A Novel Approach to Predict Core Residues on Cancer-Related DNA-Binding Domains

Ka-Chun Wong

Department of Computer Science, City University of Hong Kong, Kowloon Tong, Hong Kong.

Supplementary Issue: Integrative Analysis of Cancer Genomic Data

ABSTRACT: Protein–DNA interactions are involved in different cancer pathways. In particular, the DNA-binding domains of proteins can determine where and how gene regulatory regions are bound in different cell lines at different stages. Therefore, it is essential to develop a method to predict and locate the core residues on cancer-related DNA-binding domains. In this study, we propose a computational method to predict and locate core residues on DNA-binding domains. In particular, we have selected the cancer-related DNA-binding domains for in-depth studies, namely, winged Helix Turn Helix family, homeodomain family, and basic Helix-Loop-Helix family. The results demonstrate that the proposed method can predict the core residues involved in protein–DNA interactions, as verified by the existing structural data. Given its good performance, various aspects of the method are discussed and explored: for instance, different uses of prediction algorithm, different protein domains, and hotspot threshold setting.

KEYWORDS: SNPdryad, Protein-DNA binding interactions, applied machine learning, DNA-binding domains, bioinformatics, cancer, big data analytics

SUPPLEMENT: Integrative Analysis of Cancer Genomic Data

CITATION: Wong. A Novel Approach to Predict Core Residues on Cancer-Related DNA-Binding Domains. *Cancer Informatics* 2016:15(S2) 1–7 doi: 10.4137/CIN.S39366.

TYPE: Original Research

RECEIVED: February 24, 2016. **RESUBMITTED:** May 04, 2016. **ACCEPTED FOR PUBLICATION:** May 08, 2016.

ACADEMIC EDITOR: J. T. Efird, Editor in Chief

PEER REVIEW: Four peer reviewers contributed to the peer review report. Reviewers' reports totaled 1,017 words, excluding any confidential comments to the academic editor.

FUNDING: The work described in this article was partially supported by the grant from City University of Hong Kong (Project No. 7200444/CS) and Amazon Web Service Research Grant. The author confirms that the funders had no influence over the study design, content of the article, or selection of this journal.

COMPETING INTERESTS: Author discloses no potential conflicts of interest.

CORRESPONDENCE: kc.w@cityu.edu.hk

Introduction

The protein–DNA binding interactions are essential activities in gene transcription. In recent years, it has been recognized that gene transcription plays a more significant role than previously thought in the context of protein level control.[1] Thus, there is increasing interest in deciphering the protein–DNA binding interactions, which can be determined by the bacterial one-hybrid system in the past.

With the advent of next-generation sequencing and other modern biotechnology, the protein–DNA binding interaction studies have been accelerated from single DNA-binding protein study to the proteome-wide level; for instance, Weirauch et al have applied protein-binding microarray to comprehensively determine the eukaryotic transcription factor DNA-binding specificity in sequence level.[2] Wong et al have applied multiple expectation maximization for Motif Elicitation[3] to generate the coupling DNA motifs on chromatin interactions in human being.[4] Jolma et al have also applied systematic evolution of ligands by exponential enrichment and chromatin immunoprecipitation sequencing to characterize the DNA-binding specificities of human transcription factors, resulting in 239 distinct binding profiles.[5] The sequence and chromatin determinants surrounding protein–DNA binding interactions have also been studied in the context of transcription factors across different cell lines comprehensively.[6] Tremendous data have been accumulated with the potential for deciphering protein–DNA binding interactions further.

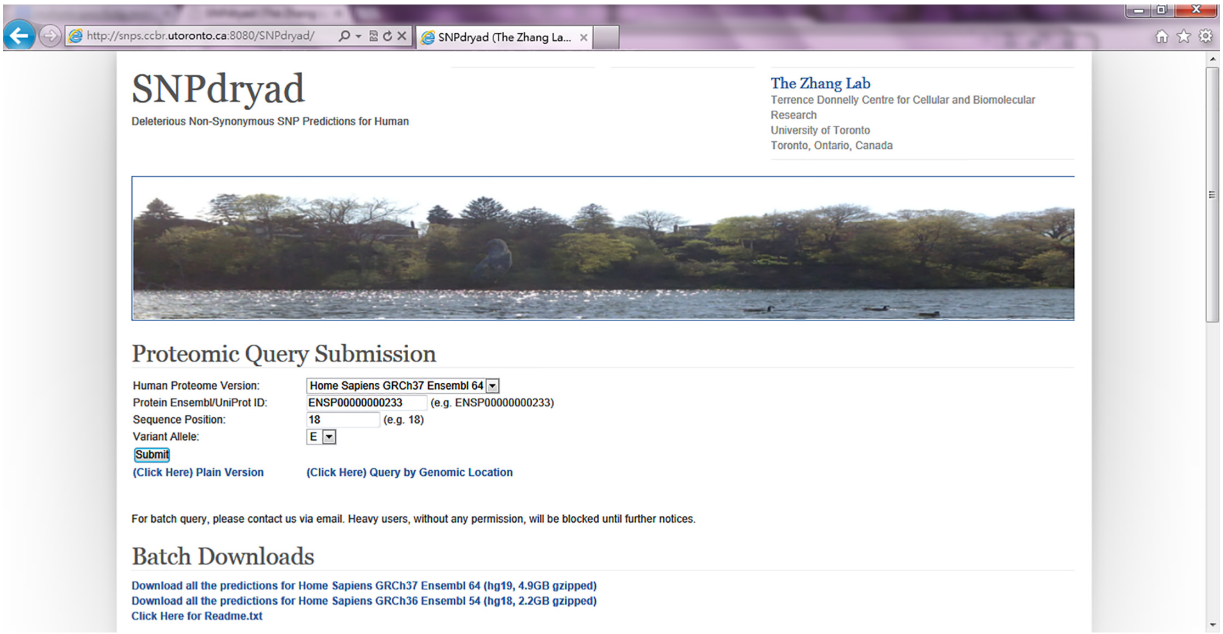
Therefore, it is important to harness and leverage the existing big data to shed proteome-wide lights on the protein–DNA-binding studies; for instance, Wong et al have proposed a computational framework to learn and predict the specificity-determining residue–nucleotide interactions across different DNA-binding families.[7] Pelossof et al have also proposed an approach for learning the recognition models across different DNA-binding families.[8] Wong et al have proposed an evolutionary computational approach for learning the combinatorial protein–DNA-binding sequence patterns.[9]

Objective

In this study, we propose to adopt SNPdryad[10] for the prediction of DNA-binding residues. In other words, given a protein sequence, we would like to predict and locate which residues are DNA binding, as verified by its protein–DNA complex structural information from Protein Data Bank (PDB).

Methodology

To predict the DNA-binding residues for a specific DNA-binding domain family, we are interested in the positional distribution of the harmful nonsynonymous single-nucleotide polymorphisms (nsSNPs) across the DNA-binding domain



http://snps.ccbr.utoronto.ca:8080/SNPdryad/

SNPdryad (The Zhang La...

SNPdryad

Deleterious Non-Synonymous SNP Predictions for Human

The Zhang Lab
Terrence Donnelly Centre for Cellular and Biomolecular Research
University of Toronto
Toronto, Ontario, Canada

Proteomic Query Submission

Human Proteome Version: Home Sapiens GRCh37 Ensembl 64
Protein Ensembl/UniProt ID: ENSP00000000233 (e.g. ENSP00000000233)
Sequence Position: 18 (e.g. 18)
Variant Allele: E

Submit

(Click Here) Plain Version (Click Here) Query by Genomic Location

For batch query, please contact us via email. Heavy users, without any permission, will be blocked until further notices.

Batch Downloads

Download all the predictions for Home Sapiens GRCh37 Ensembl 64 (hg19, 4.9GB gzipped)
Download all the predictions for Home Sapiens GRCh36 Ensembl 54 (hg18, 2.2GB gzipped)
Click Here for Readme.txt

Figure 1. Snapshot of the SNPdryad website in June 2014. The users can download the entire predictions of the human proteomes hg19 and hg18.

family because nsSNPs are known to be able to directly alter the encoded protein functions. Therefore, we have performed multiple sequence alignment on the protein-coding sequences for each domain and count the number of harmful nsSNPs on these sequences.

The algorithm and the prediction results can be accessed from the website.[a] A snapshot of the website is shown in Figure 1. In particular, we would like to note that the precise number of deleterious amino acid substitutions on the human proteome is yet to be determined. Thus, the website is designed to remind the users about it when they query the SNPdryad website as shown in Figure A1.

In particular, we would like to note that SNPdryad is trained on the existing literature annotations (Harvard HumDiv dataset) that may be limited in size if we consider all the possible amino acid substitutions on a human proteome, although this is also the approach the existing methods (eg, Harvard PolyPhen2) have taken.

Results

Fully deleterious substitutions. In total, we scanned 92,012 human proteins (including protein isoforms) and 36,935,804 amino acid positions; a total of 10,120,155 substitutions (about 1.4%) were predicted to be fully deleterious (with the SNPdryad prediction score of 1).

DNA-binding family choices. In this study, we have selected three DNA-binding protein families for in-depth studies. (1) The ETS domain has been selected (Pfam ID: PF00178) since it is important in different tissue developments and cancer progression for metazoans.[11] (2) The homeodomain family is selected because it was demonstrated to be related to multiple cancers: breast cancer,[12] prostate cancer,[13] and non-muscle invasive bladder cancers.[14] (3) The basic Helix-Loop-Helix (bHLH) domain family is selected because its proteins

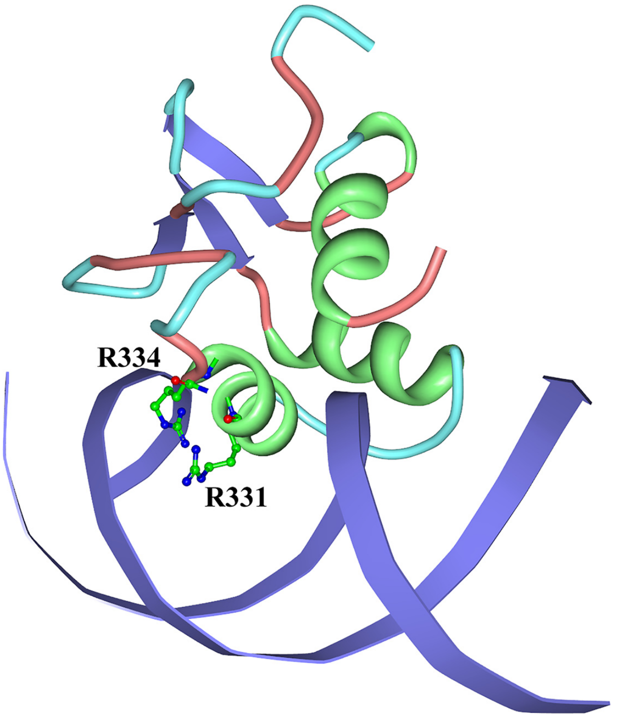


Figure 2. Crystal structure of mouse Elf3 C-terminal DNA-binding domain in complex with type II TGF-beta receptor promoter DNA (PDB ID: 3 JTG).[16] It is drawn using Protein Workshop.[21] The harmful nsSNP hotspots in Figure 3 are mapped back to the sequence positions of the structure, namely, R331 and R334.

[a] http://snps.ccbr.utoronto.ca:8080/SNPdryad/

are found to sense and respond to environmental stimulus in different cancer-related pathways.[15]

Winged Helix Turn Helix family. The ETS domain has been selected (Pfam ID: PF00178) since it is a large transcription factor family, which is important in different tissue developments and cancer progression for metazoans.[11] The ETS domain belongs to the winged Helix Turn Helix domain family. It is a DNA-binding domain that has three alpha helices and four beta strands (eg, PF03444). Especially, the domain is characterized by the alternative intervention between the helices and strands.

In particular, we extracted the amino acid sequences annotated as the ETS domains in human proteome. The positions of the harmful nsSNPs were mapped onto each sequence. To provide biological insights, we also downloaded the crystal structure of mouse Elf3 C-terminal DNA-binding domain in complex with type II TGF-beta receptor promoter DNA (PDB ID: 3 JTG).[16] The amino acid chain in

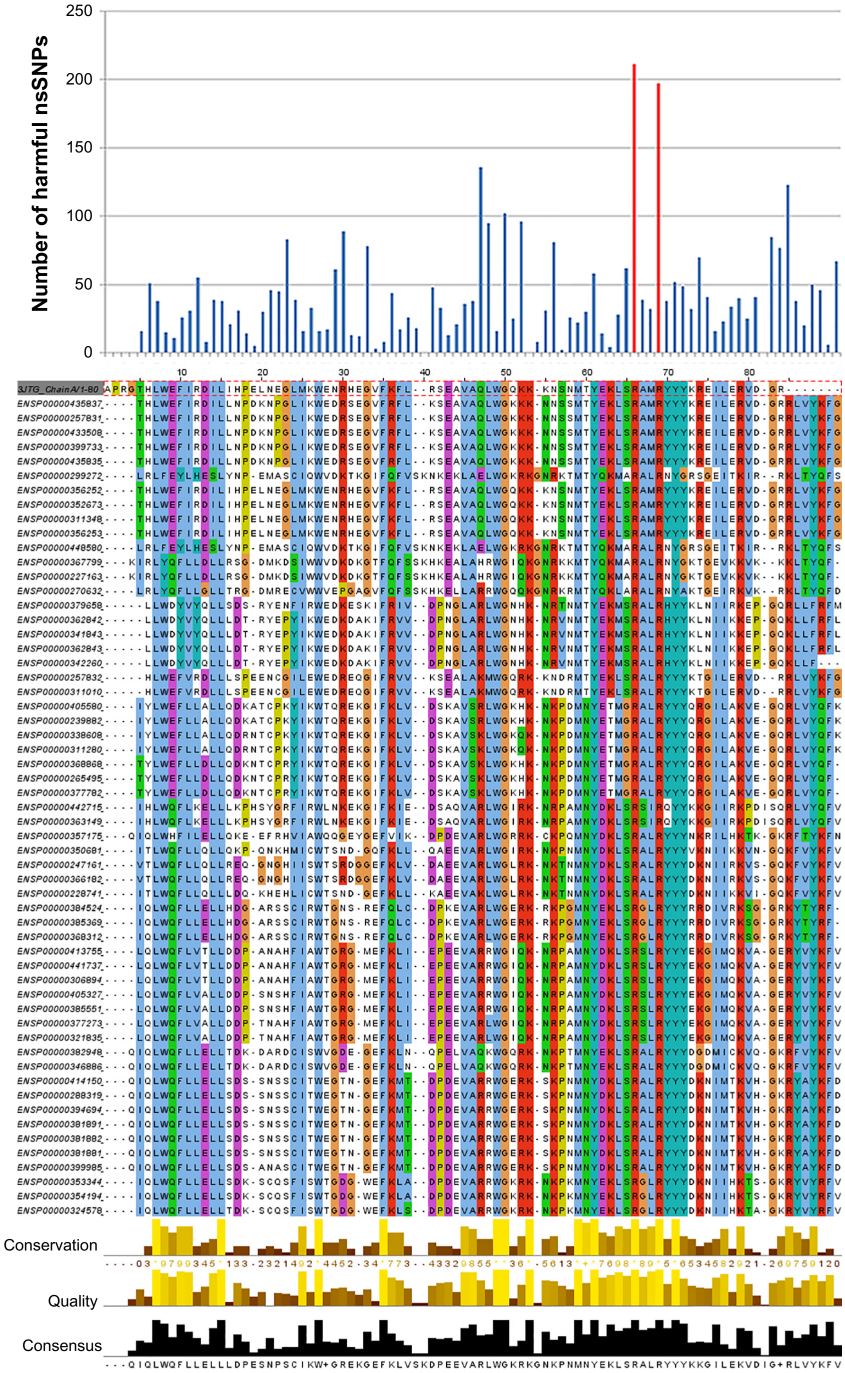


Figure 3. Multiple sequence alignment for the ETS-domain (Pfam ID: PF00178) sequences in human and the amino acid sequence extracted from the ETS-domain structure (PDB ID: 3 JTG Chain A), colored in the Clustalx color scheme.

Notes: The histogram shows the number of harmful nsSNPs for each alignment position. The peaks with *z*-scores >3 are colored in red, while the others are colored in blue.

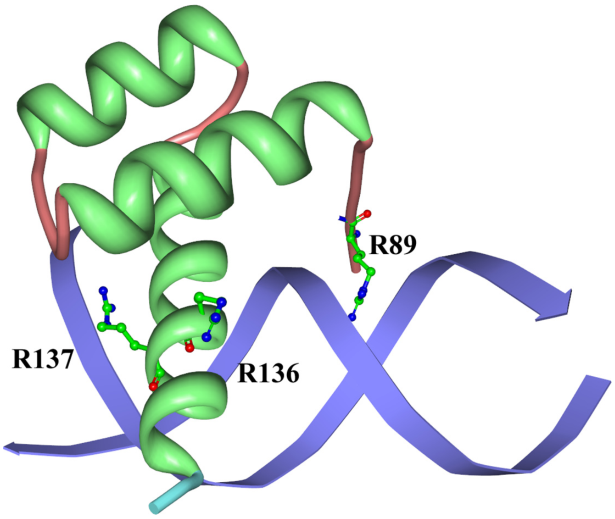


Figure 4. Three-dimensional structure of Aristaless homeodomain in complex with DNA (PDB ID: 3 LNQ).[22] It is drawn using Protein Workshop.[21]
Note: The harmful nsSNP hotspots in Figure 5 are mapped back to the sequence positions of the structure, namely, R89, R136, and R137.

that structure was aligned with the ETS domain sequences using MUSCLE with the default parameter setting.[17] The resultant multiple sequence alignment and the mapped number of harmful nsSNPs are depicted in Figure 3. Interestingly, it can be observed that the position distribution of the harmful nsSNPs is not uniform across the domain. To distinguish harmful nsSNP hot spots (tall peaks) from the rest, the mean and standard deviation of the number of harmful nsSNPs are calculated. A position is called a harmful nsSNP hot spot when the number of harmful nsSNPs exceeds the mean plus three standard deviations (ie, *z*-scores >3). Based on such a setting, two alignment positions are found as the hotspots. They are highlighted in red. It can be observed that those harmful nsSNP hot spots are located at the conserved positions. If we map the positions back to the structural data as shown in Figure 2, interestingly, they correspond to the region where the protein-DNA binding occurs, explaining the enrichment of harmful nsSNPs predicted by SNPdryad.

Homeodomain family. The homeodomain family is a DNA-binding domain that has three alpha helices connected by short loop structures (eg, PF00046). The defining feature is that one of the alpha helix is found nearly perpendicular to the plane formed by the other two alpha helices. The HOX genes (which belong to the homeodomain family) have been demonstrated to be related to different human cancers[18]; for instance, the induction of HOXA5 can cause around 300 genes to be

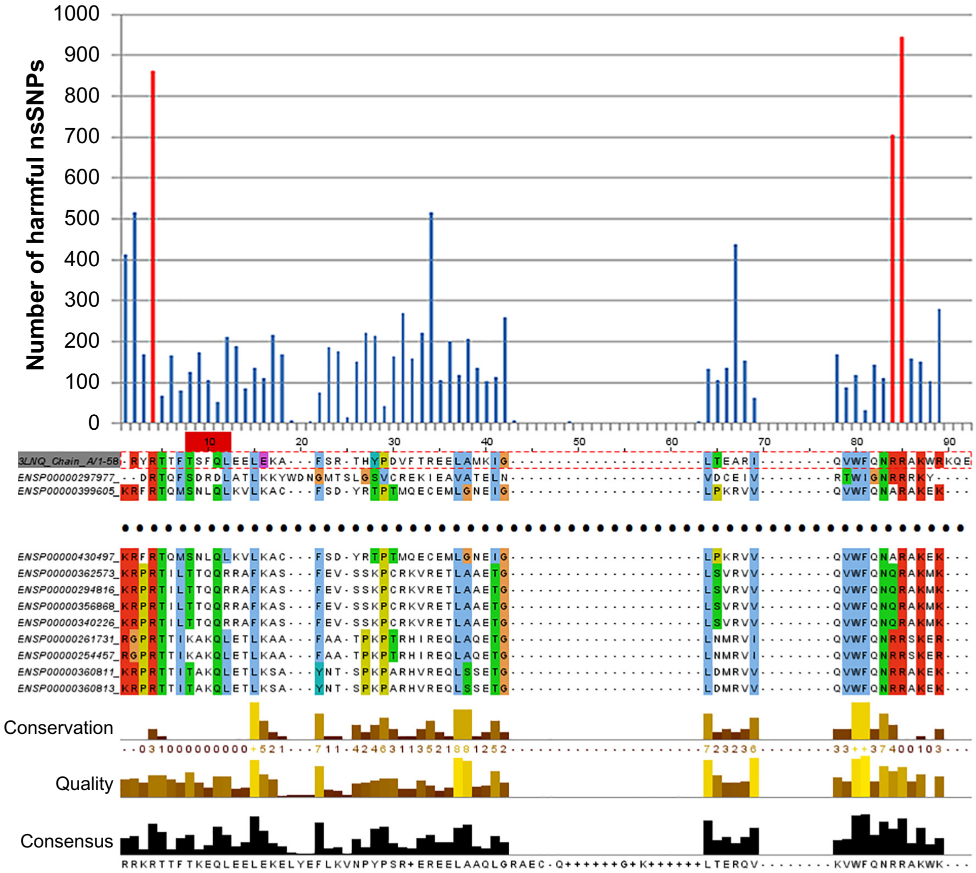


Figure 5. Multiple sequence alignment for the homeodomain domain (Pfam ID: PF00048) sequences in human and the amino acid sequence extracted from the homeodomain structure (PDB ID: 3 LNQ Chain A), colored in the Clustalx color scheme.
Notes: The histogram shows the number of harmful nsSNPs for each alignment position. The peaks with *z*-scores >3 are colored in red, while the others are colored in blue.

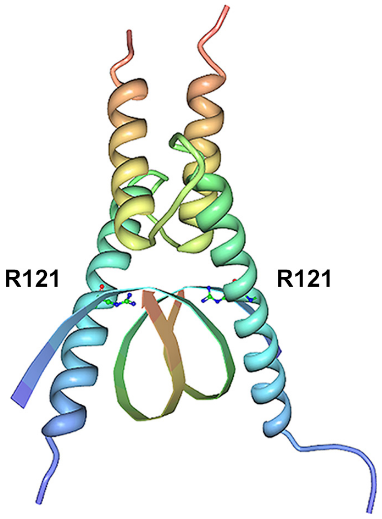


Figure 6. Crystal structure of MyoD bHLH domain bound to DNA (PDB ID: 1MDY).[23] It is drawn using Protein Workshop.[21]
Note: The harmful nsSNP hotspot in Figure 7 is mapped back to the sequence position of the structure, namely, R121.

unregulated in breast cancer cell lines[12]; the overexpression of HOXC8 was found to be related to the loss of differentiation in prostate cancer cells[13]; HOXA9 was demonstrated to be an independent indicator of prognosis in nonmuscle invasive bladder cancers.[14]

As an illustrative example, we have selected the structure of Aristaless homeodomain (PDB ID: 3 LNQ) as shown in Figure 4. Similar to the previous section, we have exhaustively extracted all homeodomain sequences in human being and performed the multiple sequence alignment. Once aligned, we count how many harmful nsSNPs can be found at each alignment position as shown in Figure 5. To distinguish harmful nsSNP hot spots (tall peaks) from the rest, the mean and standard deviation of the number of harmful nsSNPs are calculated. A position is called a harmful nsSNP hot spot when the number of harmful nsSNPs exceeds the mean plus three standard deviations (ie, *z*-scores > 3). Based on such a setting, three hotspots are identified. If we map the three hotspots back to the structure in Figure 4, we can

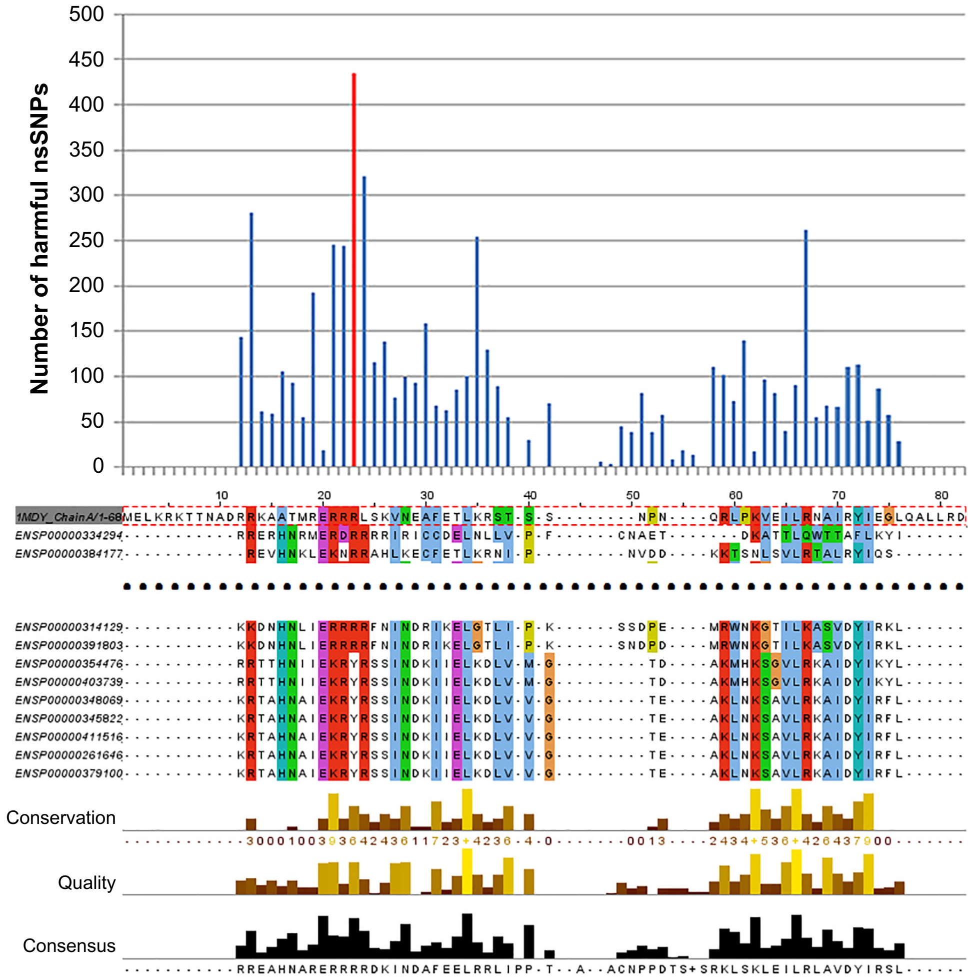


Figure 7. Multiple sequence alignment for the bHLH domain (Pfam ID: PF00010) sequences in human and the amino acid sequence extracted from the bHLH structure (PDB ID: 1MDY Chain A), colored in the Clustalx color scheme.
Notes: The histogram shows the number of harmful nsSNPs for each alignment position. The peak with *z*-scores >3 is colored in red, while the others are colored in blue.

observe that the core DNA-binding residues are predicted as hotspots.

bHLH family. The bHLH domain family is a DNA-binding domain that consists of two alpha helices connected by a loop structure (eg, PF00010). In most cases, they bind to DNA as a dimer (ie, two domains). In addition, they are shown to be related to different cancers; for instance, the transcription factors c-Myc and HIF-1 can interact with each other and promote metabolic advantages to tumor cells[19] and adjust adaptive responses to hypoxic environments.[20] bHLH-PAS proteins are also believed as multitasking family of transcription factors that can sense and respond to environmental stimulus in the cancer-related pathways.[15]

Therefore, it is very important to identify the core DNA-binding residues for the bHLH family. In particular, we have selected the crystal structure of MyoD bHLH domain as an example visualized in Figure 6. Similar to the previous sections, we have collected and aligned all the bHLH sequences from the human proteome, resulting in the multiple sequence alignment profile as shown in Figure 7. It can be observed that only one alignment position remains after setting the *z*-scores cutoff to 3. If we map that position back to the structure, we can observe that it is indeed a core DNA-binding residue position as shown in Figure 6.

Discussion

Gene regulation is a very important step in genetics. In particular, gene transcription is responsible for nearly 70% contribution to protein levels.[1] Therefore, it is very important to decipher the gene transcription mechanism in which the protein–DNA-binding mechanism plays a significant role.

To this end, we have proposed a novel approach to predict the core DNA-binding residues on the cancer-related proteins. We propose that SNPdryad can be integrated and transformed to predict and locate core DNA-binding residues for cancer-related protein studies. The results suggest that the approach is feasible and can be explored further.

In the future, we seek to have comprehensive benchmarking on the prediction performance than the existing case studies, although the additional computational burden has to be provisioned carefully. Another interesting direction is to see if other prediction algorithms such as PolyPhen2 and MutationTaster can be applied in a similar fashion. It would be helpful if the existing methods can complement to each other, contributing to accurate ensemble prediction of DNA-binding residues on cancer-related proteins. The peak threshold setting is another interesting direction to be explored for the proposed method in the future.

Author Contributions

Conceived and designed the experiments: KW. Analyzed the data: KW. Wrote the first draft of the manuscript: KW. Developed the structure and arguments for the paper: KW. Made critical revisions: KW. The author reviewed and approved of the final manuscript.

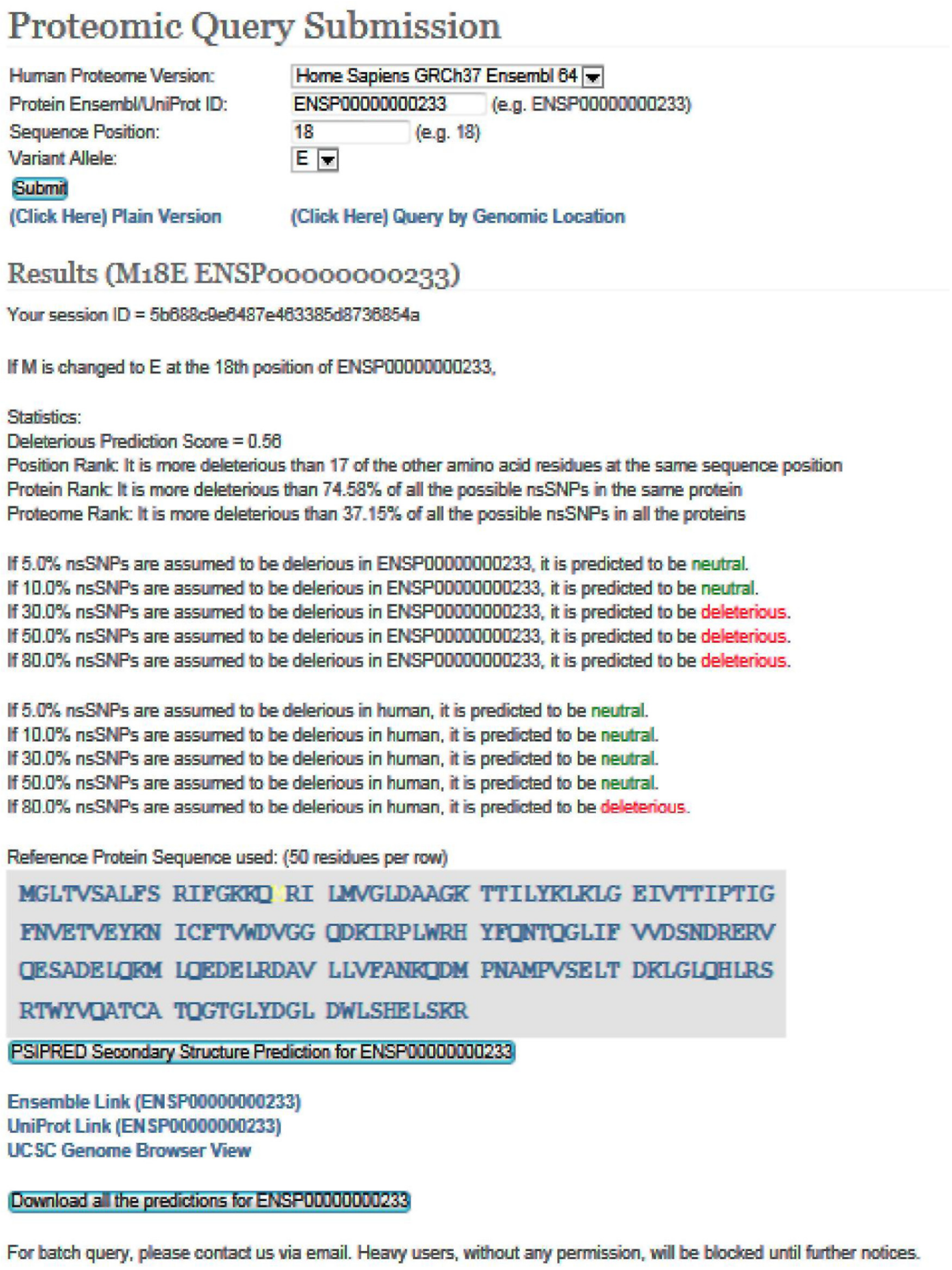
Acknowledgments

The author thanks the four anonymous reviewers for their constructive comments. The author also thanks Tak-Ming Chan for his works in surveying on protein–DNA binding. Last but not the least, the author thanks Zhaolei Zhang for his insightful comments.

REFERENCES

1. Li JJ, Biggin MD. Gene expression. Statistics requantitates the central dogma. *Science*. 2015;347(6226):1066–7.
2. Weirauch MT, Yang A, Albu M, et al. Determination and inference of eukaryotic transcription factor sequence specificity. *Cell*. 2014;158(6):1431–43.
3. Bailey TL, Johnson J, Grant CE, Noble WS. The MEME suite. *Nucleic Acids Res*. 2015;43(W1):39–49.
4. Wong KC, Li Y, Peng C. Identification of coupling DNA motif pairs on long-range chromatin interactions in human K562 cells. *Bioinformatics*. 2016;32(3):321–4.
5. Jolma A, Yan J, Whitington T, et al. DNA-binding specificities of human transcription factors. *Cell*. 2013;152(1):327–39.
6. Arvey A, Agius P, Noble WS, Leslie C. Sequence and chromatin determinants of cell-type-specific transcription factor binding. *Genome Res*. 2012;22(9):1723–34.
7. Wong KC, Li Y, Peng C, Moses AM, Zhang Z. Computational learning on specificity-determining residue-nucleotide interactions. *Nucleic Acids Res*. 2015;43(21):10180–9.
8. Pelossof R, Singh I, Yang JL, Weirauch MT, Hughes TR, Leslie CS. Affinity regression predicts the recognition code of nucleic acid-binding proteins. *Nat Biotechnol*. 2015;33(12):1242–9.
9. Wong KC, Peng C, Wong MH, Leung KS. Generalizing and learning protein-dna binding sequence representations by an evolutionary algorithm. *Soft Comput*. 2011;15(8):1631–42.
10. Wong KC, Zhang Z. Snpdryad: predicting deleterious non-synonymous human snps using only orthologous protein sequences. *Bioinformatics*. 2014;30(8):1112–9.
11. Sharrocks AD. The ETS-domain transcription factor family. *Nat Rev Mol Cell Biol*. 2001;2(11):827–37.
12. Arderiu G, Cuevas I, Chen A, Carrio M, East L, Boudreau NJ. HoxA5 stabilizes adherens junctions via increased Akt1. *Cell Adh Migr*. 2007;1(4):185–95.
13. Waltregny D, Alami Y, Clausse N, de Leval J, Castronovo V. Over-expression of the homeobox gene HOXC8 in human prostate cancer correlates with loss of tumor differentiation. *Prostate*. 2002;50(3):162–9.
14. Kim YJ, Yoon HY, Kim JS, et al. HOXA9, ISL1 and ALDH1A3 methylation patterns as prognostic markers for nonmuscle invasive bladder cancer: array-based DNA methylation and expression profiling. *Int J Cancer*. 2013;133(5):1135–42.
15. Bersten DC, Sullivan AE, Peet DJ, Whitelaw ML. bhlh-pas proteins in cancer. *Nature Reviews Cancer*. 2013;13(12):827–41.
16. Agarkar VB, Babayeva ND, Wilder PJ, Rizzino A, Tahirov TH. Crystal structure of mouse Elf3 C-terminal DNA-binding domain in complex with type II TGF-beta receptor promoter DNA. *J Mol Biol*. 2010;397(1):278–89.
17. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*. 2004;32(5):1792–7.
18. Bhatlekar S, Fields JZ, Boman BM. HOX genes and their role in the development of human cancers. *J Mol Med*. 2014;92(8):811–23.
19. Dang CV, Kim JW, Gao P, Yustein J. The interplay between MYC and HIF in cancer. *Nat Rev Cancer*. 2008;8(1):51–6.
20. Gordan JD, Thompson CB, Simon MC. HIF and c-Myc: sibling rivals for control of cancer cell metabolism and proliferation. *Cancer Cell*. 2007;12(2):108–13.
21. Moreland JL, Gramada A, Buzko OV, Zhang Q, Bourne PE. The Molecular Biology Toolkit (MBT): a modular platform for developing molecular visualization applications. *BMC Bioinformatics*. 2005;6:21.
22. Miyazono K, Zhi Y, Takamura Y, et al. Cooperative DNA-binding and sequence-recognition mechanism of aristaless and clawless. *EMBO J*. 2010;29(9):1613–23.
23. Ma PC, Rould MA, Weintraub H, Pabo CO. Crystal structure of MyoD bHLH domain-DNA complex: perspectives on DNA recognition and implications for transcriptional activation. *Cell*. 1994;77(3):451–9.

Appendix



Proteomic Query Submission

Human Proteome Version: Homo Sapiens GRCh37 Ensembl 64
Protein Ensembl/UniProt ID: ENSP00000000233 (e.g. ENSP00000000233)
Sequence Position: 18 (e.g. 18)
Variant Allele: E

Submit
(Click Here) Plain Version
(Click Here) Query by Genomic Location

Results (M18E ENSP00000000233)

Your session ID = 5b688c9e6487e463385d8736854a

If M is changed to E at the 18th position of ENSP00000000233,

Statistics:
Deleterious Prediction Score = 0.56
Position Rank: It is more deleterious than 17 of the other amino acid residues at the same sequence position
Protein Rank: It is more deleterious than 74.58% of all the possible nsSNPs in the same protein
Proteome Rank: It is more deleterious than 37.15% of all the possible nsSNPs in all the proteins

If 5.0% nsSNPs are assumed to be delerious in ENSP00000000233, it is predicted to be neutral.
If 10.0% nsSNPs are assumed to be delerious in ENSP00000000233, it is predicted to be neutral.
If 30.0% nsSNPs are assumed to be delerious in ENSP00000000233, it is predicted to be deleterious.
If 50.0% nsSNPs are assumed to be delerious in ENSP00000000233, it is predicted to be deleterious.
If 80.0% nsSNPs are assumed to be delerious in ENSP00000000233, it is predicted to be deleterious.

If 5.0% nsSNPs are assumed to be delerious in human, it is predicted to be neutral.
If 10.0% nsSNPs are assumed to be delerious in human, it is predicted to be neutral.
If 30.0% nsSNPs are assumed to be delerious in human, it is predicted to be neutral.
If 50.0% nsSNPs are assumed to be delerious in human, it is predicted to be neutral.
If 80.0% nsSNPs are assumed to be delerious in human, it is predicted to be deleterious.

Reference Protein Sequence used: (50 residues per row)

```
MGLTVSALFS RIFGKKQMRI LMVGLDAAGK TTILYKLKLG EIVTTIPTIG
FNVETVEYKN ICFTVWDVGG QDKIRPLWRH YFQNTQGLIF VVDSNDRERV
QESADELQKM LQEDELRDAV LLVFANKQDM PNAMPVSELT DKLGLQHLRS
RTWYVQATCA TQGTGLYDGL DWLSHELSKR
```

Ensemble Link (ENSP00000000233)
UniProt Link (ENSP00000000233)
UCSC Genome Browser View

For batch query, please contact us via email. Heavy users, without any permission, will be blocked until further notices.

Figure A1. Demo query on the SNPdryad website in June 2014. In the demo query, we queried SNPdryad whether the amino acid substitution from M to E at the 18th position of the input protein (ENSP00000000233) is deterious or not.