technically sampled from all residues in the proteins with observed SAVs). Computation of Fisher's exact test showed that SAVs were observed less than expected at ProNA2020-predicted binding interface residues (odds ratio $= 0.98$, p value $= 2.2 \times 10^{-16}$, Additional File 1: Table S2, Supporting Online Material, SOM). This trend was

Journal : BMCOne 12859 | Dispatch : 7-10-2020 | Pages : 17
Article No : 3759 | ☐ LE | ☐ TYPESET
MS Code : | ☑ CP | ☑ DISK

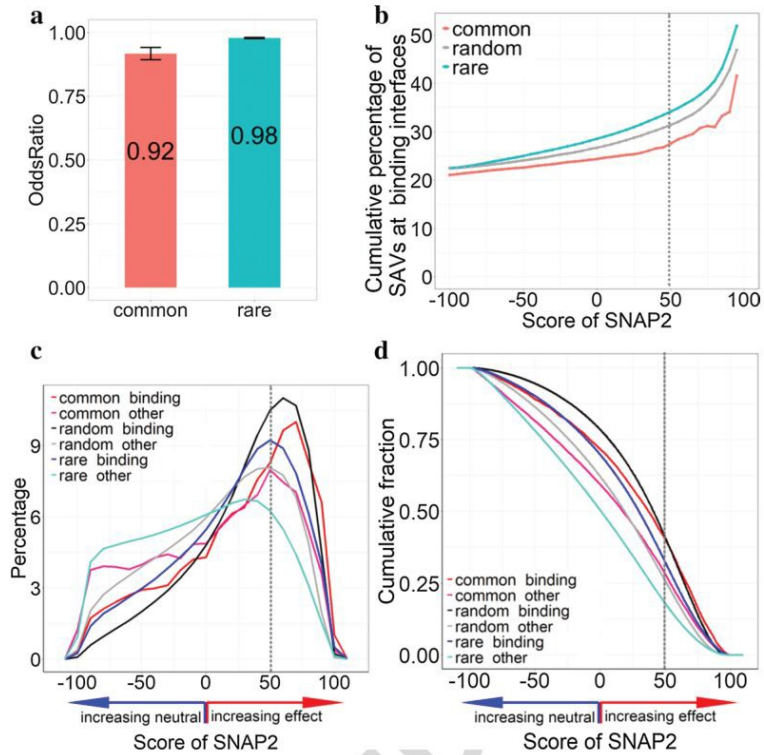


Fig. 1 Macro-molecular binding SAVs. All results were based on the ExAC data from 60 k individuals [5]; SNAP2 [15, 16] predicted effects on molecular protein function, and ProNA2020 [4] predicted residues at ProNA-binding interfaces (binding either other proteins, DNA, or RNA). (**a** demonstrates the degree to which SAVs (Single Amino acid Variants) are predicted more or less often than expected by chance (Methods) in ProNA-binding interfaces by the method ProNA2020 [4]. In particular, common SAVs (observed in > 5% of population) and rare SAVs (observed in < 1% of population) were significantly under-represented in ProNA-binding. The lines below and above the bars for the odds ratios marked the 95% confidence intervals taken from Fisher's exact test computed on the number of SAVs predicted as binding/non-binding in each class (common or rare; note the error bar for the rare SAVs is so small that it appears as a single horizontal line). **b** Zooms into the subset of all SAVs predicted as ProNA-binding. The y-axis gives the cumulative percentage of SAVs predicted above a certain SNAP2-score (x-axis) [15, 16] predicted to be in ProNA-binding interfaces. This score reflects the strength of predicting SAVs to affect molecular protein function (+100 strongest prediction of effect) or to be neutral (−100 strongest prediction of neutrality). Random (gray line) was based on the average over all possible 19-non-native mutations computed in silico (Method). Computing Kolmogorov–Smirnov *p* values between all pairs of lines revealed that the differences between common and all others were extremely significant (common vs. rare: *p* value $< 2.2 \times 10^{-16}$ and common vs. random: *p* value $< 2.7 \times 10^{-15}$). The *p* value between random and rare was not quite significant (*p* value $< 2 \times 10^{-2}$, Additional File 1: Table S1; **c**, **d** distinguish distributions between SAVs at residue positions predicted in ProNA-binding interfaces (dubbed *binding*) and non-binding (dubbed *other*) for different SNAP2-score thresholds. While **c** shows the raw distribution, **c** highlighted the cumulative distribution (as in **b**). The differences between all pairwise curves were statistically significant (Additional File 1: Table S1). For instance, for very reliable effect predictions with SNAP2-scores ≥ 50 (dashed vertical lines), about 40% of all common SAVs were predicted to affect molecular function and to be in a residue predicted or observed (ProNA2020 [4] uses whatever is available, either a homology-based inference from experimental information or machine learning prediction) to be in an interface binding a large molecule (protein, DNA, or RNA)

underscored by tests distinguishing different types of SAVs (common/rare) and different binding classes (protein-, DNA-, RNA-binding). Both common and rare SAVs were predicted less often than expected on ProNA-binding interface residues (Fig. 1a, Additional File 1: Fig. S1, p value$_{\text{common}} = 5.5 \times 10^{-11}$ and p value$_{\text{rare}} = 2.2 \times 10^{-16}$; Additional File 1:

Journal : BMCOne 12859 | Dispatch : 7-10-2020 | Pages : 17
Article No : 3759 | ☐ LE | ☐ TYPESET
MS Code : | ☑ CP | ☑ DISK

Table S3, note this defined the limit of the calculation using the software environment R [22]). The same trend held for each of the type of ProNA-binding, namely for protein, DNA, and RNA binding (Additional File 1: Table S3).

All SAVs existing in the human population might sample almost all human residues. In particular rare SAVs may ultimately sample all positions comprehensively. If so, rare SAVs should be observed in ProNA-binding interfaces exactly as expected by chance. Our results did not contradict this assumption. Although given the data set size, an odds ratio of 0.98 was distinctly below 1, this might be explained by the fact that not all SAVs can be observed in healthy individuals. ExAC sampled only people who survived to the point of becoming sequenced, i.e. SAVs so deleterious that their cells would not replicate were already selected against. While the direction of this effect (<1) is evident, its magnitude cannot be measured by our analysis, i.e. there might be some other effect to explain the difference between 0.98 and 1. However, the ProNA-binding positions predicted with the highest SNAP2-scores were clearly avoided by rare SAVs (black curve for random binding shifted to right of blue curve for rare binding in Fig. 1c and upwards in Fig. 1d). The fact that common SAVs were substantially less likely to be at ProNA-binding interfaces than expected by chance (odds ratio 0.92, Fig. 1a) was again extremely significant, as was the difference between rare and common, the latter appeared selected for avoiding ProNA-binding.

SAVs with higher effect prediction scores more likely to bind

SNAP2 [15, 16] predicts the impact of SAVs upon molecular protein function. SNAP2-scores range from $+100$ implying strong predictions of effect on molecular protein function and correlating with strong effects [17] to SNAP2-scores $=-100$ implying strong predictions of neutrality/no effect on molecular protein function. For increasing SNAP2-scores, the fractions of the residues predicted to be at ProNA-binding interface increased (Fig. 1b, Additional File 1: Table S1). The curve for rare SAVs remained above the random background, while that for common SAVs remained below random (Fig. 1b). For instance, at SNAP2-scores ≥ 50 (highly reliable effect prediction/strong effect), 34% of the rare SAVs were predicted to be at ProNA-binding interface residues. For these rare SAVs with strongly predicted effect, all types of ProNA-binding were highly over-represented with respect to random (Odds ratios clearly above 1 with Fisher's exact test p values consistently extremely significant, Additional File 1: Table S4). The situation was largely inverted for common SAVs: all odds ratios for common SAVs (ProNA, protein, DNA, and RNA) were statistically significantly below 1 (implying that binding predictions were under-represented with respect to chance) and 28% of the common SAVs were predicted at ProNA-binding interface residues for SNAP2-scores ≥ 50 (Additional File 1: Table S4). These two results indicated that, on the one hand, the SNAP2-score distributions differed substantially (and statistically significantly, Additional File 1: Table S1) between binding SAVs and non-binding SAVs for both common and rare SAVs (Fig. 1c, Additional File 1: Table S1). On the other hand, the difference in the distributions between binding and non-binding was smaller for common than for rare SAVs (Fig. 1b, rare curve above common curve). Over half of all SAVs predicted with very high SNAP2-scores (≥ 95) were predicted by *ProNA2020* as binding (Fig. 1b: rare SAVs in blue dominate the count). We also confirmed the above results for the subset of all SAVs

Journal : BMCOne 12859 | Dispatch : 7-10-2020 | Pages : 17
Article No : 3759 | ☐ LE | ☐ TYPESET
MS Code : | ☑ CP | ☑ DISK

with very strong ProNA2020 predictions for binding (|ProNA2020-scores| ≥ 50, Additional File 1: Fig. S1) This finding was consistent with results suggesting cancer SAVs to frequently hit protein-binding sites leading to loss-of function [12].

ProNA-binding SAVs stronger predicted with effect than non-binding

Next we analyzed the distribution of SAVs according to the strength of the effect prediction (SNAP2-score). Firstly, for residues predicted at ProNA-binding interfaces, the average over all possible SAVs (representing random; *19-non-native*), largely, had the highest SNAP2-scores (Fig. 1d dark line highest except for SNAP2-scores above 65); the 2nd highest was the curve for common binding SAVs (Fig. 1d). The difference between the two curves was statistically highly significant (Kolmogorov–Smirnov p value $< 2.2 \times 10^{-16}$, Additional File 1: Table S1). SAVs so deadly that they kill the carrier before birth are a subset of 19-non-native, but are removed from all ExAC SAVs. Thus, the random curves including such disruptive SAVs are expected to be shifted to the right for the distribution (Fig. 1c) and upward for the cumulative distribution (Fig. 1d). Secondly, we confirmed earlier findings [1] that common SAVs were predicted to affect molecular protein function more often than rare SAVs (Fig. 1d: common_binding higher than rare_binding and common_non-binding higher than rare_non-binding; Kolmogorov–Smirnov p value $< 2.2 \times 10^{-16}$ for both common and rare SAVs, Additional File 1: Table S1). Limiting the analysis to residues predicted as ProNA-binding with highest reliability, i.e. those predicted more strongly (|ProNA2020-scores| ≥ 50), confirmed the same tendency (Additional File 1: Fig. S1D).

Both for common and rare SAVs, SAVs at binding interfaces were predicted with stronger effect scores than non-binding SAVs (Fig. 1d: red above magenta and blue above cyan; Kolmogorov–Smirnov p value $< 2.2 \times 10^{-16}$ for common and rare SAVs, Additional File 1: Table S1). Although most common SAVs were predicted not at binding interfaces (Fig. 1d: magenta), the common SAVs predicted as ProNA-binding were predicted with higher SNAP2-scores than rare SAVs predicted as ProNA-binding (Fig. 1d: red higher than blue for SNAP2-scores > -25; Kolmogorov–Smirnov p value $< 2.2 \times 10^{-16}$, Additional File 1: Table S1). Only rare non-binding SAVs were predicted with levels of effect below that for random SAVs (Fig. 1d, only cyan below green, Additional File 1: Table S1). The combination of the findings that SAVs were predicted to be under-represented in binding interface residues (Fig. 1a) and that SAVs at binding interfaces were strongly predicted to have effect (Fig. 1d) both confirmed one aspect of our initial hypothesis: SAVs avoid ProNA-binding interface residues and when they hit those, they are likely to affect molecular protein function.

Common non-binding SAVs were predicted, on average, with higher SNAP2-scores (more likely as effect) than rare non-binding SAVs (Fig. 1d; statistical significance of difference: Kolmogorov–Smirnov p value $< 2.2 \times 10^{-16}$, Additional File 1: Table S1) and common non-binding SAVs reached effect predictions close to random SAVs (Fig. 1d: gray vs. magenta). Some of those common non-binding SAVs might be crucial for binding small molecules, i.e. be involved in signaling, or they might be related to protein stability. In fact, I-Mutant2 [10] predicted the fraction of stability-affecting SAV to be almost the same between residues predicted by *ProNA2020* as binding (84.8%) and non-binding (84.6%).

Journal : BMCOne 12859 | Dispatch : 7-10-2020 | Pages : 17
Article No : 3759 | ☐ LE | ☐ TYPESET
MS Code : | ☑ CP | ☑ DISK

Common SAVs predicted with effect but not predicted at ProNA-binding interfaces explained why rare SAVs remained below common SAVs for increasing SNAP2-scores (Fig. 1b: red below blue): rare binding SAVs tended to be predicted with higher SNAP2-scores than rare non-binding, leading to a big difference in the SNAP2-distributions for rare SAVs (Fig. 1c: blue and cyan differ; Fig. 1b: cyan highest, Additional File 1: Table S1). In contrast, common SAVs tend to have stronger effects, binding or not binding, leading to a small difference in the SNAP2-curves (Fig. 1c: red and magenta similar, Fig. 1b: red curve lowest—essentially the quotient between red and magenta in Fig. 1c, Additional File 1: Table S1). The same observation explained the under-representation of binding SAVs for very strong predictions (SNAP2-scores ≥ 50) reflected by Fisher's exact tests (Additional File 1: Table S4).

The trend that the strongest effect predictions were obtained for ProNA-binding residues, was most pronounced for protein binding (Additional File 1: Fig. S3). Of the SAVs occurring at multiple macro-molecules binding interfaces, those SAVs at protein, DNA and RNA binding interfaces, were predicted with the strongest SNAP2-scores (Additional File 1: Fig. S3, blue line, Kolmogorov–Smirnov p value $< 2.2 \times 10^{-16}$).

Validation of approach through experimental annotations

Our basic hypothesis was that SAVs at ProNA-binding interfaces more likely affect molecular protein function than those of non-binding residues. As proof of principle, we analyzed experimental annotations using proteins for which high-resolution structures of macro-molecule binding interfaces were available from the PDB [3] and superposed SAVs affecting molecular function so strongly that they cause disease (OMIM [19]). First, we mapped the SAVs from ExAC [5] upon proteins with experimentally known 3D structures [3] and experimentally known ProNA-binding sites. This procedure matched about 70 K SAVs (~1%, Table 1). For those, the fraction of ProNA-binding interface residues with predicted effect was higher than that for non-binding. Furthermore, higher fractions of common than of rare SAVs were predicted with effect, and common SAVs at binding interfaces were predicted, on average, with higher SNAP2-scores (three panels in the last row of Additional File 1: Fig. S4). The high difference between the SNAP2-score distributions of rare binding/non-binding SAVs was confirmed for the subset of SAVs with PDB annotations (first panels in the first and last row of Additional File 1: Fig. S4). This implied that the 1% of the data with high-resolution 3D information about ProNA-binding interfaces completely confirmed the trends cast by the ProNA2020 prediction method (Additional File 1: Fig. S4), but they were not statistically significant due to the small amount of data (Additional File 1: Table S5). For SAVs with experimental effect annotations (from OMIM, HumVar and PMD), rare binding SAVs were over-represented, while common binding SAVs were under-represented (Additional File 1: Table S6) confirming the finding for predictions with SNAP2-scores ≥ 50 (Fig. 1b, Additional File 1: Fig S2).

Amongst the ExAC SAVs with experimental annotations, only 392 SAVs had experimental annotations for both binding and effect (of about 6.7 m, i.e. $< 0.006\%$); none of those fell into the class common + binding. For rare SAVs, 25.4% were at protein-, 13.3% RNA-, and 29.8% DNA-binding interfaces. All these fractions exceeded those obtained for ProNA2020 and SNAP2 (at SNAP-score ≥ 50; three panels in first row of Additional

Journal : BMCOne 12859 | Dispatch : 7-10-2020 | Pages : 17
Article No : 3759 | ☐ LE | ☐ TYPESET
MS Code : | ☑ CP | ☑ DISK

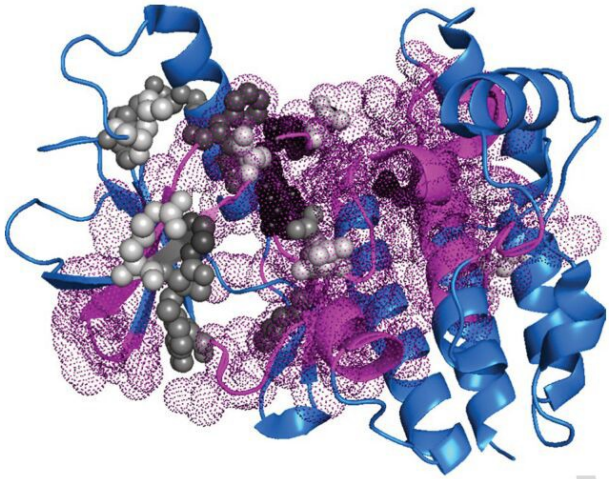


Fig. 2 SAVs in ProNA-binding interfaces predicted strongly with effect. The crystal structure of the BRAF kinase domain in complex with MEK1 (PDB identifier 4MNF [36]) illustrated a typical example for residues predicted to bind with known and predicted effect. Residues in magenta-colored dots were predicted as *ProNA-binding*; residues in gray and black spheres marked effect variants (SAVs/missense SNVs/missense mutations) annotated by experiments (from either OMIM [21], HumVar [22], or PMD [23]); the gray/black shading was proportional to the SNAP2-score (prediction of effect), from white (SNAP2-score around 0, i.e. low likelihood of effect) to black (SNAP2-score > 90, i.e. high likelihood of effect predicted). For this representative example, 86% of the SAVs predicted strongly to have effect (SNAP2-score > 90) were predicted on binding residues, i.e. were covered by magenta-colored dots

File 1: Fig S2: protein binding:17%, RNA binding: 12% and DNA binding:17.9%). The crystal structure of BRAF kinase domain in complex with MEK1 (PDB identifier 4MNF [23]) gave an example, how to imagine such an over-representation of binding residues (Fig. 2): almost 86% of the SAVs with very strong effect predictions were observed on binding interface residues.

Overall, the experimental annotations suggested the same conclusions as the prediction methods SNAP2 (for effect) and ProNA2020 (for binding). However, due to the small data size, none of those results were statistically significant (Additional File 1: Tables S5, S6), and the distinction between rare and common SAVs could not be resolved, at all. Although this cannot prove the validity of our approach, even slightly differing results could have been taken as proof-of-principle given the tiny overlaps (e.g. fraction of ExAC SAVs with experimental annotations of binding interface and effect $< 0.6*10^{-4}$, i.e. fewer than one in ten thousands).

SAVs at binding interfaces differ substantially between tissue types

Suspecting that the type of binding might differ between tissues, we investigated all proteins expressed differentially according to the Human Protein Atlas (HPA [24]). For proof-of-principle, we focused on SAVs strongly predicted to affect molecular function (SNAP2 > 50). For these, the distribution of SAVs predicted by ProNA2020 at binding interfaces, differed substantially between common and rare SAVs for all three binding classes (Fig. 3). For instance, rare SAVs predicted with strong effect occurred more often at predicted binding interfaces than expected by chance in leukocytes which play an

Journal : BMCOne 12859 | Dispatch : 7-10-2020 | Pages : 17
Article No : 3759 | ☐ LE | ☐ TYPESET
MS Code : | ☑ CP | ☑ DISK

Overall, the experimental annotations suggested the same conclusions as the prediction methods SNAP2 (for effect) and ProNA2020 (for binding). However, due to the small data size, none of those results were statistically significant (Additional File 1: Tables S5, S6), and the distinction between rare and common SAVs could not be

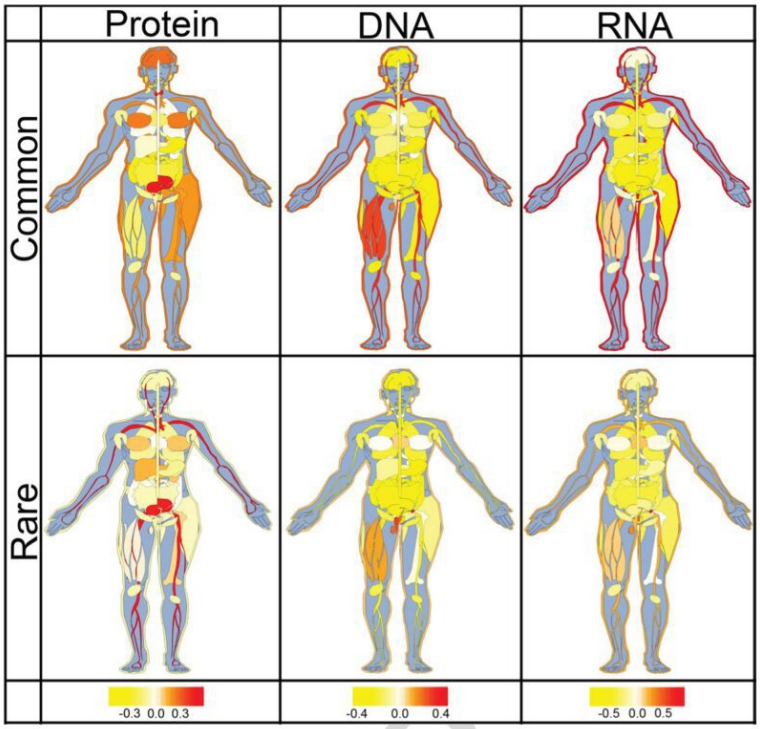


Fig. 3 Predicted ProNA-binding interface SAVs differed between human tissues. The sketches capture to which extent SAVs at residues predicted in the interfaces of protein-binding (left column), DNA-binding (middle column), or RNA-binding (right column) were over-represented in particular human tissue types (taken from HPA, the Human Protein Atlas [24]). Top row: common SAVs (> 5% of population); bottom row: rare SAVs (< 1% of population; note non-extremes between 1 and 5% were ignored). The values in each tissue were calculated as: $(PERC_{tissue}-PERC_{overall})/PERC_{overall}$ (Methods). Values around 0 (white) represented observations as expected by chance, values < 0 (yellow) indicated under-representation, and values > 0 (red) over-representation. For instance, common SAVs predicted in DNA-binding interfaces were under-represented in lung tissue, but over-represented in the skin

import role for the immune response. An intact immune response includes contributions from many subsets of leukocytes [25], e.g. from the B-cells that produce immunoglobulins (Ig) also known as antibodies. The N-termini (amino termini) of the heavy and light chains of vary between lg molecules, this variability is crucial for binding bacterial and viral pathogens. In other words, we expect to observe many binding SAVs in these regions to differ in function to adopt to many pathogens, and many of those differences would be rare as they differ between people.

Common SAVs predicted at DNA binding interfaces were enriched in skin, skeletal muscle, thyroid gland, leukocytes and testes. On the other hand, rare SAVs predicted at DNA binding interfaces were over-represented in the tissues of the reproductive system (ovaries, testes, vagina, seminal vesicle and endometrium). The latter might be explained by those tissues being more active in gene expression regulation [26, 27]. Common SAVs predicted at RNA binding interfaces were enriched in leukocytes, vagina, skin, and adrenal gland, while rare SAVs predicted at RNA binding interfaces were not over-represented in any tissue. With respect to the respiratory system, we found rare protein binding SAVs were slightly over-represented in lung.

Journal : BMCOne 12859 | Dispatch : 7-10-2020 | Pages : 17
Article No : 3759 | ☐ LE | ☐ TYPESET
MS Code : | ☑ CP | ☑ DISK

Overall, both common and rare effect SAVs predicted at macro-molecular binding interfaces were under-represented in most of internal organs such as stomach, colon and lung but over-represented in skin and leukocytes. Only SAVs at nucleotide binding (DNA or RNA) interfaces were over-represented in reproductive organs. Protein binding SAVs were over-represented in urinary bladder and brain.

Discussion

Approach limited by privacy concerns preventing access to individual genomes

Our approach had two major limitations. Due to privacy and data security the ExAC data does not allow the analysis for an individual. This has two implications: firstly, we cannot investigate compensatory mutations [28–32], i.e. instances in which two effect SAVs cancel each other out. Secondly, we cannot analyze anything such as the sum over all SAVs in a binding site. Given that we needed to base our analysis on sequence-based predictions to ascertain results of statistical significance and that SNAP2 predictions fail to identify binding sites and evolutionary couplings [33] for almost 99% of the data, these limitations did not matter for our findings. However, if we could drop privacy concerns and if we had more 3D structures, it seems almost evident by definition that random changes—as rare SAVs are expected to be—are less likely to be evolutionarily coupled than common SAVs that have been selected for in evolution. Thus dropping the limitations would most likely increase the evidence that some fraction of the difference in effect on molecular protein function between common and rare SAVs was explained by ProNA-binding.

Conclusion

A higher fraction of common SAVs (single amino acid/missense variants observed in > 5% of the population) has been predicted by the method SNAP2 [16] to affect molecular protein function than that of rare SAVs (< 1%) [1]. We hypothesized that this might be caused by common SAVs affecting interfaces binding other proteins, DNA, or RNA (dubbed *ProNA-binding*) in order to change some aspects of molecular protein function in a sub-population specific manner. Using predictions from the method ProNA2020 that combined machine learning and homology-based inference [4], we tested our hypothesis. Overall, SAVs were less likely to occur at predicted ProNA-binding interfaces than expected by chance (Fig. 1a: odds ratios < 1 with statistically extremely significant p values, Additional File 1: Tables S2–S4), common even less so than rare SAVs (Fig. 1a, b). The under-representation of common SAVs in ProNA-binding was even more pronounced for the subset of most reliably predicted binding residues (Additional File 1: Fig. S1: odds ratio 0.88). At the same time, SAVs predicted to affect molecular function by SNAP2 often coincided with ProNA-binding. Importantly, common SAVs predicted at ProNA-binding interfaces were more likely to be predicted with high SNAP2-scores than other SAVs (Fig. 1d: red curve highest for SNAP2-score > 60). In terms of binding type protein-binding SAVs were predicted with higher SNAP2-scores than nucleotide-binding SAVs, and SAVs predicted at interfaces to more than one type of binding (protein&DNA | protein&RNA | DNA&RNA | protein&DNA&RNA) were shifted most toward effect (Additional File 1: Fig. S3, blue line). All results obtained for prediction methods were essentially confirmed by explicitly using experimental annotations. However, results based on experimental data remained statistically insignificant, as fewer

Journal : BMCOne 12859 Dispatch : 7-10-2020 Pages : 17
Article No : 3759 ☐ LE ☐ TYPESET
MS Code : ☑ CP ☑ DISK

than 2‰ (0.14%) of the ExAC SAVs had reliable experimental annotations about binding interfaces (Table 1: 18 + 9194); and even fewer had experimental effect annotations (0.11%) (Table 1: 149 + 7198). Finally, we observed that ProNA-binding SAVs occurred differentially between tissue types (Fig. 3). Rare SAVs were predicted more than expected in protein-binding residues of urinary bladder tissue, and in nucleotide-binding residues of the reproductive system (ovary, testis, vagina, seminal vesicle and endometrium). Overall, the results supported our initial hypothesis that the higher fraction of common than rare SAVs with effect is partially explained by ProNA-binding (strictly speaking: the results did not refute the hypothesis). Essentially, the complex finding was that while, common SAVs were under-represented in ProNA-binding interfaces, common binding SAVs had the highest odds of affecting function. According to our hypothesis, they are the primary candidate for explaining different phenotypes in sub-populations. Rare binding SAVs also had very strong effects, consistent with the interpretation that they are not selected for in evolution (they are *rare*) because they disrupt binding. One example for the extraordinary importance of common SAVs was the differential expression of RNA-binding, in particular, in skin tissues (Fig. 3).

Methods

Data variants (SAVs)

SAVs (single amino acid variant; abbreviations found in the literature for the same include: nsSNV, nsSNP, and SAAV) were collected by the Exome Aggregation Consortium (ExAC) at the Broad Institute from 60,706 exomes [5]. We extracted all SAVs from ExAC release 0.3.1 that were labelled as 'missense variant' and 'SNV' in the 'CSQ' information field. In total, these summed to 10,474,468 SAVs; for 6,699,150 of these results from both prediction methods, SNAP2 [15, 16] (impact on molecular protein function) and ProNA2020 (ProNA-binding residues), were available. 34,309 were classified as common (linkage disequilibrium allele frequency: LDAF ≥ 0.05), 25,217 as uncommon ($0.01 \leq$ LDAF < 0.05), and 6,639,624 as rare (LDAF < 0.01).

Experimental annotations

To motivate our analysis based on predictions, we began with a collection of SAVs with experimental binding annotations based on the PDB [3]. SIFTS [34] was used to map UniProtKB sequences [35] onto PDB annotations. Binding interface residues were considered only when the closest pair of atoms between two proteins (or between protein and DNA/RNA) was within 6 Å (0.6 nm; Table 1).

A combination of OMIM, HumVar and PMD provided variant effect annotations. We extracted 22,858 human disease-associated variants/SAVs in 3537 proteins from OMIM [19] and HumVar [20], and another 3192 from PMD [21]. We mapped those variants onto ExAC SAVs. Overall 7347 variants/SAVs were experimentally annotated as effect (Table 1).

Implicitly, the PDB annotations of ProNA-binding interface residues (all residues observed in interfaces between the protein analyzed and another protein, DNA, or RNA) were used to compare trends between ProNA-binding residues experimentally known and predicted by ProNA2020 [4]. Similarly, experimental annotated SAVs from OMIM [19], HumVar [20] and PMD [21] served to compare observed SAV effects to

Journal : BMCOne 12859 | Dispatch : 7-10-2020 | Pages : 17
Article No : 3759 | ☐ LE | ☐ TYPESET
MS Code : | ☑ CP | ☑ DISK

those predicted by SNAP2 [15, 16]. Results based exclusively on experimental annotations did not provide statistically significant differences due to small counts (~1% of the SAVs had experimental binding annotations—Table 1; 0.3% had effect annotations, and 0.006% had experimental annotations for binding and effect, corresponding to 392 residue positions with observed SAVs). In particular, only ten (10!) common SAVs had annotations for effect and binding/non-binding (Table 1), rendering comparisons between common and rare SAVs impossible without predictions.

Tissue-enriched variants

Tissue-enriched variants were defined by protein expression data from *The Human Protein Atlas* (HPA https://www.proteinatlas.org) [24, 36]. As tissue-enriched variants, we considered all SAVs with an expression levels ≥ 1 (TPM or FPKM) which also were at least four-fold enriched in a particular tissue compared to the average over all other tissues. The percentage of ProNA-binding variants in each tissue were normalized as: $(PERC_{tissue}-PERC_{overall})/PERC_{overall}$. For common DNA binding variants in heart, for example, $PERC_{tissue}$ was the percentage of enriched common SAVs predicted as DNA-binding in proteins expressed in heart and $PERC_{overall}$ was the percentage of all enriched common SAVs predicted as DNA-binding (in any of the tissues considered).

Effect predictions (SNAP2

Effect scores for SAVs in all sets were computed using SNAP2 [15, 16]. SNAP2 uses a protein sequence and a list of SAVs as input to predict the effect of each substitution on molecular protein function. SNAP2 is based on a standard feed-forward neural network (often referred to as ANN) using as input biophysical amino acid properties, predicted 1D structure (incl. secondary structure, solvent accessibility from PROF [37] and ReProf [38], residue flexibility [39]), and—most importantly—evolutionary information from multiple sequence alignments generated by PSI-BLAST [40]. Cross-validated on about 100 k experimentally annotated variants, SNAP2 significantly outperformed other methods, attaining a two-state accuracy (effect/neutral) of 83% [16]. The prediction scores range from − 100 (strongly predicted as neutral) to + 100 (strongly predicted as effect). Generally, the least reliable predictions have SNAP2-scores around 0, while the most reliable ones have SNAP2-scores closer to |100|, and higher scores correlate with stronger effects [17]. This implies that the higher the SNAP2-score, the more likely the SAV with this score is (1) predicted correctly, (2) likely to have a stronger effect than another correctly predicted effect-SAV with lower score, and (3) more likely to have an effect than an effect-SAV with lower score. Largely, SNAP2 captures effects upon molecular protein function much better than effects on biological processes, and less likely over-predicts disease-affecting SAVs than other methods [16, 18, 41], although capturing OMIM-like variants with high specificity [41, 42]. Assessing the performance of SNAP2 against data from DMS studies (deep mutational scanning), suggests that the method tends to over-predict effect when assessed using a binary threshold at SNAP2-score > 0 as effect prediction [18, 43]. This had been noted earlier [44] and suggested using higher thresholds (SNAP2-score > 20) in order to distinguish effect/neutral. In our analysis, we have addressed this by mostly consider the entire spectrum of the SNAP2-score, or using thresholds even higher than this (SNAP2-score ≥ 50) for binary analyses.

Journal : BMCOne 12859 | Dispatch : 7-10-2020 | Pages : 17
Article No : 3759 | ☐ LE | ☐ TYPESET
MS Code : | ☑ CP | ☑ DISK

ProNA-binding predictions (ProNA2020

The ProNA2020 [4] method predicted for each SAV whether or not the amino acid "native" at the corresponding residue position (according to the UniProtKB/Swiss-Prot sequence [35]) is in a ProNA-binding interface, i.e. binding either to another protein, DNA, or RNA (or any combination of the three). ProNA2020 is a state-of-the-art sequence-based prediction method trained on data for binding taken from low- and high-resolution experiments on the per-protein level (protein binds or not), and from high-resolution 3D structures on the per-residue level (which residue binds). It uses a combination of different machine-learning devices and homology-based inference (if the protein is sequence similar to proteins for which experimental knowledge about binding is available). The per-residue modules learned to identify all residues in the query protein close to any atom of another protein, DNA, or RNA (closest atom within 6.5 Å = 0.6 nm of substrate; note: we referred to all of those as to ProNA-binding residues). The part of the method based on machine learning cannot identify binding sites, i.e. it cannot distinguish between two residues predicted to bind that are in the same or in two different binding sites. Overall, the machine-learning-based part of ProNA2020 reached sustained performance levels of a two-state per-residue accuracy of Q2 = 81% for DNA, Q2 = 80% for RNA, and Q2 = 69% for protein–protein interactions. In analogy to SNAP2, ProNA2020 also puts out a score ranging from −100 (strongly predicted as non-binding) to +100 (strongly predicted as binding). The default threshold for ProNA2020 [35] (ProNA2020 score > 0: binding) stroke a balance between over- or under-prediction. Consequently, the ratio of false positives/false negatives (number of residues expected to be incorrectly predicted as binding/number of residues expected to be incorrectly predicted as non-binding for ProNA2020-score > 0). For the three per-residue prediction tasks, the ratios were: 1.02 for protein-binding (minute over-prediction), 0.99 for DNA-binding (tiny under-prediction), and 0.94 for RNA-binding (slight under-prediction).

Random background predictions

We experimented with a variety of models for the random background, i.e. for establishing how much an observation differed from the expected. The problem was that all models for random sampling maintained bias from the extreme difference in the number of rare and common SAVs. Ultimately, the only viable solution was to compute all possible SAVs, i.e. all amino acid variants (all 19 non-native amino acids) at each SAV position (dubbed: 19 non-native). These *19 non-native* SAVs constituted the background. Although Deep Mutational Scanning (DMS) experiments test the effect of 19 non-native SAVs [43], not all these 19 can be accessed by changing a single nucleotide, i.e. by a SNV.

Fisher's exact test

Fisher's exact test was applied to the per-residue predictions in the following way. For instance, for DNA binding: with Ncb as the number of common SAVs predicted to bind DNA (3731), Ncn that of common SAVs not to bind DNA (30,018), Nrb the number of rare SAVs predicted to bind DNA (2,776,214), and Nrn that of rare SAVs not to bind DNA (19,661,312), we obtain:

Journal : BMCOne 12859 | Dispatch : 7-10-2020 | Pages : 17
Article No : 3759 | ☐ LE | ☐ TYPESET
MS Code : | ☑ CP | ☑ DISK

$$\text{Odd-ratio} = \frac{Ncb/Ncn}{Nrb/Nrn} = 0.88$$

The resulting p value for Fisher's exact test was calculated by the standard function *fisher.test* in the R package [22].

Error estimates

Error rates for the evaluation measures were estimated by bootstrapping [45] (without replacement to render more conservative estimates), i.e. by re-sampling the set of residues used for the evaluation 1000 times and calculating the standard deviation over those 1000 different results. Each of these sample sets contained 50% of the original residues (picked randomly, again: without replacement).

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s12859-020-03759-0.

Additional file 1. The statistical analysis results for Protein-, DNA- and RNA-binding SAVs respectively and the details for Fisher's exact tests.

Abbreviations

ExAC: Exome Aggregation Consortium; PPI: Protein-protein interaction: interactions between transiently binding different proteins; ProNA-binding residues: Describing all residues that bind proteins, DNA, or RNA; SAVs: Single amino acid variants (often also referred to as missense/non-synonymous point mutations, or missense/non-synonymous SNVs—Single Nuclear Variants); LDAF: Allele frequency as inferred from the haplotype estimation.

Acknowledgements

We thank Tim Karl for technical and Inga Weise (both TUM) for administrative assistance. Particular thanks to all who make databases available and all those who contribute their experimental data to such public resources.

Authors' contributions

J.Q. designed and performed the analysis, and writing the manuscript; D.N. prepared part of dataset and helped in manuscript revision; B.R. designed and guided the analysis and revised the manuscript. All authors have read and approved the final manuscript.

Funding

Open Access funding enabled and organized by Projekt DEAL. Financial support is from the program of China Scholarships Council (CSC201606230244). This work was supported by a grant from the Alexander von Humboldt foundation through the German Ministry for Research and Education (BMBF: Bundesministerium fuer Bildung und Forschung), as well as by the Bavarian Ministry for Education.

Availability of data and materials

We upload our dataset at: https://github.com/Rostlab/ProNA2020/tree/master/DataSet

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

None.

Author details

[1] Department of Informatics, I12-Chair of Bioinformatics and Computational Biology, Technical University of Munich (TUM), Boltzmannstrasse 3, 85748 Garching, Munich, Germany. [2] TUM Graduate School, Center of Doctoral Studies in Informatics and Its Applications (CeDoSIA), 85748 Garching, Germany. [3] Institute of Advanced Study (TUM-IAS), Lichtenbergstr. 2a, 85748 Garching, Munich, Germany. [4] Institute for Food and Plant Sciences (WZW) Weihenstephan, Alte Akademie 8, 85354 Freising, Germany. [5] Biobank of Ninth People's Hospital, Shanghai Ninth People's Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai 200125, China.

Received: 1 May 2020 Accepted: 16 September 2020

	Journal : BMCOne 12859	Dispatch : 7-10-2020	Pages : 17
	Article No : 3759	☐ LE	☐ TYPESET
	MS Code :	☑ CP	☑ DISK

References

1. Mahlich Y, Reeb J, Hecht M, Schelling M, De Beer TAP, Bromberg Y, Rost B. Common sequence variants affect molecular function more than rare variants? Sci Rep. 2017;7(1):1608.
2. Yates CM, Filippis I, Kelley LA, Sternberg MJ. SuSPect: enhanced prediction of single amino acid variant (SAV) phenotype using network features. J Mol Biol. 2014;426(14):2692–701.
3. Burley SK, Berman HM, Bhikadiya C, Bi C, Chen L, Di Costanzo L, Christie C, Dalenberg K, Duarte JM, Dutta S, et al. RCSB Protein Data Bank: biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology and energy. Nucleic Acids Res. 2019;47(D1):D464–74.
4. 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(7):2428–43.
5. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, O'Donnell-Luria AH, Ware JS, Hill AJ, Cummings BB, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature. 2016;536:285–91.
6. Peng Y, Alexov E. Investigating the linkage between disease-causing amino acid variants and their effect on protein stability and binding. Proteins. 2016;84(2):232–9.
7. de Beer TA, Laskowski RA, Parks SL, Sipos B, Goldman N, Thornton JM. Amino acid changes in disease-associated variants differ radically from variants observed in the 1000 genomes project dataset. PLoS Comput Biol. 2013;9(12):e1003382.
8. Yue P, Li Z, Moult J. Loss of protein structure stability as a major causative factor in monogenic disease. J Mol Biol. 2005;353(2):459–73.
9. Martelli PL, Fariselli P, Savojardo C, Babbi G, Aggazio F, Casadio R. Large scale analysis of protein stability in OMIM disease related human protein variants. BMC Genomics. 2016;17(Suppl 2):397.
10. Capriotti E, Fariselli P, Casadio R. I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. Nucleic Acids Res. 2005;33(Web Server issue):W306–10.
11. Zhong Q, Simonis N, Li QR, Charloteaux B, Heuze F, Klitgord N, Tam S, Yu H, Venkatesan K, Mou D, et al. Edgetic perturbation models of human inherited disorders. Mol Syst Biol. 2009;5:321.
12. Engin HB, Kreisberg JF, Carter H. Structure-based analysis reveals cancer missense mutations target protein interaction interfaces. PLoS ONE. 2016;11(4):e0152929.
13. Raimondi F, Betts MJ, Lu Q, Inoue A, Gutkind JS, Russell RB. Genetic variants affecting equivalent protein family positions reflect human diversity. Sci Rep. 2017;7(1):12771.
14. Llorian-Salvador O, Bernhofer M, Mahlich Y, Rost B. An exhaustive analysis of single amino acid variants in helical transmembrane proteins. In: bioRxiv. bioRxiv; 2019.
15. Bromberg Y, Rost B. SNAP: predict effect of non-synonymous polymorphisms on function. Nucleic Acids Res. 2007;35(11):3823–35.
16. Hecht M, Bromberg Y, Rost B. Better prediction of functional effects for sequence variants. BMC Genomics. 2015;16(Suppl 8):S1.
17. Bromberg Y, Rost B. Comprehensive in silico mutagenesis highlights functionally important residues in proteins. Bioinformatics. 2008;24(ECCB Proceedings):i207–12.
18. Reeb J, Wirth T, Rost B. Variant effect predictions capture some aspects of deep mutational scanning experiments. BMC Bioinform. 2020;21(1):107.
19. Amberger JS, Bocchini CA, Schiettecatte F, Scott AF, Hamosh A. OMIM.org: Online Mendelian Inheritance in Man (OMIM(R)), an online catalog of human genes and genetic disorders. Nucleic Acids Res. 2015;43(Database issue):D789-798.
20. Capriotti E, Calabrese R, Casadio R. Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information. Bioinformatics. 2006;22(22):2729–34.
21. Kawabata T, Ota M, Nishikawa K. The protein mutant database. Nucleic Acids Res. 1999;27(1):355–7.
22. Team RC. R: a language and environment for statistical computing. Vienna: R Foundation for Statistical Computing; 2013.
23. Haling JR, Sudhamsu J, Yen I, Sideris S, Sandoval W, Phung W, Bravo BJ, Giannetti AM, Peck A, Masselot A, et al. Structure of the BRAF-MEK complex reveals a kinase activity independent role for BRAF in MAPK signaling. Cancer Cell. 2014;26(3):402–13.
24. Uhlen M, Fagerberg L, Hallstrom BM, Lindskog C, Oksvold P, Mardinoglu A, Sivertsson A, Kampf C, Sjostedt E, Asplund A, et al. Proteomics. Tissue-based map of the human proteome. Science. 2015;347(6220):1260419.
25. Chaplin DD. Overview of the immune response. J Allergy Clin Immunol. 2010;125(2 Suppl 2):S3-23.
26. Houshdaran S, Zelenko Z, Irwin JC, Giudice LC. Human endometrial DNA methylome is cycle-dependent and is associated with gene expression regulation. Mol Endocrinol. 2014;28(7):1118–35.
27. Shima JE, McLean DJ, McCarrey JR, Griswold MD. The murine testicular transcriptome: characterizing gene expression in the testis during the progression of spermatogenesis. Biol Reprod. 2004;71(1):319–30.
28. Altschuh D, Lesk AM, Bloomer AC, Klug A. Correlation of co-ordinated amino acid substitutions with function in viruses related to tobacco mosaic virus. J Mol Biol. 1987;193:693–707.
29. Pollock DD, Taylor WR. Effectiveness of correlation analysis in identifying protein residues undergoing correlated evolution. Protein Eng. 1997;10:647–57.
30. Taylor WR, Hatrick K. Compensating changes in protein multiple sequence alignment. Protein Eng. 1994;7:341–8.
31. Goebel U, Sander C, Schneider R, Valencia A. Correlated mutations and residue contacts in proteins. Proteins Struct Funct Genet. 1994;18(4):309–17.
32. Marks DS, Hopf TA, Sander C. Protein structure prediction from sequence variation. Nat Biotechnol. 2012;30(11):1072–80.
33. Hopf TA, Ingraham JB, Poelwijk FJ, Scharfe CP, Springer M, Sander C, Marks DS. Mutation effects predicted from sequence co-variation. Nat Biotechnol. 2017;35(2):128–35.

Journal : BMCOne 12859	Dispatch : 7-10-2020	Pages : 17
Article No : 3759	☐ LE	☐ TYPESET
MS Code :	☑ CP	☑ DISK

34. Velankar S, Dana JM, Jacobsen J, Van Ginkel G, Gane PJ, Luo J, Oldfield TJ. O'donovan C, Martin M-J, Kleywegt GJ: SIFTS: structure integration with function, taxonomy and sequences resource. Nucleic Acids Res. 2012;41(D1):D483–9.
35. Boutet E, Lieberherr D, Tognolli M, Schneider M, Bansal P, Bridge AJ, Poux S, Bougueleret L, Xenarios I. UniProtKB/Swiss-Prot, the manually annotated section of the UniProt knowledgebase: how to use the entry view. Methods Mol Biol. 2016;1374:23–54.
36. Uhlen M, Oksvold P, Fagerberg L, Lundberg E, Jonasson K, Forsberg M, Zwahlen M, Kampf C, Wester K, Hober S, et al. Towards a knowledge-based Human Protein Atlas. Nat Biotechnol. 2010;28(12):1248–50.
37. Rost B. Protein secondary structure prediction continues to rise. J Struct Biol. 2001;134:204–18.
38. Kloppmann E, Hönigschmid P, Reeb J, Rost B. Protein secondary structure prediction in 2018. In: Roberts GCK, Watts A, editors. Encylopedia of Biophysics. Vienna: European Biophyscial Societies' Association; 2019.
39. Schlessinger A, Yachdav G, Rost B. PROFbval: predict flexible and rigid residues in proteins. Bioinformatics. 2006;22:891–3.
40. Altschul SF, Madden TL, Schaeffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped Blast and PSI-Blast: a new generation of protein database search programs. Nucleic Acids Res. 1997;25:3389–402.
41. Reeb J, Hecht M, Mahlich Y, Bromberg Y, Rost B. Predicted molecular effects of sequence variants link to system level of disease. PLoS Comput Biol. 2016;12(8):e1005047. https://doi.org/10.1371/journal.pcbi.1005047.
42. Schaefer C, Bromberg Y, Achten D, Rost B. Disease-related mutations predicted to impact protein function. BMC Genomics. 2012;13(Suppl 4):S11.
43. Livesey BJ, Marsh JA. Using deep mutational scanning to benchmark variant effect predictors and identify disease mutations. Mol Syst Biol. 2020;16(7):e9380.
44. Bromberg Y, Kahn PC, Rost B. Neutral and weakly nonneutral sequence variants may define individuality. Proc Natl Acad Sci USA. 2013;110(35):14255–60.
45. Efron B, Tibshirani R. Statistical data analysis in the computer age. Science. 1991;353:390–5.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Journal : **BMCOne 12859**	Dispatch : **7-10-2020**	Pages : **17**
Article No : **3759**	☐ LE	☐ TYPESET
MS Code :	☑ CP	☑ DISK

Chapter 4

4 Conclusion

The interactions between proteins and other large macro-molecules: DNA, RNA, and proteins participate in all essential biological processes. And mutations or sequence variants on those binding residues will cause strong phenotype and even serious diseases. However, experiment-based binding residue identification methods are not suitable for high-throughput binding site analysis, so it is necessary to establish the computational based binding prediction methods.

In this thesis, we establish a sequence based comprehensive protein-DNA, -RNA and -protein binding prediction system: ProNA2020. ProNA2020 is a two-level prediction system which uses only protein sequence as input. In the first level (protein level), it predicts whether the input protein is a binding protein or not. And we combine the alignment based profile kernel with neutral language based ProtVec for protein level prediction. Profile-kernel has a better performance for the proteins from large families with more sequence alignments, while ProtVec is much better at proteins from small families with less sequence alignments. In the second level (residue level), for those predicted binding proteins, ProNA2020 further decides which residues is bound on the input protein. ProNA2020 is the first comprehensive system which combines protein level and residue level prediction, and it outperforms other state-of-the-art methods in particular tasks during independent test.

Overall, this thesis provides a new comprehensive protein binding prediction system which makes high-throughput binding sites researches with high accuracy to be possible. And our analyses on human SAVs indicate those SAVs with functional effects

are enriched on macro-molecular binding residues. And the SAVs on residues which bind all three macro-molecules (DNA, RNA and protein) are found to be the most effective SAVs. Thus, our research about the binding residues can benefit future biology and medicine research (e.g. precision medicine) in both methodology and theory way.

REFERENCES

Adeli, E., Wu, G., Saghafi, B., An, L., Shi, F., and Shen, D. (2017). Kernel-based Joint Feature Selection and Max-Margin Classification for Early Diagnosis of Parkinson's Disease. Sci Rep 7, 41069.

Adzhubei, I.A., Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A., Bork, P., Kondrashov, A.S., and Sunyaev, S.R. (2010). A method and server for predicting damaging missense mutations. Nat Methods 7, 248-249.

Asgari, E., and Mofrad, M.R. (2015). Continuous Distributed Representation of Biological Sequences for Deep Proteomics and Genomics. PLoS One 10, e0141287.

Baspinar, A., Cukuroglu, E., Nussinov, R., Keskin, O., and Gursoy, A. (2014). PRISM: a web server and repository for prediction of protein-protein interactions and modeling their 3D complexes. Nucleic Acids Res 42, W285-289.

Berggard, T., Linse, S., and James, P. (2007). Methods for the detection and analysis of protein-protein interactions. Proteomics 7, 2833-2842.

Bomba, L., Walter, K., and Soranzo, N. (2017). The impact of rare and low-frequency genetic variants in common disease. Genome Biol 18, 77.

Bressin, A., Schulte-Sasse, R., Figini, D., Urdaneta, E.C., Beckmann, B.M., and Marsico, A. (2019). TriPepSVM: de novo prediction of RNA-binding proteins based on short amino acid motifs. Nucleic acids research 47, 4406-4417.

Bromberg, Y., and Rost, B. (2007). SNAP: predict effect of non-synonymous polymorphisms on function. Nucleic Acids Res 35, 3823-3835.

Burley, S.K., Berman, H.M., Kleywegt, G.J., Markley, J.L., Nakamura, H., and Velankar, S. (2017). Protein Data Bank (PDB): The Single Global Macromolecular Structure Archive. Methods Mol Biol 1607, 627-641.

Capriotti, E., Calabrese, R., and Casadio, R. (2006). Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information. Bioinformatics 22, 2729-2734.

Collins, K., and Mitchell, J.R. (2002). Telomerase in the human organism. Oncogene 21, 564-579.

del Sol, A., Balling, R., Hood, L., and Galas, D. (2010). Diseases as network perturbations. Curr Opin Biotechnol 21, 566-571.

Dhole, K., Singh, G., Pai, P.P., and Mondal, S. (2014). Sequence-based prediction of protein-protein interaction sites with L1-logreg classifier. Journal of theoretical biology 348, 47-54.

Dosztanyi, Z. (2018). Prediction of protein disorder based on IUPred. Protein Sci 27, 331-340.

Eisenberg, E., and Levanon, E.Y. (2018). A-to-I RNA editing - immune protector and transcriptome diversifier. Nat Rev Genet 19, 473-490.

Esmaielbeiki, R., Krawczyk, K., Knapp, B., Nebel, J.C., and Deane, C.M. (2016). Progress and challenges in predicting protein interfaces. Brief Bioinform 17, 117-131.

Fellouse, F.A., Barthelemy, P.A., Kelley, R.F., and Sidhu, S.S. (2006). Tyrosine plays a dominant functional role in the paratope of a synthetic antibody derived from a four amino acid code. J Mol Biol 357, 100-114.

Franklin, D. (2019). P152R Mutation Within MeCP2 Can Cause Loss of DNA-Binding Selectivity. Interdiscip Sci 11, 10-20.

Gibson, G. (2012). Rare and common variants: twenty arguments. Nat Rev Genet 13, 135-145.
Goodwin, S., McPherson, J.D., and McCombie, W.R. (2016). Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet 17, 333-351.
Grechkin, M., Logsdon, B.A., Gentles, A.J., and Lee, S.I. (2016). Identifying Network Perturbation in Cancer. PLoS Comput Biol 12, e1004888.
Gurdeep Singh, K.D., Priyadarshini P. Pai and Sukanta Mondal (2014). SPRINGS: Prediction of Protein-Protein Interaction Sites Using Artificial Neural Networks PeerJ PrePrints.
Hamp, T., Goldberg, T., and Rost, B. (2013). Accelerating the Original Profile Kernel. PLoS One 8, e68459.
Hecht, M., Bromberg, Y., and Rost, B. (2015). Better prediction of functional effects for sequence variants. BMC Genomics 16 Suppl 8, S1.
Helgason, H., Sulem, P., Duvvari, M.R., Luo, H., Thorleifsson, G., Stefansson, H., Jonsdottir, I., Masson, G., Gudbjartsson, D.F., Walters, G.B., et al. (2013). A rare nonsynonymous sequence variant in C3 is associated with high risk of age-related macular degeneration. Nat Genet 45, 1371-1374.
Hönigschmid, P. (2012). Improvement of DNA- and RNA- Protein Binding Prediction. In Informatics (Munich: Technical University Munich).
Huang, Y.F., Chen, S.C., Chiang, Y.S., Chen, T.H., and Chiu, K.P. (2012). Palindromic sequence impedes sequencing-by-ligation mechanism. BMC Syst Biol 6 Suppl 2, S10.
Jeong, J.C., Lin, X., and Chen, X.W. (2011). On position-specific scoring matrix for protein function prediction. IEEE/ACM Trans Comput Biol Bioinform 8, 308-315.
Jia, J., Liu, Z., Xiao, X., Liu, B., and Chou, K.C. (2016). Identification of protein-protein binding sites by incorporating the physicochemical properties and stationary wavelet transforms into pseudo amino acid composition. Journal of biomolecular structure & dynamics 34, 1946-1961.
Jonsson, T., Atwal, J.K., Steinberg, S., Snaedal, J., Jonsson, P.V., Bjornsson, S., Stefansson, H., Sulem, P., Gudbjartsson, D., Maloney, J., et al. (2012). A mutation in APP protects against Alzheimer's disease and age-related cognitive decline. Nature 488, 96-99.
Jonsson, T., Stefansson, H., Steinberg, S., Jonsdottir, I., Jonsson, P.V., Snaedal, J., Bjornsson, S., Huttenlocher, J., Levey, A.I., Lah, J.J., et al. (2013). Variant of TREM2 associated with the risk of Alzheimer's disease. N Engl J Med 368, 107-116.
Jubb, H.C., Pandurangan, A.P., Turner, M.A., Ochoa-Montano, B., Blundell, T.L., and Ascher, D.B. (2017). Mutations at protein-protein interfaces: Small changes over big surfaces have large impacts on human health. Prog Biophys Mol Biol 128, 3-13.
Keskin, O., Tuncbag, N., and Gursoy, A. (2016). Predicting Protein-Protein Interactions from the Molecular to the Proteome Level. Chem Rev 116, 4884-4909.
Krogh, A., Larsson, B., von Heijne, G., and Sonnhammer, E.L. (2001). Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol 305, 567-580.
Kuang, R., Ie, E., Wang, K., Wang, K., Siddiqi, M., Freund, Y., and Leslie, C. (2005). Profile-based string kernels for remote homology detection and motif extraction. J Bioinform Comput Biol 3, 527-550.
Kumar, M., Gromiha, M.M., and Raghava, G.P. (2007). Identification of DNA-binding proteins using support vector machines and evolutionary profiles. BMC Bioinformatics 8, 463.

Lango Allen, H., Estrada, K., Lettre, G., Berndt, S.I., Weedon, M.N., Rivadeneira, F., Willer, C.J., Jackson, A.U., Vedantam, S., Raychaudhuri, S., et al. (2010). Hundreds of variants clustered in genomic loci and biological pathways affect human height. Nature 467, 832-838.
Lappalainen, T., Scott, A.J., Brandt, M., and Hall, I.M. (2019). Genomic Analysis in the Age of Human Genome Sequencing. Cell 177, 70-84.
Levy, S.E., and Myers, R.M. (2016). Advancements in Next-Generation Sequencing. Annu Rev Genomics Hum Genet 17, 95-115.
Liu, G.H., Shen, H.B., and Yu, D.J. (2016). Prediction of Protein-Protein Interaction Sites with Machine-Learning-Based Data-Cleaning and Post-Filtering Procedures. The Journal of membrane biology 249, 141-153.
Lupas, A., Van Dyke, M., and Stock, J. (1991). Predicting coiled coils from protein sequences. Science 252, 1162-1164.
Luscombe, N.M., Austin, S.E., Berman, H.M., and Thornton, J.M. (2000). An overview of the structures of protein-DNA complexes. Genome Biol 1, REVIEWS001.
Ma, X., Guo, J., Liu, H.D., Xie, J.M., and Sun, X. (2012). Sequence-based prediction of DNA-binding residues in proteins with conservation and correlation information. IEEE/ACM Trans Comput Biol Bioinform 9, 1766-1775.
Mahlich, Y., Reeb, J., Hecht, M., Schelling, M., De Beer, T.A.P., Bromberg, Y., and Rost, B. (2017). Common sequence variants affect molecular function more than rare variants? Sci Rep 7, 1608.
Mayor, S. (2007). Genome sequence of one individual is published for first time. BMJ 335, 530-531.
Mikolov, T., Sutskever, I., Chen, K., Corrado, G.s., and Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. Advances in Neural Information Processing Systems 26.
Mishra, A., Pokhrel, P., and Hoque, M.T. (2019). StackDPPred: a stacking based prediction of DNA-binding protein from sequence. Bioinformatics 35, 433-441.
Muckenthaler, M., Gray, N.K., and Hentze, M.W. (1998). IRP-1 binding to ferritin mRNA prevents the recruitment of the small ribosomal subunit by the cap-binding complex eIF4F. Mol Cell 2, 383-388.
Mukherjee, S., and Zhang, Y. (2011). Protein-protein complex structure predictions by multimeric threading and template recombination. Structure 19, 955-966.
Ng, P.C., Henikoff, J.G., and Henikoff, S. (2000). PHAT: a transmembrane-specific substitution matrix. Predicted hydrophobic and transmembrane. Bioinformatics 16, 760-766.
Ng, P.C., and Henikoff, S. (2003). SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Res 31, 3812-3814.
Nielsen, H., Engelbrecht, J., Brunak, S., and von Heijne, G. (1997). Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. Protein Eng 10, 1-6.
Nishino, T., and Morikawa, K. (2002). Structure and function of nucleases in DNA repair: shape, grip and blade of the DNA scissors. Oncogene 21, 9022-9032.
Nooren, I.M., and Thornton, J.M. (2003). Structural characterisation and functional significance of transient protein-protein interactions. J Mol Biol 325, 991-1018.
Northey, T.C., Baresic, A., and Martin, A.C.R. (2018). IntPred: a structure-based predictor of protein-protein interaction sites. Bioinformatics 34, 223-229.
Ofran, Y., and Rost, B. (2003). Predicted protein-protein interaction sites from local sequence information. FEBS Lett 544, 236-239.

Ofran, Y., and Rost, B. (2007). ISIS: interaction sites identified from sequence. Bioinformatics 23, e13-16.
Pal, G., Kouadio, J.L., Artis, D.R., Kossiakoff, A.A., and Sidhu, S.S. (2006). Comprehensive and quantitative mapping of energy landscapes for protein-protein interactions by rapid combinatorial scanning. J Biol Chem 281, 22378-22385.
Peng, Z., and Kurgan, L. (2015). High-throughput prediction of RNA, DNA and protein binding regions mediated by intrinsic disorder. Nucleic acids research 43, e121.
Peng, Z., Wang, C., Uversky, V.N., and Kurgan, L. (2017). Prediction of Disordered RNA, DNA, and Protein Binding Regions Using DisoRDPbind. Methods Mol Biol 1484, 187-203.
Petschnigg, J., Snider, J., and Stagljar, I. (2011). Interactive proteomics research technologies: recent applications and advances. Curr Opin Biotechnol 22, 50-58.
Qiu, J., Bernhofer, M., Heinzinger, M., Kemper, S., Norambuena, T., Melo, F., and Rost, B. (2020). ProNA2020 predicts protein-DNA, protein-RNA, and protein-protein binding proteins and residues from sequence. J Mol Biol 432, 2428-2443.
Ramensky, V., Bork, P., and Sunyaev, S. (2002). Human non-synonymous SNPs: server and survey. Nucleic Acids Res 30, 3894-3900.
Re, A., Joshi, T., Kulberkyte, E., Morris, Q., and Workman, C.T. (2014). RNA-protein interactions: an overview. Methods Mol Biol 1097, 491-521.
Redon, C., Pilch, D., Rogakou, E., Sedelnikova, O., Newrock, K., and Bonner, W. (2002). Histone H2A variants H2AX and H2AZ. Curr Opin Genet Dev 12, 162-169.
Reichmann, D., Rahat, O., Cohen, M., Neuvirth, H., and Schreiber, G. (2007). The molecular architecture of protein-protein binding sites. Curr Opin Struct Biol 17, 67-76.
Res, I., Mihalek, I., and Lichtarge, O. (2005). An evolution based classifier for prediction of protein interfaces without using protein structures. Bioinformatics 21, 2496-2501.
Rost, B., and Sander, C. (1993). Improved prediction of protein secondary structure by use of sequence profiles and neural networks. Proc Natl Acad Sci U S A 90, 7558-7562.
Rost, B., and Sander, C. (1994). Combining evolutionary information and neural networks to predict protein secondary structure. Proteins 19, 55-72.
Sarkar, A., and Nandineni, M.R. (2018). Association of common genetic variants with human skin color variation in Indian populations. Am J Hum Biol 30.
Schlessinger, A., Yachdav, G., and Rost, B. (2006). PROFbval: predict flexible and rigid residues in proteins. Bioinformatics 22, 891-893.
Shatsky, M., Nussinov, R., and Wolfson, H.J. (2004). A method for simultaneous alignment of multiple protein structures. Proteins 56, 143-156.
Singh, M., and Tyagi, S.C. (2018). Genes and genetics in eye diseases: a genomic medicine approach for investigating hereditary and inflammatory ocular disorders. Int J Ophthalmol 11, 117-134.
Steinthorsdottir, V., Thorleifsson, G., Sulem, P., Helgason, H., Grarup, N., Sigurdsson, A., Helgadottir, H.T., Johannsdottir, H., Magnusson, O.T., Gudjonsson, S.A., et al. (2014). Identification of low-frequency and rare sequence variants associated with elevated or reduced risk of type 2 diabetes. Nat Genet 46, 294-298.
Styrkarsdottir, U., Thorleifsson, G., Sulem, P., Gudbjartsson, D.F., Sigurdsson, A., Jonasdottir, A., Jonasdottir, A., Oddsson, A., Helgason, A., Magnusson, O.T., et al. (2013). Nonsense mutation in the LGR4 gene is associated with several human diseases and other traits. Nature 497, 517-520.

Su, H., Liu, M., Sun, S., Peng, Z., and Yang, J. (2019). Improving the prediction of protein-nucleic acids binding residues via multiple sequence profiles and the consensus of complementary methods. Bioinformatics 35, 930-936.
Sui, H., Chen, Q., and Imamichi, T. (2020). A pull-down assay using DNA/RNA-conjugated beads with a customized competition strategy: An effective approach to identify DNA/RNA binding proteins. MethodsX 7, 100890.
Szilagyi, A., and Skolnick, J. (2006). Efficient prediction of nucleic acid binding function from low-resolution protein structures. Journal of molecular biology 358, 922-933.
Tarpey, P.S., Raymond, F.L., Nguyen, L.S., Rodriguez, J., Hackett, A., Vandeleur, L., Smith, R., Shoubridge, C., Edkins, S., Stevens, C., et al. (2007). Mutations in UPF3B, a member of the nonsense-mediated mRNA decay complex, cause syndromic and nonsyndromic mental retardation. Nat Genet 39, 1127-1133.
The UniProt, C. (2017). UniProt: the universal protein knowledgebase. Nucleic Acids Res 45, D158-D169.
Thomas, P.D., Campbell, M.J., Kejariwal, A., Mi, H., Karlak, B., Daverman, R., Diemer, K., Muruganujan, A., and Narechania, A. (2003). PANTHER: a library of protein families and subfamilies indexed by function. Genome Res 13, 2129-2141.
van Straalen, N.M., and Roelofs, T.F.M. (2006). An Introduction to Ecological Genomics (Oxford: Oxford University press).
Viswanathan, R., Fajardo, E., Steinberg, G., Haller, M., and Fiser, A. (2019). Protein-protein binding supersites. PLoS Comput Biol 15, e1006704.
Walia, R.R., Xue, L.C., Wilkins, K., El-Manzalawy, Y., Dobbs, D., and Honavar, V. (2014). RNABindRPlus: a predictor that combines machine learning and sequence homology-based methods to improve the reliability of predicted RNA-binding residues in proteins. PloS one 9, e97725.
Wang, B., Chen, P., Huang, D.S., Li, J.J., Lok, T.M., and Lyu, M.R. (2006). Predicting protein interaction sites from residue spatial sequence profile and evolution rate. FEBS Lett 580, 380-384.
Wang, E.T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., Kingsmore, S.F., Schroth, G.P., and Burge, C.B. (2008). Alternative isoform regulation in human tissue transcriptomes. Nature 456, 470-476.
Wang, G.S., and Cooper, T.A. (2007). Splicing in disease: disruption of the splicing code and the decoding machinery. Nat Rev Genet 8, 749-761.
Wei, Z.S., Yang, J.Y., Shen, H.B., and Yu, D.J. (2015). A Cascade Random Forests Algorithm for Predicting Protein-Protein Interaction Sites. IEEE Trans Nanobioscience 14, 746-760.
Yachdav, G., Kloppmann, E., Kajan, L., Hecht, M., Goldberg, T., Hamp, T., Honigschmid, P., Schafferhans, A., Roos, M., Bernhofer, M., et al. (2014). PredictProtein--an open resource for online prediction of protein structural and functional features. Nucleic Acids Res 42, W337-343.
Yan, J., Friedrich, S., and Kurgan, L. (2016). A comprehensive comparative review of sequence-based predictors of DNA- and RNA-binding residues. Brief Bioinform 17, 88-105.
Yan, J., and Kurgan, L. (2017). DRNApred, fast sequence-based method that accurately predicts and discriminates DNA- and RNA-binding residues. Nucleic Acids Res 45, e84.
Yang, Y., Zhao, H., Wang, J., and Zhou, Y. (2014). SPOT-Seq-RNA: predicting protein-RNA complex structure and RNA-binding function by fold recognition and binding affinity prediction. Methods in molecular biology 1137, 119-130.

Zhang, J., Ma, Z., and Kurgan, L. (2017). Comprehensive review and empirical analysis of hallmarks of DNA-, RNA- and protein-binding residues in protein chains. Briefings in bioinformatics.
Zhang, J., Ma, Z., and Kurgan, L. (2019). Comprehensive review and empirical analysis of hallmarks of DNA-, RNA- and protein-binding residues in protein chains. Brief Bioinform 20, 1250-1268.
Zhang, X., and Liu, S. (2017). RBPPred: predicting RNA-binding proteins from sequence using SVM. Bioinformatics 33, 854-862.
Zhi-Sen Wei, K.H., Jing-Yu Yang,Hong-Bin Shen,Dong-Jun Yu (2016). Protein-protein interaction sites prediction by ensembling SVM and sample-weighted random forests. Neurocomput 193, 201-212.