Pattern Recognition 135 (2023) 109134

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

Sequence patterns and HMM profiles to predict proteome wide zinc finger motifs

Chakkarai Sathyaseelan, L Ponoop Prasad Patro, Thenmalarchelvi Rathinavelan*

Department of Biotechnology, Indian Institute of Technology Hyderabad, Kandi, Telangana 502284, India

ARTICLE INFO

Article history:
Received 2 March 2022
Revised 2 October 2022
Accepted 22 October 2022
Available online 29 October 2022

Keywords:
Zinc finger classification
Zinc finger motif
Zinc finger proteome
Pfam HMM profile
Zinc finger prediction

ABSTRACT

Zinc finger (ZnF) is an important class of nucleic acid and protein recognition domain, wherein, zinc ion is the inorganic co-factor that forms a tetrahedral geometry with the cysteine and/or histidine residues. ZnF domains take up diverse architectures with different ZnF motifs and have a wide range of biological functions. Nonetheless, predicting the ZnF motif(s) from the sequence is quite challenging. To this end, 74 unique ZnF sequence patterns are collected from the literature and classified into 32 different classes. Since the shorter length of ZnF sequence patterns leads to inaccurate predictions, ZnF domain Pfam HMM profiles defined under 6 ZnF Pfam clans (215 HMM profiles) and a few undefined Pfam clans (74 HMM profiles) are used for the prediction. A web server, namely, ZnF-Prot (https://project.iith.ac.in/znprot/) is developed which can predict the presence of 31 ZnF domains in a protein/proteome sequence of any organism. The use of ZnF sequence patterns and Pfam HMM profiles resulted in an accurate prediction of 610 test cases (taken randomly from 249 organisms) considered here. Additionally, the application of ZnF-Prot is demonstrated by considering *Arabidopsis thaliana, Homo sapiens, Saccharomyces cerevisiae, Caenorhabditis elegans* and *Ciona intestinalis* proteomes as test cases, wherein, 87–96% of the predicted ZnF motifs are cross-validated.

1. Introduction

Metal ions play a pivotal role in catalyzing the biological functions and stabilizing the biomacromolecular structures. Among the metal ion co-factors, D-block metal ions play widespread biological functions [1]. Previous studies have shown that 4–10% of the genes in an organism encode for zinc binding proteins, *viz.*, roughly 3000 human proteins, to perform the catalytic activity as well as to stabilize the protein structure [1,2]. Zinc ion binding enzymes are present in all the six major enzyme classes, namely, oxidoreductases, transferases, hydrolases, lyases, isomerases and ligases [2]. Thus, zinc deficiency causes several diseases in humans [3]. Zinc deficiency is indeed concurrent with iron deficiency anaemia (IDA) since zinc ion binding enzymes play a role in iron metabolism [4,5].

Zinc finger (ZnF) proteins are a special class of zinc binding proteins which consist of a small protein domain, wherein, Zn(II) is the inorganic co-factor that forms a tetrahedral geometry with the cysteine and/or histidine residues [6–8]. ZnF proteins play an important role in protein catalytic activity, stability and folding. Thus, any alterations in the amino acids present in the ZnF domain of a protein lead to several diseases [9]. The term zinc finger has been assigned based on the finger-like appearance seen in the zinc ion binding domain of *Xenopus laevis* transcription factor IIIA [10]. This is referred to as the classical C2H2 zinc finger (ZnF) domain and is the most commonly found to date. Over the time, many additional zinc finger domains have been discovered. ZnF domains are compact and have a diverse architecture. They play a wide range of essential biological functions. For instance, zinc finger proteins have a key role in several tissue development and differentiation [9]. Zinc finger proteins are one of the most commonly occurring transcription factors and play an extensive role in gene regulation [9]. Due to their diverse and essential functional roles, zinc finger proteins are potential drug targets to treat viral infections [11], cancer [12], inflammatory conditions [13], parasitic diseases [13] and neurological disorders [13]. Besides, programmed ZnF nucleases have emerged as a prominent gene editing tool [14] and ZnF is also being used in diagnostics [15,16].

In addition to the classical C2H2 zinc finger domains, non-classical zinc finger domains also occur in nature. The ZnF domains differ from each other in terms of their motifs (ZnF motifs) as the cysteine and histidine amino acids combinations and compositions vary among them. An earlier investigation has classified ZnF domains into eight classes [17] based on their 3D conformation. The human genome organization (HUGO) gene nomenclature

* Corresponding author.
E-mail address: tr@bt.iith.ac.in (T. Rathinavelan).

https://doi.org/10.1016/j.patcog.2022.109134
0031-3203/© 2022 Elsevier Ltd. All rights reserved.

committee has classified human ZnF into 30 different domains based on their structure [9].

Sequence pattern matching algorithms enable the identification of conserved DNA or amino acid sequence patterns present in a set of sequences [18]. Pattern matching helps in the identification of biomarkers [19] and, conserved patterns in the non-coding regions [20], protein coding regions [20] and promoter regions [20]. Similarly, it is useful in identifying the amino acid patterns with functional relevance [20], establishing out the evolutionary relationship among the sequences [21], protein family classification [20], *etc.*

Profile Hidden Markov model (HMM) is widely used in computational biology to analyze the sequences [22] and aids in the gene annotation, protein classification, motif detection, *etc.* by overcoming the shortcomings in pairwise alignment, similarity search and multiple sequence alignment methods [22]. It has long been used in aligning distantly related sequences [23] as the multiple sequence alignment (MSA) of distantly related sequences becomes challenging [24]. Profile HMM provides the position-specific probabilities of the occurrence of amino acids/nucleotides in a given set of sequences. The most common application of profile HMM is to identify the distance homologues as well as the known sequence domains present in the new sequences. Here, both sequence patterns and profile HMMs are employed to identify various ZnF domains/motifs present in a query protein/proteome.

The rest of the paper is organized as follows: Section 2 outlines the related work. Problem definition is described in the Section 3. Section 4 defines the proposed work. Sections 5–7 detail the experimental setup, results and applications respectively. Section 8 illustrates the limitation. The conclusions are given in the Section 9.

2. Related work

Considering the wide range of biological roles, identifying the ZnF sequence motifs present in the ZnF domain(s) of a protein/proteome is of great importance. Since it is time-consuming to experimentally determine the ZnF protein domains/motifs [25], computational prediction comes as an alternative. Although there are tools such as Ion Com [26] and Zincbindpredict [27] to predict the zinc binding sites present in a protein (note that zinc binding proteins [28] are different from zinc finger proteins [29]) sequence using pattern search approach, there is no well-established tool to predict the presence of zinc finger motif(s) in a given protein sequence. Indeed, the existing MIB web tool [30] can predict the zinc finger motif region, provided, the template structure is already available in the protein databank. Thus, MIB has a limitation of predicting the ZnF motifs from a given protein/proteome sequence.

3. Problem definition

From the literature survey, it is evident that there is no up to date classification of ZnF domains is available. Further, there is no dedicated tool available for the identification of ZnF domains in a given protein/proteome sequence. However, the biological significance of zinc finger domains necessitates the proper classification of all the existing zinc finger domains and the development of a single platform that can predict both the classical and non-classical zinc finger domains/motifs present in a given protein/proteome sequence. Development of a tool that uses sequence patterns and profile HMMs together may thus be useful in filling up the gap in the prediction of any type of ZnF domains/motifs present in a protein/proteome sequence. In this context, the pattern search and profile HMM approaches are discussed below.

3.1. Pattern search

Pattern recognition approach has successfully been shown to identify the nucleotide or amino acid sequence pattern(s) in biological sequences [31–34]. Let us consider a list of known amino acids patterns (P) = [p_i, p_{i+1}, p_{i+2}, p_{i+3},..., p_{i+n}] whose presence is to be identified in the list of given protein sequences (S) = [s_i, s_{i+1}, s_{i+2}, s_{i+3},..., s_{i+n}]. Note that both P and S have the patterns that consist of 20 amino acids residues [A, C, D, E, F, G, H, I, K, L, M, N, P,Q, R, S, T, V, W and Y]. The pattern search algorithms search for the presence of amino acids patterns (which is relatively shorter than S) in S. For example, a short sequence pattern (P = CMVSKLCFMSHDESFH) is searched against a longer sequence (S = MHDECHMVSCMVSKLCFMSHDESFHVSFHCKLAEGHMC) irrespective of the amino acids position. However, this approach has a shortcoming of providing false positive(s) [35].

3.2. Profile HMM

Sequence similarity search representing a domain or a protein family can be done using the profile hidden Markov model (HMM). This approach has successfully been used in detecting the remote homologs by comparing a sequence to HMM profile [23]. It is a probabilistic model that can capture position-specific information about the conservation of amino acids in each columns of a set of aligned sequences. Thus, profile hidden Markov model (HMM) can be used to identify protein sequence similarities as it can encapsulate the evolutionary changes that have occurred in a set of related sequences. Profile HMM also captures the information about to what extent the gaps and insertions occur in a set of sequences.

An HMM profile is associated with several probabilities as discussed below. A representative profile HMM model is shown in Fig. 1. In the profile HMM model, each column representing the alignment has a node, wherein, each node represents a match (M) or a gap insertion (I) or a deletion (D) state. Thus, one can describe the transition probabilities that simply represents the probability of transition from one of these three states to another (*viz.*, M to M, M to I, M to D, I to I, I to M, I to D, D to D, D to M and D to I) via a connected edge. For instance, in an ungapped model, the path through the model is strictly linear when moving from i^{th} match state node to $i + 1^{th}$ match state node. Further, there is emission probabilities which are associated with each match and insertion states and, is derived based on the probability of a given residue existing at a particular position in the given alignment.

Thus, profile HMM method can capture the match states (**M**), wherein, the probability distribution is the frequency of the amino acids in each position, the insertion states (**I**) which are used to model highly variable regions in the alignment and, the deletion states (**D**), wherein, they do not match with any residues and acts as silent states. Finally, this probabilistic model provides the estimation of not only the observed frequencies of the amino acids in each position (emission probability, (e)), but also, the transitions between the amino acids derived from the observed occupancy of each position (transition probability, (t)) in a trained multiple sequence alignment (Fig. 1B). Finally, an HMM diagram is derived for a given set of sequences which has M, I and D as the hidden states (Fig. 1C).

Subsequently, a dynamic programing algorithm is employed to align a new sequence against the generated HMM profile. Although many dynamic programing algorithms are being used to align the sequence to a HMM profile [36], Viterbi algorithm is the frequently used one [36] which is employed in the current study. As per Viterbi algorithm, the optimal alignment of a query sequence to the profile HMM can be obtained through the following equations 1-3 [37,38]. Viterbi dynamic programming algorithm creates the alignment matrix by considering the sequence of observations (for

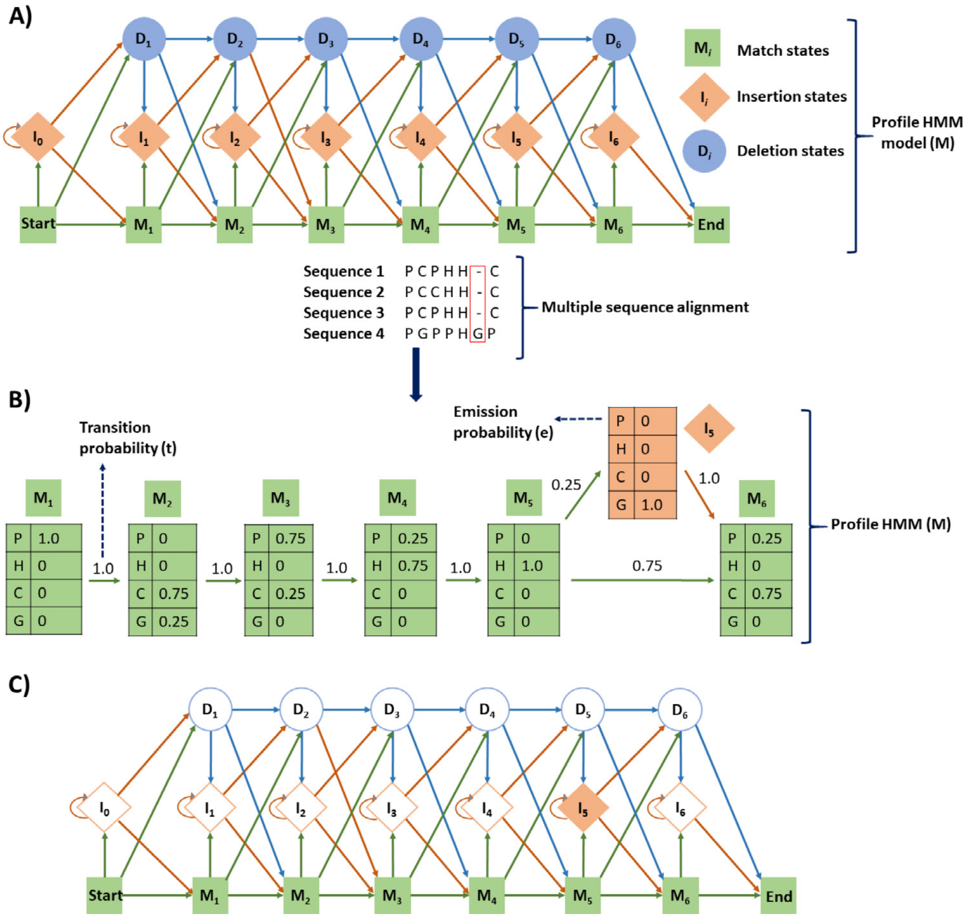


Fig. 1. Schematic illustration of profile hidden Markov model (HMM) using a toy example. (A) A generalized representation of HMM model, wherein, the match (M_i), insertion (I_i) and deletion (D_i) states are colored in green (box), orange (diamond) and blue (circle) respectively. Note that '*i*' represents the column number in the aligned sequences. (B and C) A toy example showing the generation of transition (detonated as t and indicated adjacent to each solid arrow) and emission (denoted as e and indicated in the boxes) probabilities (B) and, the corresponding HMM diagram (C). Since the transition path has only match (square shape) and insertion (diamond shape) states (B), the corresponding boxes are alone highlighted with appropriate color in (C).

example, x_1, x_2, x_3, ….,x_i) on the horizontal axis and the states (for instance, s_1, s_2, s_3, …., s_j) on the vertical axis. Each cell (i,j) in the matrix accumulates the Viterbi probability score using the following equations.

$$V_j^M(i) = log\ e_{M_j\ (x_i)} + max\{ \begin{array}{l} V_{j-1}^M(i-1) + \log a_{M_{j-1}M_j} \\ V_{j-1}^I(i-1) + \log a_{I_{j-1}M_j} \\ V_{j-1}^D(i-1) + \log a_{D_{j-1}M_j} \end{array} \qquad (1)$$

$$V_j^I(i) = log\ e_{I_j\ (x_i)} + max\{ \begin{array}{l} V_j^M(i-1) + \log a_{M_jI_j} \\ V_j^I(i-1) + \log a_{I_jI_j} \\ V_j^D(i-1) + \log a_{D_jI_j} \end{array} \qquad (2)$$

$$V_j^D(i) = max\{ \begin{array}{l} V_{j-1}^M(i) + \log a_{M_{j-1}D_j} \\ V_{j-1}^I(i) + \log a_{I_{j-1}D_j} \\ V_{j-1}^D(i) + \log a_{D_{j-1}D_j} \end{array} \qquad (3)$$

wherein, $V_j^M(i)$, $V_j^I(i)$ and $V_j^D(i)$ are the log-odds values of the highest scoring alignment (*viz.*, the most probable path) correspond to the subsequence of an observation. The $e_{M_j(x_i)}$ and $e_{I_j(x_i)}$ are the emission probabilities of the observation (x_i) at the current state M_j and I_j respectively. The Viterbi path probability from the previous step are: $V_{j-1}^M(i-1)$, $V_{j-1}^I(i-1)$ and $V_{j-1}^D(i-1)$ for the match state, $V_j^M(i-1)$, $V_j^I(i-1)$ and $V_j^D(i-1)$ for the insert state and $V_{j-1}^M(i)$, $V_{j-1}^I(i)$ and $V_{j-1}^D(i)$ for the deletion state. $a_{M_{j-1}M_j}$, $a_{I_{j-1}M_j}$ and $a_{D_{j-1}M_j}$ are the transition probabilities to a match (M_j), from a match (M_{j-1}) or an insertion (I_{j-1}) or a deletion (D_{j-1}) state. Similarly, $a_{M_jI_j}$, $a_{I_jI_j}$ and $a_{D_jI_j}$ are the transition probabilities to an insertion state (I_j) from a match (M_j) or an insertion (I_j) or a deletion (D_j) state and $a_{M_{j-1}D_j}$, $a_{I_{j-1}D_j}$ and $a_{D_{j-1}D_j}$ are the transition probabilities to a deletion state (D_j) from a match (M_{j-1}) or an insertion (I_{j-1}) or a deletion (D_{j-1}) state. Finally, a backtracking procedure is followed using a maximum Viterbi probability score to visit each column in the matrix until the first column is reached.

4. Proposed work

Based on the literature review and problem statement, following workflow has been formulated:

(1) Classification of the zinc finger domains and the creation of zinc finger motif sequence patterns repository. Prior to the implementation of the proposed problem, a well-defined classification of ZnF domains and the concurrent sequence motif information are required. Thus, previously published ZnF domain classifications [9,17,39] have been carefully examined to create a repository of zinc finger sequence motifs corresponding to different ZnF domains.

(2) Collection of the HMM profiles corresponding to different ZnF domains and the creation of their local repository.

(3) Implementation of ZnF domain prediction web tool that employs sequence pattern search and profile HMM approaches.

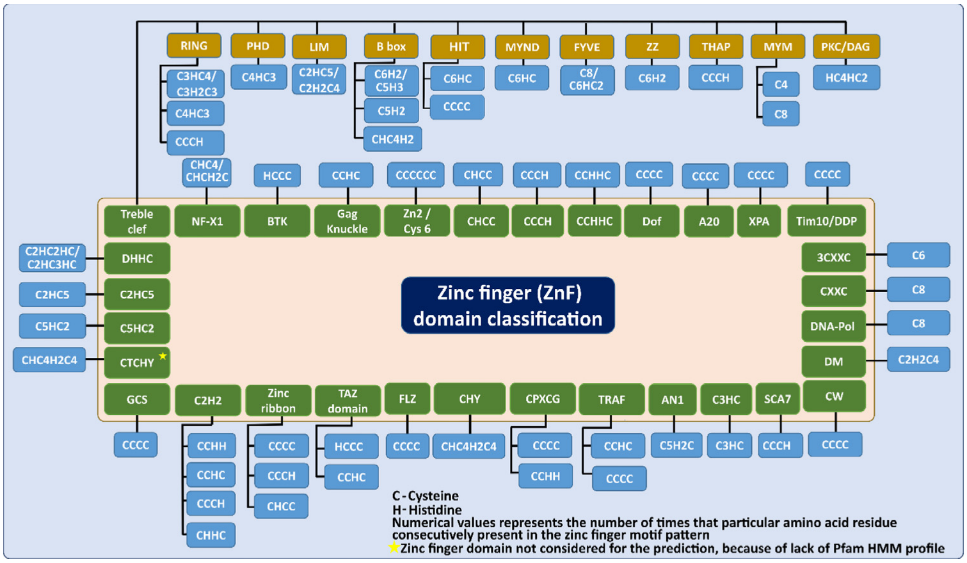


Fig. 2. Classification of 32 zinc finger domains. The sequence motif patterns under each zinc finger domain (green colored box) are shown in blue colored box. Note that treble clef zinc finger domains were sub-classified into 11 different types (orange colored box), namely, RING, PHD, LIM, B-box, HIT, MYND, ZZ, THAP, MYM, FYVE and PKC/DAG domains. C and H in the sequence pattern represent cysteine and histidine amino acid residues that coordinate with the zinc ion and the numerical values given in the sequence pattern represent the number of times a particular amino acid occurs consecutively in the zinc finger motif pattern. The yellow color star represents that CTCHY zinc finger domain which is not considered for the prediction as it doesn't have any Pfam HMM profile (See Experimental setup section).

(4) Validation of the efficacy of the proposed approach.
(5) Demonstration of the application of the proposed approach.

5. Experimental setup

5.1. Classification of zinc finger domains

By considering the structural diversity of the ZnF domains, a systematic classification of zinc finger domains was initially done through a thorough literature survey. Since 2003, four investigations have discussed in detail about the classification of zinc finger domains. The first ever ZnF domain classification was done in 2003 based on their 3-dimensional structure [17]. In this study, ZnF domains were classified into eight types, namely, C2H2 zinc finger, gag knuckle (zinc knuckle), treble clef zinc finger, zinc ribbon, Zn2/Cys6 zinc finger, TAZ2 domain, zinc binding loop and metallothionein. In 2017, ZnF domains related to human health and diseases were classified into 30 different types in accordance with the approval of the human genome organization (HUGO) gene nomenclature committee [9]. ZnF domains were classified into 8 major classes in an another study [40]: $\beta\beta\alpha$ zinc finger (C2H2 zinc finger), gag knuckle (zinc knuckle) zinc finger, zinc ribbon, CCCH zinc finger, CCHHC zinc finger, treble clef zinc finger (subclasses: RING domain, LIM domain, FYVE domain, PHD domain, B-box domain and HIT domain), Zn2/Cys6 zinc finger and TAZ domain zinc finger. Additionally, the treble clef zinc finger was classified into to 6 sub-classes in the same study (see above). Notably, the zinc binding loop and metallothionein domain which were originally classified as ZnF domain in 2003 [17] were not considered as the ZnF domains in the recent classifications [9,40]. In another study, ZnF domains were classified into 51 types based on the protein functional domain classification of InterPro and SMART databases [39]. A detailed analysis was carried out in the current study to derive a unified and comprehensive classification of ZnF domains by considering the abovementioned classifications (Supplementary Table S1). Here, the zinc finger domains were classified into 32 major classes (Fig. 2), namely, C2H2, CCCH, Zinc ribbon, zinc knuckle (gag knuckle), TAZ domain, Zn2/Cys6, treble clef (RING, PHD, LIM, B-box, HIT, MYND, ZZ, THAP, MYM, FYVE and PKC/DAG domains), CCHHC, CW, 3CXXC, DHHC, CXXC, CPXCG, CHCC, CHY, CTCHY, TRAF, FLZ, A20, AN1, C3HC, C2HC5, C5HC2, Dof, DNA-pol, SCA7, GCS, BTK, Tim10/DDP, DM, NF-X1 and XPA. Further, a total of 56 different zinc finger motif types were grouped under the above-mentioned 32 major classes based on the zinc ion coordinating cysteine and/or histidine sequence motifs (Fig. 2). For instance, CCHH (i.e. 2 cysteines followed by 2 histidines), CCHC, CCCH and CHHC motif types were grouped under C2H2 zinc finger. Similarly, zinc ribbon domain classification encompasses CCCC, CCCH and CHCC motif types. Notably, treble clef was sub-classified into 11 types and 17 sequence motifs were grouped under them.

5.2. Creation of a repository of zinc finger motif sequence patterns

Through an extensive literature survey, 74 different zinc finger motif sequence patterns that represent 56 motif types of 32 ZnF domains (Table 1) were collected and stored in a local database. The patterns were stored in such a way that the amino acids connecting the cysteine (C) and/or histidine (H) residues and coordinate with zinc ion were represented by “X”. For instance, in C-X_2-C-X_{11-14}-C-X_2-C zinc finger motif pattern, which belongs to the zinc ribbon domain identified in L37 and L37a Ribosomal protein in *Haloarcula marismortui* [41], the amino acids connecting the 4 cysteine residues were indicated by “X”. Further, any 2 amino acids (X_2) can be present in between the 1st and 2nd cysteines and 3rd and 4th cysteines. Similarly, in between 2nd and 3rd cysteines any amino acids combination with a length of 11 to 14 (X_{11-14}) can be present. In the local repository, 10, 1, 6, 1, 2, 1, 21, 1, 1, 1, 3, 1 and 2 motif sequence patterns were listed under C2H2, CCCH, Zinc ribbon, zinc knuckle (gag knuckle), TAZ domain, Zn2/Cys6, treble clef, CCHHC, CW, 3CXXC, DHHC, CXXC and CPXCG domains respectively (Table 1). Similarly, CHCC, CHY, TRAF, FLZ, A20, AN1, C3HC, C2HC5, C5HC2, Dof, DNA-pol, SCA7, GCS, BTK, Tim10/DDP, DM, NF-X1 and XPA zinc finger domains contain one zinc finger motif sequence pattern each (Table 1).

5.3. Collection of Pfam hidden Markov model (HMM) profiles

The HMM profile corresponding to 31 zinc finger domains were collected from Pfam (protein families) database [42]

Table 1
List of zinc finger motif sequence patterns that represent 56 motif types of 32 ZnF domains. Column 2 lists 32 zinc finger domain classification, column 3 lists the sub-classifications (only in treble clef), column 4 lists 56 motif types and column 5 lists the 74 different zinc finger motif sequence patterns.

S. No	Classification	Sub-classification	Motif types	Sequence motif patterns
1	C2H2		CCHH	C-X$_{2-4}$-C-X$_{4-5}$-H-X$_{2-4}$-H
				C-X$_{2-5}$-C-X$_{12}$-H-X$_{3-5}$-H
				C-X$_{6-12}$-C-X$_{17-22}$-H-X-H
				C-X$_{2}$-C-X$_{11-15}$-H-X$_{3}$-[CH]
				C-X$_{4}$-C-X$_{12}$-H-X$_{16}$-H
				C-X$_{4-5}$-C-X$_{21-24}$-H-X-H
			CCHC	C-X$_{2}$-C-X$_{4}$-H-X$_{4}$-C
				C-X$_{5}$-C-X$_{12}$-H-X$_{4}$-C
				C-X$_{2}$-C-X$_{11}$-H-X$_{3}$-C
				C-X$_{2}$-C-X$_{16}$-H-X$_{6}$-C
				C-X$_{3–8}$-C-X$_{23-31}$-H-X$_{1-2}$-C
			CCCH	C-X$_{1}$-C-X$_{6-25}$-C-X$_{1}$-H
				C-X$_{5}$-C-X$_{5}$-C-X$_{3}$-H
			CHHC	C-X$_{5-7}$-H-X$_{7-9}$-H-X$_{2-4}$-C
2	CCCH		CCCH	C-X$_{3-17}$-C-X$_{3-10}$-C-X$_{1-5}$-H
3	Zinc ribbon		CCCC	C-X$_{2}$-C-X$_{21-24}$-C-X$_{2}$-C
				C-X$_{2-4}$-C-X$_{10-20}$-C-X$_{2}$-C
				C-X$_{2}$-C-X$_{12}$-C-X$_{9}$-C
				C-X$_{2}$-[CH]-X$_{15-17}$-C-X$_{2}$-C
			CCCH	C-X$_{2}$-C-X$_{12}$-C-X$_{4}$-H
			CHCC	C-X$_{5}$-H-X$_{53}$-C-X$_{2}$-C
4	Zinc knuckle		CCHC	C-X$_{2-4}$-C-X$_{4}$-H-X$_{4}$-C
5	TAZ domain		CCHC	C-X$_{4}$-C-X$_{8-9}$-H-X$_{3}$-C
			HCCC	H-X$_{3}$-C-X$_{4-12}$-C-X$_{2-4}$-C
6	Zn2/Cys6		C6	C-X$_{2}$-C-X$_{6}$-C-X$_{5-12}$-C-X$_{2}$-C-X$_{6-9}$-C
7	Treble clef	RING	C3HC4/C3H2C3	C-X$_{2}$-C-X$_{9-39}$-C-X$_{1-3}$-H-X$_{2-4}$-[CH]-X$_{2}$-C-X$_{4-74}$-C-X$_{2}$-C
			C4HC3	C-X$_{2}$-C-X$_{9-21}$-C-X$_{2-4}$-C-X$_{2}$-H-X$_{2}$-C-X$_{7-74}$-C-X$_{2}$-C
			CCCH	C-X$_{2}$-C-X$_{16-18}$-C-X$_{6-12}$-H
		PHD	C4HC3	C-X$_{1-2}$-C-X$_{7-21}$-C-X$_{2-4}$-C-X$_{4-5}$-H-X$_{2}$-C-X$_{12-46}$-C-X$_{2}$-C
		LIM	C2HC5/C2H2C4	C-X$_{2}$-C-X$_{16-23}$-[CH]-X$_{2}$-[CH]-X$_{2}$-C-X$_{2}$-C-X$_{15-30}$-C-X$_{1-3}$-[CHD]
				C-X$_{2}$-C-X$_{17-19}$-H-X$_{2}$-C-X$_{2}$-C-X$_{2}$-C-X$_{15-19}$-C
				C-X$_{2}$-C-X$_{17-19}$-H-X$_{2}$-C-X$_{2}$-C-X$_{2}$-C-X$_{7-21}$-C-X$_{2}$-[CHDE]
		B-box	C6H2/C5H3	C-X$_{2}$-C-X$_{6-17}$-C-X$_{2}$-C-X$_{4-8}$-C-X$_{2-3}$-[CH]-X$_{3-4}$-H-X$_{5-10}$-H
				C-X$_{2}$-C-X$_{7-11}$-C-X$_{2}$-[CD]-X$_{4-5}$-C-X$_{2}$-C-X$_{4-5}$-H-X$_{2-8}$-H
				C-X$_{2}$-C-X$_{8}$-C-X$_{7}$-C-X$_{2}$-C-X$_{4}$-H-X$_{6-8}$-H
				C-X$_{2}$-C-X$_{7-12}$-C-X$_{2}$-C-X$_{4}$-C-X$_{2}$-[CH]-X$_{3-4}$-H-X$_{4-9}$-H
			CHC4H2	C-X$_{2-4}$-H-X$_{7-10}$-C-X$_{1-4}$-[CDE]-X$_{4-7}$-C-X$_{2}$-C-X$_{3-6}$-H-X$_{2-5}$-H
		HIT	C6HC	C-X$_{2-4}$-C-X$_{7-11}$-C-X$_{2}$-C-X$_{4}$-C-X$_{3}$-C-X$_{3}$-H-X$_{2-5}$-C
			CCCC	C-X$_{2-4}$-C-X$_{15-19}$-C-X$_{3}$-C
		MYND	C6HC	C-X$_{2-4}$-C-X$_{7-12}$-C-X$_{0-2}$-C-X$_{5}$-C-X$_{3}$-C-X$_{7-9}$-H-X$_{3}$-C
		ZZ	C6H2	C-X$_{2}$-C-X$_{5-11}$-C-X$_{2}$-C-$_{5-8}$-C-X$_{2}$-C-$_{2-14}$-H-X$_{1-7}$-H
		THAP	CCCH	C-X$_{2-4}$-C-X$_{35-53}$-C-X$_{2}$-H
		MYM	C8	C-X$_{2}$-C-X$_{19-22}$-C-X$_{3}$-C-X$_{13-19}$-C-X$_{2}$-C-X$_{19-25}$-C-X$_{2}$-C
			CCCC	C-X$_{2}$-C-X$_{19-24}$-[FY]-C-X$_{3}$-C
		FYVE	C8/C6HC2	C-X$_{1-4}$-C-X$_{11-23}$-C-X$_{2}$-C-X$_{4}$-C-X$_{2-6}$-[CH]-X$_{13-40}$-C-X$_{2}$-C
		PKC/DAG	HC4HC2	H-X$_{11-12}$-C-X$_{2}$-C-X$_{12-14}$-C-X$_{2}$-C-X$_{4}$-H-X$_{2}$-C-X$_{6-7}$-C
8	CCHHC		CCHHC	C-X$_{4}$-C-X$_{4}$-H-X$_{7}$-H-X$_{5}$-C
9	CW		CCCC	C-X$_{2-4}$-C-X$_{18-21}$-C-X$_{10-15}$-C
10	3CXXC		3 CXXC	3 C-X$_{2}$-C
11	DHHC		CCHC	C-X$_{2}$-C-X$_{9}$-H-X$_{6}$-C
			C2HC2HC	C-X$_{2}$-C-X$_{9}$-H-C-X$_{2}$-C-X$_{4}$-D-H-H-C-X$_{5}$-C
			C2HC3HC	C-X$_{2}$-C-X$_{9}$-H-C-X$_{2}$-C-X$_{2}$-C-X$_{4}$-D-H-H-C-X$_{5}$-C
12	CXXC		C8	C-X$_{2}$-C-X$_{2}$-C-X$_{4-5}$-C-X$_{2}$-C-X$_{2}$-C-X$_{9-15}$-C-X$_{4}$-C
13	CPXCG		CCCC	C-X$_{2}$-C-X$_{17-20}$-C-X$_{2}$-C
			CCHH	C-X$_{2}$-C-X$_{25}$-H-X$_{2}$-H
14	CHCC		CHCC	C-X$_{8}$-H-X$_{14}$-C-X$_{2}$-C
15	CHY		CHC4H2C4	C-X-H-X$_{10}$-CC-X$_{5}$-C-X$_{2}$-CH-X$_{5}$-H-X$_{11}$-C-X$_{2}$-C-X$_{9}$-C-X$_{2}$-C
16	TRAF		CCCC/CCHC	C-X$_{2-5}$-C-X$_{11}$-[HC]-X$_{3-4}$-C
17	FLZ		CCCC	C-X$_{2}$-C-X$_{17-20}$-C-X$_{3}$-C
18	A20		CCCC	C-X$_{2-4}$-C-X$_{11}$-C-X$_{2}$-C
19	AN1		C5H2C	C-X$_{2}$-C-X$_{9-12}$-C-X$_{1-2}$-C-X$_{4}$-C-X$_{2}$-H-X$_{5}$-H-X-C
20	C3HC		C3HC	C-X$_{7}$-C-X$_{2}$-C-X$_{41}$-H-X$_{3}$-C
21	C2HC5		C2HC5	C-X$_{1}$-C-X$_{4}$-H-X$_{6-8}$-C-X$_{2}$-C-X$_{4}$-C-X$_{3-8}$-C-X$_{2}$-C
22	C5HC2		C3H	C-X$_{2}$-C-X$_{18-19}$-C-X$_{2}$-H
23	Dof		CCCC	C-X$_{2}$-C-X$_{21}$-C-X$_{2}$-C
24	DNA-pol		C8	C-X$_{2-4}$-C-X$_{9-34}$-C-X$_{2-4}$-C-X$_{30-35}$-C-X$_{2-4}$-C-X$_{11-20}$-C-X$_{1-5}$-C
25	SCA7		CCCH	C-X$_{9-10}$-C-X$_{5}$-C-X$_{2}$-H
26	GCS		CCCC	C-X$_{2}$-C-X$_{16-17}$-C-X$_{2}$-C
27	BTK		HCCC	H-X$_{10}$-C-C-X$_{9}$-C
28	Tim10/DDP		CCCC	C-X$_{3}$-C-X$_{15-21}$-C-X$_{3}$-C
29	DM		C2H2C4	C-X$_{2}$-C-X$_{2}$-H-X$_{8}$-H-X$_{3-4}$-C-X$_{4}$-C-X$_{1}$-C-X$_{2-3}$-C
30	NF-X1		CHC4/CHCH2C	C-X$_{1-6}$-H-X-C-X$_{3}$-[HC]-X$_{3-4}$-[HC]-X$_{1-10}$-C
31	XPA		CCCC	C-X$_{2}$-C-X$_{17}$-C-X$_{2}$-C
32	CTCHY		C2HC3HC3HC	C-X$_{2}$-C-X$_{10}$-HC-X$_{2}$-C-X$_{2}$-C-X$_{9}$-HC-X$_{2}$-C-X$_{2}$-C-X$_{8}$-H-X$_{1}$-C

(https://pfam.xfam.org/). Pfam has a massive collection of protein families which are represented by multiple sequence alignment (MSA) and HMM profiles along with the functional domain information. A total of 289 Pfam HMM profiles were collected, among which, 259 belong~~s~~ to above classified 31 zinc finger domains and 30 belong~~s~~ to other zinc finger domains (not a part of above classification and don't have any sequence motif patterns) (Supplementary Table S2). It is noteworthy that the Pfam HMM profiles corresponding to the zinc finger domains were searched through the keyword "zinc finger" in the keyword search option of the Pfam database. Subsequently, the grouping of the collected Pfam HMM profiles according to 31 different zinc finger classes was done as described below. Initially, the Pfam HMM profiles belonging to a particular ZnF clan in the Pfam database were directly assigned under a zinc finger domain. For example~~s~~, the 52 different Pfam HMM profiles that correspond to the Pfam clan C2H2-zf (CL0361) were categorized as C2H2 zinc finger HMM profile. However, categorizing all the HMM profiles corresponding to the 31 different ZnF domains was not straightforward since many HMM profiles were not defined under a particular ZnF clan. Such HMM profiles were assigned under a zinc finger domain based on the "description" or "name" field given in the HMM profiles or through the literature survey. For instance, zf-B_box (PF00643) is not categorized under any clan in the Pfam database. However, it was categorized under treble clef zinc finger domain in the local repository based on the HMM profile "description" field. The collected Pfam HMM profiles were then named according to their zinc finger domain classification (e.g. C2H2 zinc finger, treble clef zinc finger, *etc.*). Finally, 289 Pfam (including the 30 zinc finger domains that don't have sequence motif patterns) HMM profiles were compiled into a single file to facilitate the sequence to Pfam HMM profile search. It is noteworthy that due to the unavailability of Pfam HMM profile(s) corresponding to CTCHY zinc finger domain, this ZnF domain was excluded for the prediction.

5.4. Web server implementation

To facilitate the automated prediction of ZnF domains/motifs present in a protein/proteome using the sequence motif patterns and Pfam HMM profiles, the ZnF-Prot web server was developed. ZnF-Prot (**Z**i **n**c **F**inger domains in **Pr**oteins/Pr **ot**eomes, https://project.iith.ac.in/znprot/) web tool was implemented in an Ubuntu Linux system (18.04 LTS) with the help of Apache (https://httpd.apache.org) and D3.js (https://d3js.org/). The client-side user interface was implemented using HTML and PHP. The in-house scripts used in the prediction were written in bash, python and R.

5.5. Methodology employed in the prediction of zinc finger motifs

5.5.1. Zinc finger motif pattern search

The methodology was devised in such a way that when a query protein sequence is submitted as an input, ZnF-Prot looks for the presence of one or more of the 74 zinc finger sequence motif patterns that are stored in the local database. Due to the shorter length of the ZnF sequence patterns, the output from the pattern search approach may contain false positive(s). To overcome this issue, all the zinc finger motif patterns reported from the pattern search are confirmed using the Pfam HMM profile search (See below) (Fig. 3).

5.5.2. Sequence to Pfam HMM profile search

Simultaneously, ZnF-Prot compares the query sequence with 289 Pfam HMM profiles to avoid the false positive(s) and increase the ZnF domain/motif prediction accuracy. The HMMER [43] software is employed in ZnF-Prot for the query sequence to Pfam HMM profile search. The *hmmpress* command is used in the conversion of the text format of the HMM profiles to a binary format to facilitate the use of *hmmscan* and *hmmsearch* commands. The *hmmscan* command has been used to search the query protein sequence against binary formatted HMM profile database to find the significant ZnF functional domain. The –cut_ga ("ga" stands for the gathering threshold which is generally used to define the significance of the query sequence match against a particular Pfam domain) option is used along with the *hmmscan* command to report only the significant ZnF Pfam domain. Finally, the ZnF domain Pfam families that have the gathering threshold greater than or equal to the bit score are alone considered as the significant ones and the others are considered as an insignificant (*viz.*, the gathering threshold of the ZnF domain Pfam family is lesser than or equal to the bit score of the ZnF domain Pfam family) and are ignored. If no Pfam families are reported through the above criteria, an E-value cut-off of 1e-04 is used to shortlist the Pfam families. Note that the E-value cut-off of 1e-04 is chosen based on the systematic analysis of 610 test cases, wherein, the E-value less than 1e-04 is observed for the successful predictions (Supplementary Table S3). To further eliminate the false negative predictions by retaining the accuracy of prediction, in the next round, a relaxation of 1e-01 is given for the E-value cut-off to search the Pfam HMM profile. Note that this relaxation is incorporated since one of the 610 test cases exhibited the E-value cut-off of 1e-03. Finally, the Pfam families with the E-value >1e-01 are excluded for the next round of ZnF domain prediction (Fig. 3).

5.5.3. ZnF motif prediction

During the next round of ZnF domain prediction, the results obtained from the zinc finger motif pattern search approach and sequence to Pfam HMM profile search approach are compared with each other to remove the false positives from both approaches. If both approaches report the presence of one or more common zinc finger motif types in the query sequence, then the zinc finger domain/motif prediction is considered to be accurate. Notably, two different types of pattern search and Pfam HMM profile search comparisons are done. The first one is a straight forward method, wherein, the residue numbers of the pattern search and Pfam HMM profile search matches (normal (NRML) in Fig. 3). There may also be a situation with a shift in the residue numbers of the ZnF motif predicted using the pattern search which falls partially within the ZnF domain predicted using Pfam HMM profile search. Since this still represents the presence of the ZnF motif, this is also considered as an accurate prediction, but with a shift (shift (SHFT) in Fig. 3).

Further, there may be insertion(s) and/or deletion(s) and/or mutation(s) in the finger motif amino acids across different organisms due to the evolution, thus, the pattern search may not predict it as a ZnF motif. However, the Pfam HMM profile search can still report the presence of the ZnF motif under such scenario. Thus, to reduce the false-negative predictions under such circumstances, ZnF-Prot proceeds with the prediction through Pfam HMM profile-based prediction. Here, the pairwise sequence alignment of HMM profile to query sequence is carried out to find out the presence of conserved zinc coordinating residues (cysteine(s) or histidine(s)) in the HMM profile as well as in the query sequence. If both the HMM profile and query sequence have the marker residue(s) (cysteine(s) or histidine(s)), then, it is considered that the query sequence has the potential ZnF motif (no motif (NMTF)). Since there is no ZnF pattern seen in the query, ZnF simply reports the presence of a ZnF domain along with the motif without mentioning the position of the motif. However, if there are no conserved marker amino acids found in the query, then, ZnF-Prot reports that the query has the mutated ZnF motif (not considered as ZnF motif due to the presence of mutation (XMUT)). In contrast, if the pairwise

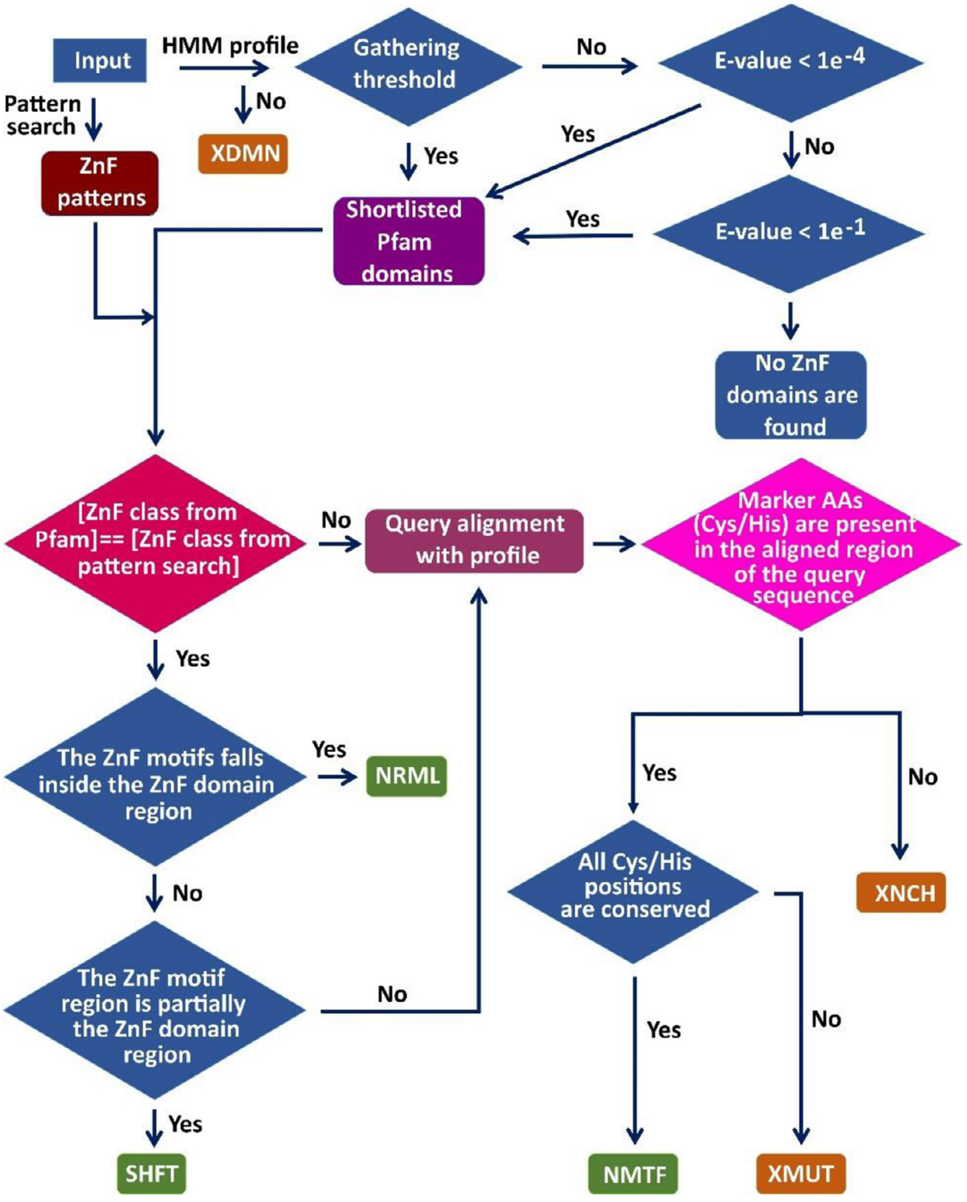

Input
HMM profile
Gathering threshold
No
E-value < 1e-4
Pattern search
No
XDMN
Yes
Yes
No
ZnF patterns
Shortlisted Pfam domains
Yes
E-value < 1e-1
No ZnF domains are found
[ZnF class from Pfam]== [ZnF class from pattern search]
No
Query alignment with profile
Marker AAs (Cys/His) are present in the aligned region of the query sequence
Yes
The ZnF motifs falls inside the ZnF domain region
Yes
NRML
Yes
No
No
All Cys/His positions are conserved
XNCH
The ZnF motif region is partially the ZnF domain region
No
Yes
No
Yes
SHFT
NMTF
XMUT

Fig. 3. Flowchart illustrating the methodology implemented in the zinc finger motif prediction. Either the protein/proteome sequence or PDB ID can be given as an input to the ZnF-Prot server. The methodology consists of 5 steps: (i) search for zinc finger motif sequence pattern in the query sequence using the ZnF motif sequence patterns stored in the local database (brown colored box), (ii) query sequence vs Pfam HMM profile search to find significant Pfam domains (violet colored box, NRML, SHFT), (iii) the choice of E-value threshold for Pfam HMM profile search during the second and third prediction rounds, (iv) comparison of the results obtained from the zinc finger motif sequence pattern and Pfam HMM profile search approaches to predict the final zinc finger motif (pink color box), (v) pairwise comparison between the HMM profile and query (NMFT, XMUT and XNCH). The ZnF motif(s) containing domains predicted with the appropriate pattern(s) is indicated as NRML and SHFT and without the pattern is indicated as NMTF. See Fig. 4 for detailed explanation regarding NRML, SHFT, NMFT, XMUT and XNCH. Note that XDMN represents the situation when there is no HMM profile found for the query sequence.

aligned HMM profile region doesn't have the conserved marker amino acids (cysteine(s) or histidine(s)), then ZnF-Prot directly reports the absence of any ZnF motif in the query (no pattern and no conserved C and H (XNCH)). Such a situation arises when a sequence pattern in the query randomly aligns with the HMM profile. Further, if there is no matching HMM profile for the given query, then, ZnF-Prot reports the absence of a ZnF domain (XDMN). The above-mentioned protocol is described in Figs. 3 and 4.

At the end, ZnF-Prot displays all the predicted zinc finger domains (*viz.*, 31 different zinc finger domains, Fig. 2) present in the query protein/proteome sequence. A detailed summary of the pattern search and Pfam HMM profile search approaches are also provided. In the case of proteome wide ZnF prediction, the overall statistics of different ZnF domains present in the proteome are also displayed. ZnF-Prot also provides the user with an option to get an email notification soon after the job is completed. This option is useful in the case of proteome wide ZnF domain prediction which requires more computational time. At the time of job submission, ZnF-Prot provides a job ID to the user which can be used to check the status of the job. The outputs will be stored in the server for a week from the date of completion of the job.

6. Results

As mentioned in the Section 5, ZnF-Prot predicts the zinc finger domain(s) present in the query protein/proteome sequence in an automated fashion.

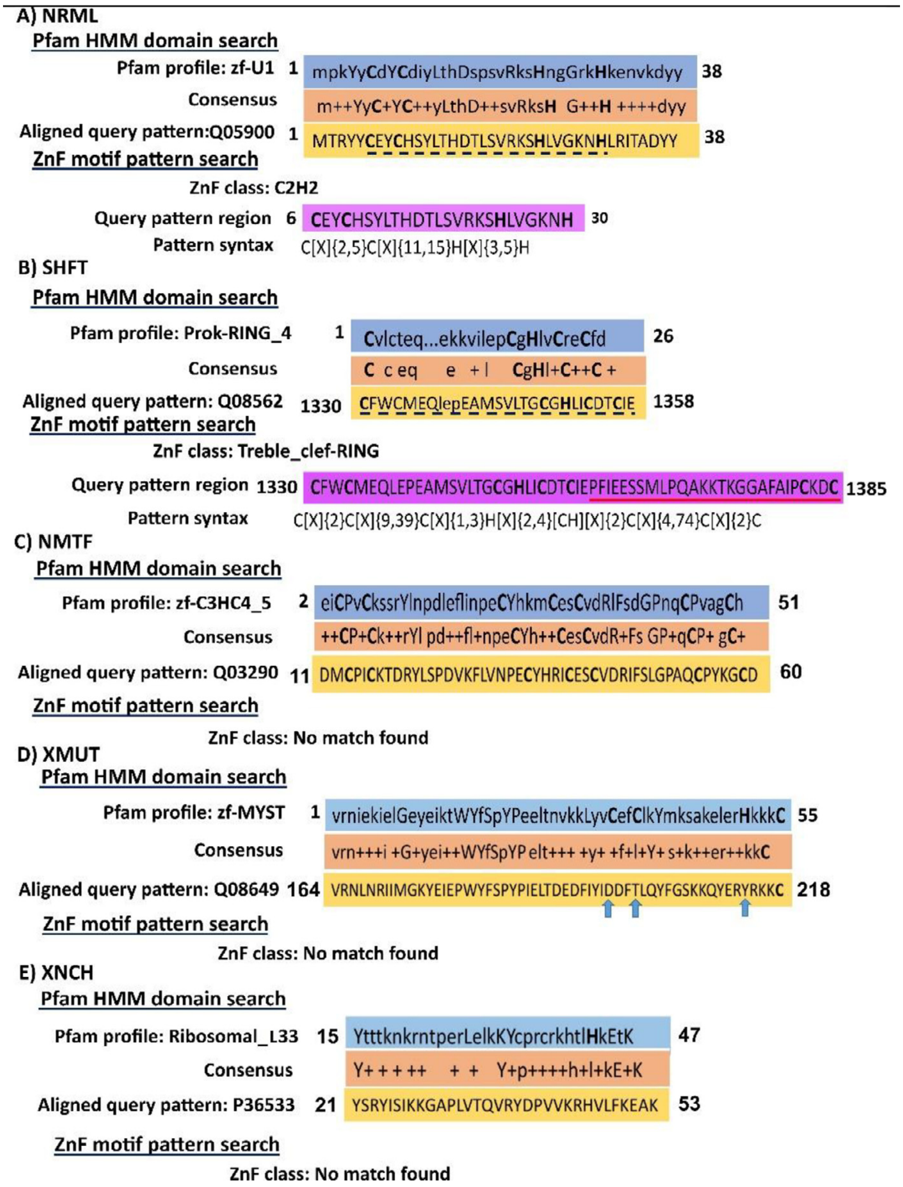

A) NRML
Pfam HMM domain search
Pfam profile: zf-U1 1 mpkYyCdYCdiyLthDspsvRksHngGrkHkenvkdyy 38
Consensus m++YyC+YC++yLthD++svRksH G++H ++++dyy
Aligned query pattern:Q05900 1 MTRYYCEYCHSYLTHDTLSVRKSHLVGKNHLRITADYY 38
ZnF motif pattern search
ZnF class: C2H2
Query pattern region 6 CEYCHSYLTHDTLSVRKSHLVGKNH 30
Pattern syntax C[X]{2,5}C[X]{11,15}H[X]{3,5}H
B) SHFT
Pfam HMM domain search
Pfam profile: Prok-RING_4 1 Cvlcteq...ekkvilepCgHlvCreCfd 26
Consensus C c eq e + l CgHl+C++C +
Aligned query pattern: Q08562 1330 CFWCMEQlepEAMSVLTGCGHLICDTCIE 1358
ZnF motif pattern search
ZnF class: Treble_clef-RING
Query pattern region 1330 CFWCMEQLEPEAMSVLTGCGHLICDTCIEPFIEESSMLPQAKKTKGGAFAIPCKDC 1385
Pattern syntax C[X]{2}C[X]{9,39}C[X]{1,3}H[X]{2,4}[CH][X]{2}C[X]{4,74}C[X]{2}C
C) NMTF
Pfam HMM domain search
Pfam profile: zf-C3HC4_5 2 eiCPvCkssrYlnpdleflinpeCYhkmCesCvdRlFsdGPnqCPvagCh 51
Consensus ++CP+Ck++rYl pd++fl+npeCYh++CesCvdR+Fs GP+qCP+ gC+
Aligned query pattern: Q03290 11 DMCPICKTDRYLSPDVKFLVNPECYHRICESCVDRIFSLGPAQCPYKGCD 60
ZnF motif pattern search
ZnF class: No match found
D) XMUT
Pfam HMM domain search
Pfam profile: zf-MYST 1 vrniekielGeyeiktWYfSpYPeeltnvkkLyvCefClkYmksakelerHkkkC 55
Consensus vrn+++i +G+yei++WYfSpYP elt+++ +y+ +f+l+Y+ s+k++er++kkC
Aligned query pattern: Q08649 164 VRNLNRIIMGKYEIEPWYFSPYPIELTDEDFIYIDDFTLQYFGSKKQYERYRKKC 218
ZnF motif pattern search
ZnF class: No match found
E) XNCH
Pfam HMM domain search
Pfam profile: Ribosomal_L33 15 YtttknkrntperLelkKYcprcrkhtlHkEtK 47
Consensus Y+ + + ++ + + Y+p++++h+l+kE+K
Aligned query pattern: P36533 21 YSRYISIKKGAPLVTQVRYDPVVKRHVLFKEAK 53
ZnF motif pattern search
ZnF class: No match found

Fig. 4. Examples representing the Pfam HMM profile alignment and the query sequence. (A) NRML showing the ZnF motif predicted through pattern search falling within the Pfam domain. (B) SHFT showing the ZnF motif predicted through pattern search falling partially out of the Pfam domain. (C) NMFT representing the case where, no motif(s) is found through pattern search, but Pfam domain is available. (D) XMUT is an extension of (C), wherein, the conserved cysteine or histidine residues are not found in the query. (E) XNCH, is an extension of (C), wherein, there is no conserved cysteine or histidine residues present in the HMM profile. Note that in all the cases, the Pfam domain, consensus sequence and the query sequence are given in blue, brown and yellow colored box respectively. Magenta colored box in (A) and (B) represent the ZnF motif pattern found in the query.

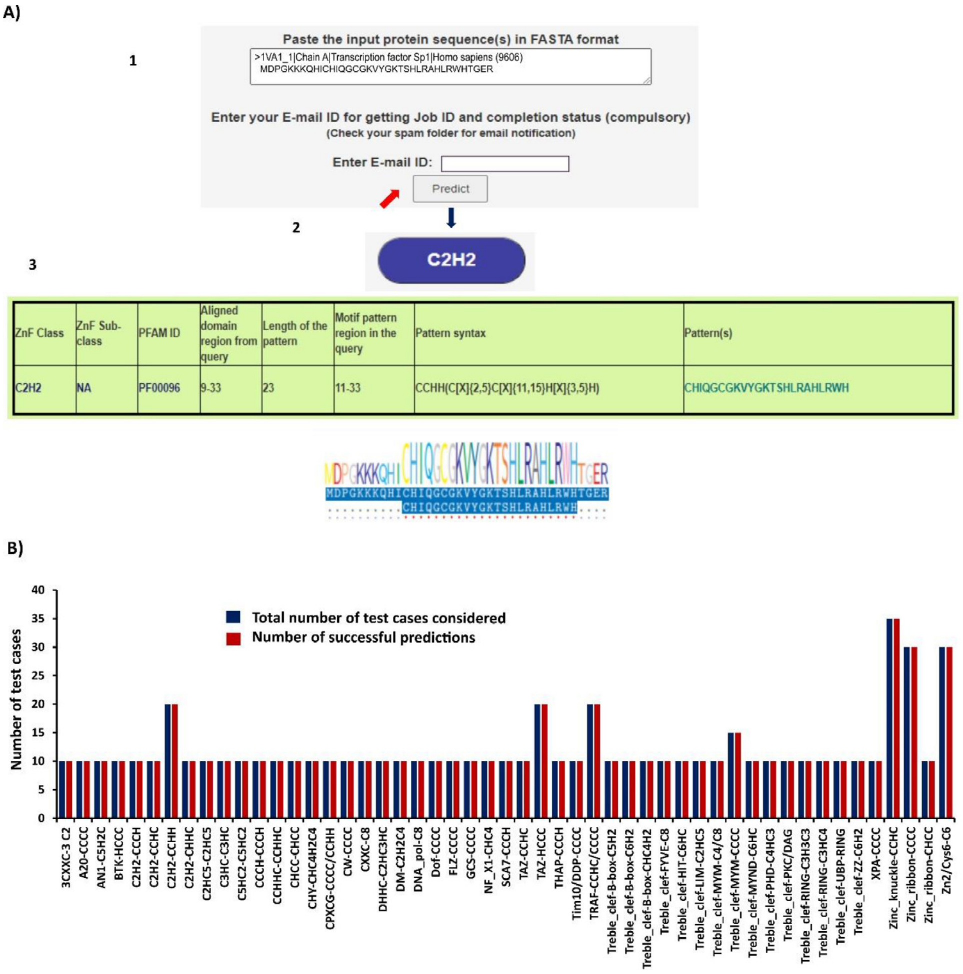
6.1. Functionality of ZnF-Prot web server

ZnF-Prot accepts the FASTA format protein sequence(s) or the PDB ID for the ZnF domain prediction. The user can either paste the protein sequence in the given sequence input window or upload the input protein sequence file. Fig. 5A depicts the web interface of ZnF-Prot, wherein, the ZnF domain prediction is displayed for a given query sequence.

6.2. ZnF domain/motif prediction accuracy of ZnF-Prot

To check the efficacy and accuracy of the prediction, a total of 610 test cases belonging to 31 different zinc finger domain classes from 249 organisms are randomly chosen and tested using the ZnF-Prot web interface. The sequences corresponding to these test cases are taken from NCBI GenBank. Among the 610 test cases, 50, 40, 30, 20, 35, 30 and 165 test cases correspond to C2H2, Zinc ribbon, TAZ domain, TRAF, Zinc knuckle (gag knuckle), Zn2/Cys6 and Treble clef (RING, PHD, LIM, B-box, HIT, MYND, ZZ, THAP, MYM, FYVE and PKC/DAG) respectively. Similarly, 10 test cases each are considered for CCHHC, CW, 3CXXC, DHHC, CXXC, CPXCG, CHCC, CHY, TRAF, FLZ, A20, AN1, C3HC, C2HC5, C5HC2, Dof, DNA-pol, SCA7, GCS, BTK, Tim10/DDP, DM, NF-X1 and XPA zinc finger domains respectively (Fig. 5B). As mentioned in the Experimental setup section, the CTCHY zinc finger is not considered for the prediction as it doesn’t have the Pfam HMM profile.

Out of the 610 test cases considered here, the ZnF-Prot web server accurately predicts 610 zinc finger domains/motifs (Fig. 5B and Supplementary Table S3).



Paste the input protein sequence(s) in FASTA format

>1VA1_1|Chain A|Transcription factor Sp1|Homo sapiens (9606)
MDPGKKKQHICHIQGCGKVYGKTSHLRAHLRWHTGER

Enter your E-mail ID for getting Job ID and completion status (compulsory)
(Check your spam folder for email notification)

Enter E-mail ID:

ZnF Class	ZnF Sub-class	PFAM ID	Aligned domain region from query	Length of the pattern	Motif pattern region in the query	Pattern syntax	Pattern(s)
C2H2	NA	PF00096	9-33	23	11-33	CCHH(C[X]{2,5}C[X]{11,15}H[X]{3,5}H)	CHIQGCGKVYGKTSHLRAHLRWH

Fig. 5. Zinc finger domain(s)/motif(s) prediction using ZnF-Prot. (A) The accurate prediction of C2H2 (2) zinc finger domain present in the PDB ID: 1VA1 (1) and the summary of sequence pattern search output (3) are shown. (B) Bar chart illustrating the prediction accuracy of ZnF-Prot for the 610 test cases correspond to 56 motifs of 31 ZnF domains. The dark blue color bar represents the total number of test cases considered under each zinc finger sub-class and red color bar represents the corresponding successful predictions.

6.3. Elimination of HMM profile based false predictions with the inclusion of sequence pattern

The combined use of sequence pattern along with the Pfam HMM profile employed in ZnF-Prot is useful in improving the ZnF motif prediction accuracy when a Pfam match is reported at an insignificant level (*viz.*, with a significant value of zero). Such insignificant Pfam matches may or may not be reliable. For example, Pfam reports a zf-PARP domain as a match at an insignificant level (Fig. S1A) for the UniProt ID: O62366. A detailed comparison between the Pfam scan, UniProt and PROSITE results indicate that the query sequence doesn't have an ZnF motif (Fig. S1). A contrary example is the UniProt ID: Q08562, wherein, the presence of zf-RING_5 domain is reported at an insignificant level which is confirmed to be a true ZnF-domain by UniProt and PROSITE (Fig. S2). The situation becomes more challenging when more than one Pfam matches are seen at the insignificant level (Fig. S1). Thus, these examples illustrate the challenges in deciding the reliability of the insignificant Pfam predictions. However, the use of sequence pattern along with Pfam HMM profile in ZnF-Prot readily distinguishes the true-positive and true-negative results (Fig. S3).

Further, utilization of ZnF sequence patterns and Pfam HMM profile helps in eliminative false-positives in mutated situations as discussed below.

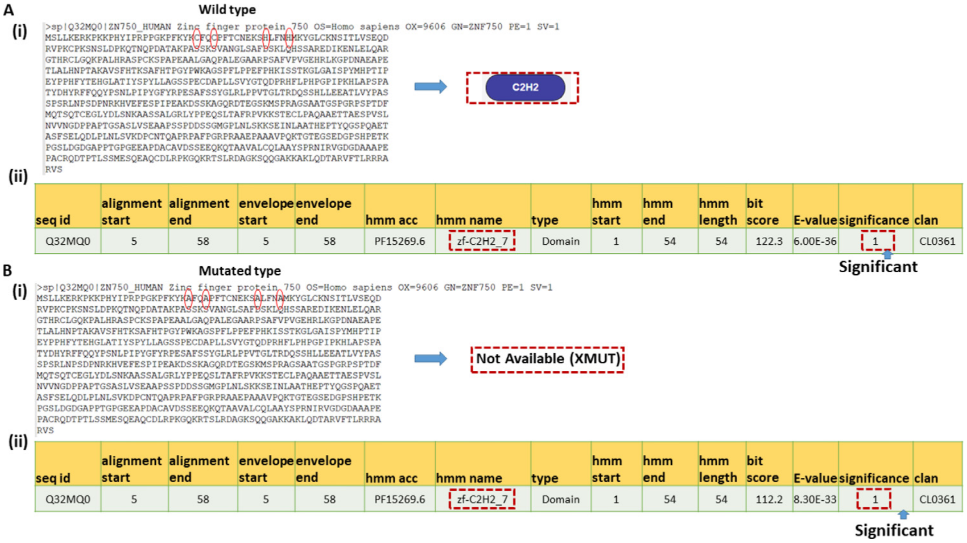
7. Applications

7.1. ZnF-Prot as a diagnostic tool to predict the mutated ZnF domains

Any alternation in the ZnF motif leads to several diseases [9]. One such example is the mutations in ZnF750 human transcription factor which leads to diseases like psoriasis [44]. ZnF-Prot accurately predicts the presence of ZnF domain in the wild-type ZnF750 human transcription factor (Uniprot ID: Q32MQ0) and its absence in the mutant (Fig. 6). This is possible because both the sequence pattern and HMM profile searches are employed in ZnF-Prot. It is noteworthy that the use of sequence to HMM profile search alone will wrongly predict the presence of ZnF domain in the mutated ZnF750 human transcription factor.

7.2. Proteome wide prediction of ZnF domains using ZnF-Prot

Besides predicting the ZnF domain(s)/motif(s) present in a protein sequence, ZnF-Prot also predicts the ZnF domains present in the whole proteome sequence of any organism. Since the methodology simply requires the sequence information to predict the ZnF domain, it can report the presence of ZnF domain(s) in a protein or proteome sequence without any structural prerequisite. Thus, the user can simply upload a text file that contains the sequences (in the FASTA format with each protein sequence separated by a ">"



```
>sp|Q32MQ0|ZN750_HUMAN Zinc finger protein 750 OS=Homo sapiens OX=9606 GN=ZNF750 PE=1 SV=1
MSLLKERKPKKPHYIPRPPGKPFKYKCFQCPFTCNEKSHLFNHMKYGLCKNSITLVSEQD
RVPKCPKSNSLDPKQTNQPDATAKPASSKSVANGLSAFDSKLQHSSAREDIKENLELQAR
GTHRCLGQKPALHRASPCKSPAPEAALGAQPALEGAARPSAFVPVGEHRLKGPDNAEAPE
TLALHNPTAKAVSFHTKSAFHTPGYPWKAGSPFLPPEFPHKISSTKGLGAISPYMHPTIP
EYPPHFYTEHGLATIYSPYLLAGSSPECDAPLLSVYGTQDPRHFLPHPGPIPKHLAPSPA
TYDHYRFFQQYPSNLPIPYGFYRPESAFSSYGLRLPPVTGLTRDQSSHLLEEATLVYPAS
SPSRLNPSDPNRKHVEFESPIPEAKDSSKAGQRDTEGSKMSPRAGSAATGSPGRPSPTDF
MQTSQTCEGLYDLSNKAASSALGRLYPPEQSLTAFRPVKKSTECLPAQAAETTAESPVSL
NVVNGDPPAPTGSASLVSEAAPSSPDDSSGMGPLNLSKKSEINLAATHEPTYQGSPQAET
ASFSELQDLPLNLSVKDPCNTQAPRPAFPGRPRAAEPAAAVPQKTGTEGSEDGPSHPETK
PGSLDGDGAPPTGPGEEAPDACAVDSSEEQKQTAAVALCQLAAYSPRNIRVGDGDAAAPE
PACRQDTPTLSSMESQEAQCDLRPKGQKRTSLRDAGKSQQGAKKAKLQDTARVFTLRRRA
RVS
```

seq id	alignment start	alignment end	envelope start	envelope end	hmm acc	hmm name	type	hmm start	hmm end	hmm length	bit score	E-value	significance	clan
Q32MQ0	5	58	5	58	PF15269.6	zf-C2H2_7	Domain	1	54	54	122.3	6.00E-36	1	CL0361

```
>sp|Q32MQ0|ZN750_HUMAN Zinc finger protein 750 OS=Homo sapiens OX=9606 GN=ZNF750 PE=1 SV=1
MSLLKERKPKKPHYIPRPPGKPFKYKAFQAPFTCNEKSALFNAMKYGLCKNSITLVSEQD
RVPKCPKSNSLDPKQTNQPDATAKPASSKSVANGLSAFDSKLQHSSAREDIKENLELQAR
GTHRCLGQKPALHRASPCKSPAPEAALGAQPALEGAARPSAFVPVGEHRLKGPDNAEAPE
TLALHNPTAKAVSFHTKSAFHTPGYPWKAGSPFLPPEFPHKISSTKGLGAISPYMHPTIP
EYPPHFYTEHGLATIYSPYLLAGSSPECDAPLLSVYGTQDPRHFLPHPGPIPKHLAPSPA
TYDHYRFFQQYPSNLPIPYGFYRPESAFSSYGLRLPPVTGLTRDQSSHLLEEATLVYPAS
SPSRLNPSDPNRKHVEFESPIPEAKDSSKAGQRDTEGSKMSPRAGSAATGSPGRPSPTDF
MQTSQTCEGLYDLSNKAASSALGRLYPPEQSLTAFRPVKKSTECLPAQAAETTAESPVSL
NVVNGDPPAPTGSASLVSEAAPSSPDDSSGMGPLNLSKKSEINLAATHEPTYQGSPQAET
ASFSELQDLPLNLSVKDPCNTQAPRPAFPGRPRAAEPAAAVPQKTGTEGSEDGPSHPETK
PGSLDGDGAPPTGPGEEAPDACAVDSSEEQKQTAAVALCQLAAYSPRNIRVGDGDAAAPE
PACRQDTPTLSSMESQEAQCDLRPKGQKRTSLRDAGKSQQGAKKAKLQDTARVFTLRRRA
RVS
```

seq id	alignment start	alignment end	envelope start	envelope end	hmm acc	hmm name	type	hmm start	hmm end	hmm length	bit score	E-value	significance	clan
Q32MQ0	5	58	5	58	PF15269.6	zf-C2H2_7	Domain	1	54	54	112.2	8.30E-33	1	CL0361

Fig. 6. Illustration of the use of ZnF-Prot as a diagnostic tool by considering the (A) wild-type and (B) mutated ZnF750 human transcription factor (Uniprot ID:Q32MQ0). Note that both ZnF-Prot (A (i)) and Pfam Scan (A (ii)) predict the ZnF domain. The absence of ZnF domain/motif is seen in the ZnF-Prot output of mutated ZnF750 human transcription factor (B (i)), whereas, Pfam Scan shows the presence of the ZnF domain (B (ii)). The cysteines (Cs) and histidines (Hs) in (A) and the mutations in the equivalent positions in (B) are circled.

header) corresponding to all the proteins present in a proteome to ZnF-Prot server. Using the methodology given in Fig. 3, the ZnF-Prot predicts the ZnF domains present in the whole proteome of an organism. This feature of ZnF-Prot is tested by considering the whole proteome sequences of *Arabidopsis thaliana, Homo sapiens, Saccharomyces cerevisiae*, *Caenorhabditis elegans* and *Ciona intestinalis.*

Initially, the experimentally determined three dimensional structures of *Arabidopsis thaliana* and *Homo sapiens* proteomes are considered as the test cases to precisely validate the ZnF domains/motifs predicted by ZnF-Prot. Using the UniProt whole proteome annotation datasheets (*H. sapiens* ID = UP000005640 and *A. thaliana* ID = UP000006548), the protein databank (PDB) IDs of the zinc finger containing protein structures of *A. thaliana* and *H. sapiens* are obtained. Subsequently, the corresponding 3D-structures are downloaded from PDB automatically with a help of an in-house python script by including the “Zn” keyword to download only the structures that have zinc. A total of 56 (*A. thaliana*) and 489 (*H. sapiens*) unique protein structures are found to have zinc finger. ZnF-Prot accurately predicts 52 (*A. thaliana*, Supplementary Table S4) and 470 (*H. sapiens*, Supplementary Table S5) ZnF domains/motifs (95.8%). A detailed inspection indicates that ZnF-Prot fails to predict the remaining ZnF domains/motifs either due to the absence of Pfam HMM profile (3.5%) or due to the presence of novel (*viz.,* less frequently observed) ZnF domains (0.7%).

Encouraged by the ZnF-Prot prediction results of *Arabidopsis thaliana* and *Homo sapiens,* ZnF domains/motifs present in the entire proteome of *Saccharomyces cerevisiae*, *Caenorhabditis elegans* and *Ciona intestinalis* are investigated using their sequences. For this, *Saccharomyces cerevisiae* proteome which contains 6050 proteins is downloaded from the UniProt database (Proteome ID: UP000002311) and is uploaded as an input in the ZnF-Prot web server for the prediction of zinc finger domains/motifs. ZnF-Prot web server predicts the presence of 341 zinc finger domains/motifs (in 334 proteins) that belong to 17 zinc finger domain classes (Fig. 7A and Supplementary Table S6). Similarly, *Caenorhabditis elegans* proteome which contains 26,620 proteins is downloaded from the UniProt database (Proteome ID: UP000001940) and subjected to ZnF domain prediction using ZnF-Protein. ZnF-Prot predicts the presence of 1652 zinc finger domains/motifs (belonging to 28 ZnF domain classes) in 1625 proteins (Fig. 7B and Supplementary Table S7). In the *Ciona intestinalis* proteome, ZnF-Prot predicts 784 zinc finger domains/motifs that belong to 24 zinc finger domains/motifs classes in 744 proteins (UniProt Proteome ID: UP000008144, contains 17,309 protein sequences) (Fig. 7C and Supplementary Table S8).

7.2.1. Validation of the prediction

The ZnF domains/motifs predicted (refer Experimental setup section for the gathering threshold and E-value used for the prediction) to be present in the proteomes of *S. cerevisiae*, *C. elegans* and *C. intestinalis* are cross-validated using UniProt [45], PROSITE [46], NCBI-CDD [47] and InterPro [48] as these tools complement each other in providing the ZnF motif containing domain information (Fig. 8 and Fig. S4). Initially, ZnF-Prot prediction output is cross-validated using UniProt annotation datasheet. Secondly, the ZnF motifs that are not annotated by UniProt, but, are reported by ZnF-Prot are validated through PROSITE. It is worth mentioning that the default normalized score (N_score) of 8.5 or an E-value of 0.32 is used for the PROSITE prediction. The ZnF-motifs that couldn’t be verified through both UniProt and PROSITE are next validated by NCBI-CDD (the default E-Value of 0.01 for the query search against position specific scoring matrix (PSSM) is used for the prediction) and subsequently by InterPro (does not provide an E-value as it uses different member databases which uses different scoring systems for the prediction, thus, may not provide a meaningful information). The remaining ZnFs are validated through literature, if available.

87–93% of ZnF domains/motifs predicted by ZnF-Prot for the whole proteomes of *Saccharomyces cerevisiae* (Supplementary Table S9), *Caenorhabditis elegans* (Supplementary Table S10) and *Ciona* intestinalis (Supplementary Table S11) are successfully cross-validated by verifying their presence in at least one of the above-mentioned online tools. In all the above three proteomes, ZnF-Prot reports 11% additional ZnF motifs/domains which are yet to be validated.

8. Limitations of ZnF-Prot

Although the ZnF motif patterns are systematically and carefully collected through literature survey, still unidentified ZnF motif patterns may exist in nature. Due to this reason, ZnF-Prot may predict

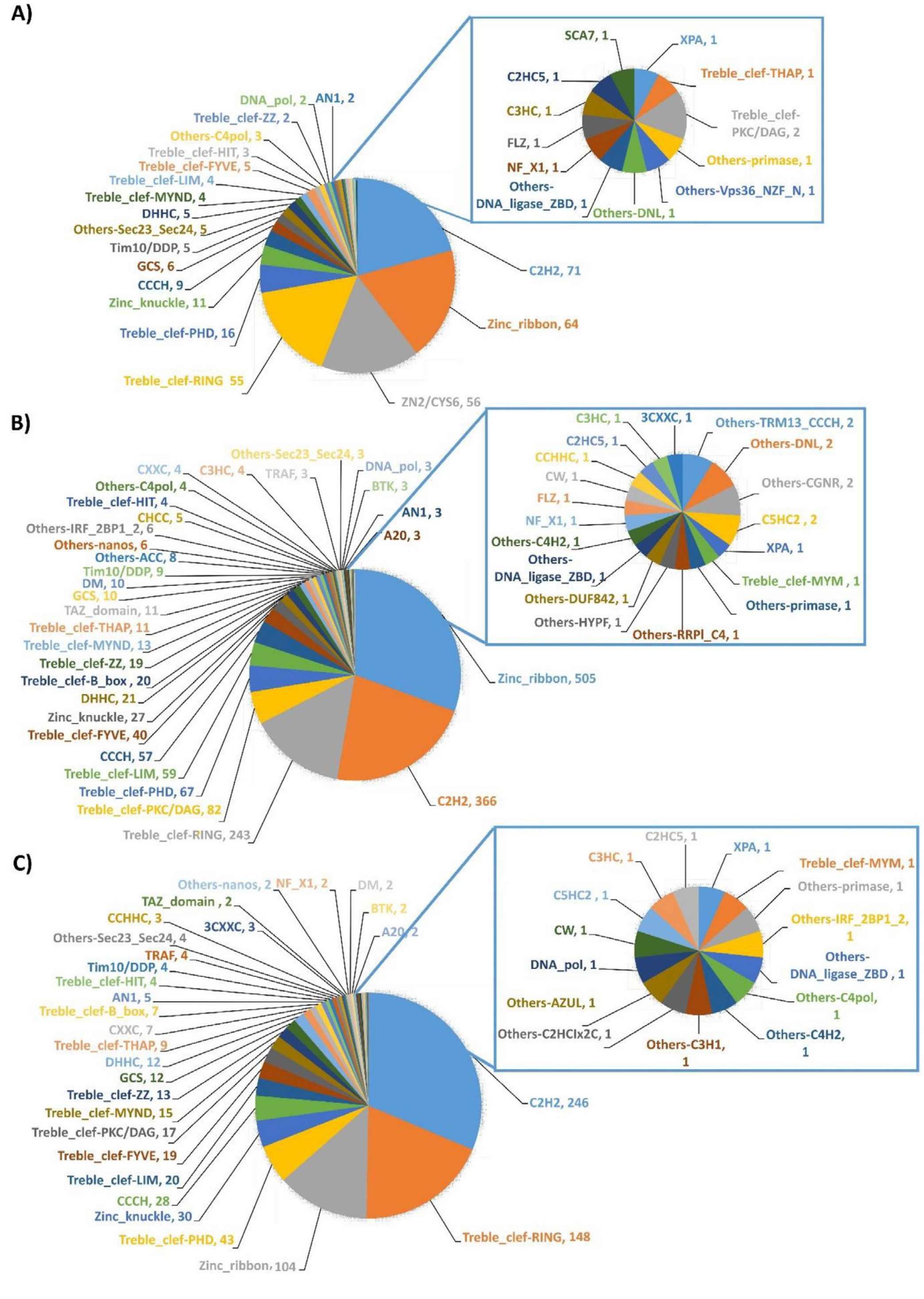


Fig. 7. Proteome wide zinc finger domain prediction using ZnF-Prot. Pie chart illustrating the statistics of different zinc finger domains predicted for (A) *Saccharomyces cerevisiae* (Proteome ID: UP000002311) (B) *Caenorhabditis elegans* (Proteome ID: UP000001940) (C) *Ciona intestinalis* (Proteome ID: UP000008144) proteomes.

a few false-negative ZnF motifs. One such example is the UniProt ID: Q8NB78, which is a novel and rare ZnF motif [49] and has not yet been classified under any ZnF motifs. The Pfam HMM profile is also not available for this motif. Thus, ZnF-Prot doesn't predict the ZnF domain containing this motif. To addresses such issues, ZnF motif sequence patterns and Pfam HMM profiles are being updated time-to-time in ZnF-Prot. One such example is Uniprot ID: Q9STM3 [50], wherein, two ZnF domains (PDB ID: 6A57) [51] together forms a novel ZnF domain with a unique sequence pattern. After identifying such a novel domain during the revision of the manuscript, the sequence pattern "C-X_4-C-X_{12}-H-X_{16}-H" has been included in the ZnF-Prot local sequence pattern library and validated.

Thus, the ZnF domain/motif prediction accuracy of ZnF-Prot will further be improved when more ZnF motif sequence patterns from multiple organisms become available.

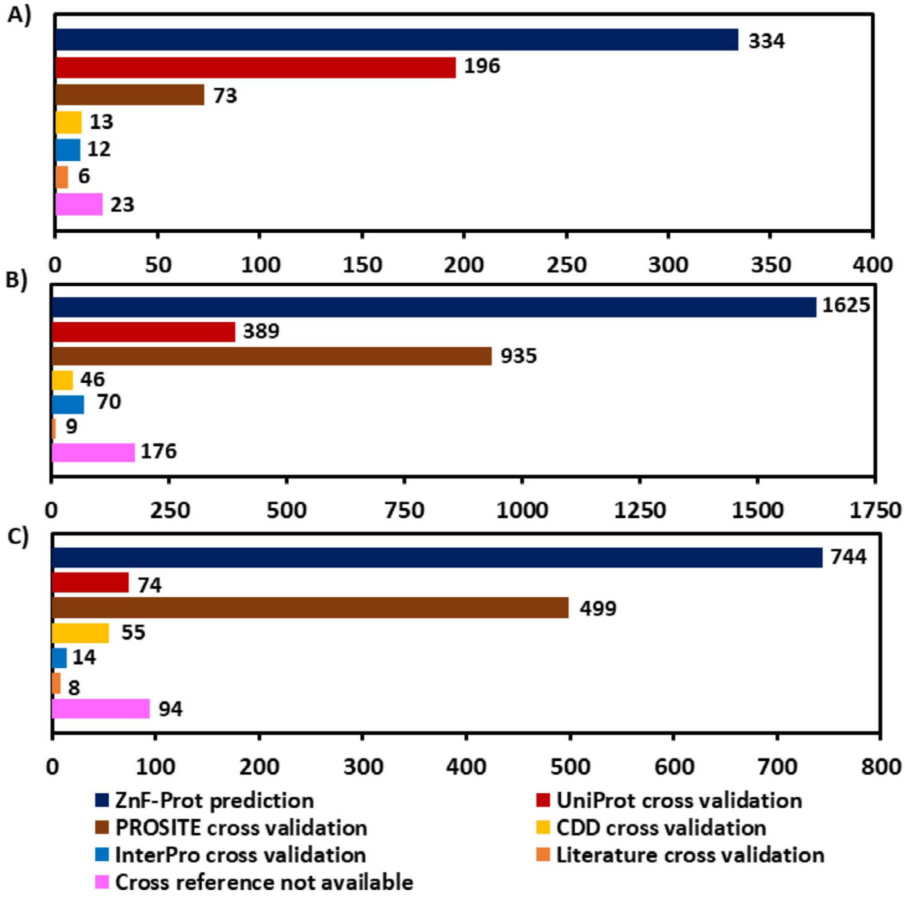


Fig. 8. Cross-validation of the proteome wide ZnF motifs predicted by ZnF-Prot. (A-C) Bar diagram illustrating the statistics of ZnF motifs predicted in (A) *S. cerevisiae*, (B) *C. elegans* and (C) *C. intestinalis* and their cross-validation. Note that the bar chart provides the comparison between ZnF-Prot prediction (colored dark blue) and, UniProt annotation (colored red), PROSITE (colored brown), NCBI-CDD (colored yellow), InterPro (colored blue) and literature (colored orange)and the corresponding ZnF motifs/domain numbers adjacent to each bar.

9. Conclusion

Zinc finger (ZnF) is one of the large families of metalloproteins, wherein, the zinc metal ion coordinates with the cysteine and/or histidine residues of the ZnF domain. Until now, there is no single platform available to predict any type of ZnF motif containing domains present in a protein sequence. To this end, a systematic classification of 32 ZnF domains is carried out and a local repository of their sequence motif patterns and Pfam HMM profiles are created. A web server, namely, ZnF-Prot which uses these sequence motif patterns and Pfam HMM profiles is developed here to predict the zinc finger domain(s)/motif(s) present in the protein/proteome of any organism. ZnF-Prot not only successfully predicts the presence of 93.1%, 87.4% and 89% ZnF domains/motifs respectively in the proteomes of *Saccharomyces cerevisiae, Caenorhabditis elegans* and *Ciona intestinalis* but also suggest the presence of 11% additional ZnF-motifs in these organisms. Thus, ZnF-Prot serves as a valuable tool in the proteome wide prediction of ZnF domains/motifs present in any organism. In the future, newly discovered ZnF domain(s) will periodically be updated to improve the prediction accuracy.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT authorship contribution statement

Chakkarai Sathyaseelan: Data curation, Methodology, Visualization, Formal analysis, Validation, Writing – original draft. **L Ponoop Prasad Patro:** Methodology, Visualization, Writing – review & editing. **Thenmalarchelvi Rathinavelan:** Conceptualization, Supervision, Writing – original draft.

Data availability

Data will be made available on request.

Acknowledgments

The authors thank IITH and CDAC for the computational resources. CS and LPPP thank Ministry of Education, Government of India for the fellowships.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.patcog.2022.109134.

References

[1] M. Padjasek, A. Kocyla, K. Kluska, O. Kerber, J.B. Tran, A. Krezel, Structural zinc binding sites shaped for greater works: structure-function relations in classical zinc finger, hook and clasp domains, J. Inorg. Biochem. 204 (2020) 110955.
[2] C. Andreini, I. Bertini, G. Cavallaro, Minimal functional sites allow a classification of zinc sites in proteins, PLoS ONE 6 (10) (2011) e26325.

[3] A.S. Prasad, Discovery of human zinc deficiency: its impact on human health and disease, Adv. Nutr. 4 (2013) 176–190.
[4] E. Kelkitli, N. Ozturk, N.A. Aslan, N. Kilic-Baygutalp, Z. Bayraktutan, N. Kurt, N. Bakan, E. Bakan, Serum zinc levels in patients with iron deficiency anemia and its association with symptoms of iron deficiency anemia, Ann. Hematol. 95 (2016) 751–756.
[5] S.M. Ireland, A.C.R. Martin, ZincBind-the database of zinc binding sites, J. Biol. Databases Curation (2019) 2019:baz006.
[6] W. Maret, Zinc biochemistry: from a single zinc enzyme to a key element of life, Adv. Nutr. 4 (2013) 82–91.
[7] N.J. Pace, E. Weerapana, Zinc-binding cysteines: diverse functions and structural motifs, Biomolecules 4 (2014) 419–434.
[8] K.S. Eom, J.S. Cheong, S.J. Lee, Structural analyses of zinc finger domains for specific interactions with DNA, J. Microbiol. Biotechnol. 26 (2016) 2019–2029.
[9] M. Cassandri, A. Smirnov, F. Novelli, C. Pitolli, M. Agostini, M. Malewicz, G. Melino, G. Raschella, Zinc-finger proteins in health and disease, Cell Death Discov. 3 (2017) 17071.
[10] A. Klug, J.W. Schwabe, Protein motifs 5. Zinc fingers, FASEB J.: Off. Publ. Fed. Am. Soc. Exp. Biol. 9 (1995) 597–604.
[11] N. Iraci, O. Tabarrini, C. Santi, L. Sancineto, NCp7: targeting a multitask protein for next-generation anti-HIV drug development part 2. Noncovalent inhibitors and nucleic acid binders, Drug Discov. Today 23 (2018) 687–695.
[12] S. Vyas, P. Chang, New PARP targets for cancer therapy, Nat. Rev. Cancer 14 (2014) 502–509.
[13] C. Abbehausen, Zinc finger domains as therapeutic targets for metal-based compounds - an update, Met.: Integr. Biometal Sci. 11 (2019) 15–28.
[14] K. Pavlovic, M. Tristan-Manzano, N. Maldonado-Perez, M. Cortijo-Gutierrez, S. Sanchez-Hernandez, P. Justicia-Lirio, M.D. Carmona, C. Herrera, F. Martin, K. Benabdellah, Using gene editing approaches to fine-tune the immune system, Front. Immunol. 11 (2020) 570672.
[15] M.S. Kim, A.G. Kini, Engineering and application of zinc finger proteins and TALEs for biomedical research, Mol. Cells 40 (2017) 533–541.
[16] S. Dutta, S. Madan, H. Parikh, D Sundar, An ensemble micro neural network approach for elucidating interactions between zinc finger proteins and their target DNA, BMC Genom. 17 (2016) 1033.
[17] S.S. Krishna, I. Majumdar, N.V. Grishin, Structural classification of zinc fingers: survey and summary, Nucleic Acids Res. 31 (2003) 532–550.
[18] Peyman Neamatollahi, Montassir Hadi, M Naghibzadeh, Simple and efficient pattern matching algorithms for biological sequences, IEEE Access 8 (2020) 23838–23846.
[19] W. Zhao, J.J. Chen, S. Foley, Y. Wang, S. Zhao, J. Basinger, W. Zou, Biomarker identification from next-generation sequencing data for pathogen bacteria characterization and surveillance, Biomark. Med. 9 (2015) 1253–1264.
[20] Brejova Brona, Dimarco Chrysanne, Vinar Tomas, Romero-Hidalgo Sandra, Holguin Gina, P Cheryl, Finding patterns in biological sequences, Technical Report CS-2000-22, University of Waterloo, 2001.
[21] Osman Ali Sadek Ibrahim, Belal A. Hamed, T.A. El-Hafeez, A new fast technique for pattern matching in biological sequences, J. Supercomput. (2022), doi:10.1007/s11227-022-04673-3.
[22] B.J. Yoon, Hidden Markov models and their applications in biological sequence analysis, Curr. Genom. 10 (2009) 402–415.
[23] L.S. Johnson, S.R. Eddy, E. Portugaly, Hidden Markov model speed heuristic and iterative HMM search procedure, BMC Bioinform. 11 (2010) 431.
[24] S. Wang, J. Peng, J. Xu, Alignment of distantly related protein structures: algorithm, bound and implications to homology modeling, Bioinformatics 27 (2011) 2537–2545.
[25] M. Jayakanthan, J. Muthukumaran, S. Chandrasekar, K. Chawla, A. Punetha, D Sundar, ZifBASE: a database of zinc finger proteins and associated resources, BMC Genom. 10 (2009) 421.
[26] X. Hu, Q. Dong, J. Yang, Y. Zhang, Recognizing metal and acid radical ion-binding sites by integrating *ab initio* modeling with template-based transferals, Bioinformatics 32 (2016) 3260–3269.
[27] S.M. Ireland, A.C.R. Martin, Zincbindpredict-prediction of zinc binding sites in proteins, Molecules 26 (4) (2021) 966.
[28] K.A. McCall, C. Huang, C.A. Fierke, Function and mechanism of zinc metalloenzymes, J. Nutr. 130 (2000) 1437S–1446S.
[29] J.H. Laity, B.M. Lee, P.E. Wright, Zinc finger proteins: new insights into structural and functional diversity, Curr. Opin. Struct. Biol. 11 (2001) 39–46.
[30] Y.F. Lin, C.W. Cheng, C.S. Shih, J.K. Hwang, C.S. Yu, C.H. Lu, MIB: metal ion-binding site prediction and docking server, J. Chem. Inf. Model. 56 (2016) 2287–2291.
[31] T. James, Cheminformatics in the service of GPCR drug discovery, Methods Mol. Biol. 1705 (2018) 395–411.
[32] J.K. Das, C. Heryakusuma, D. Susanti, P.P. Choudhury, B. Mukhopadhyay, Reduced protein sequence patterns in identifying key structural elements of dissimilatory sulfite reductase homologs, Comput. Biol. Chem. 98 (2022) 107691.
[33] Mourad Elloumi, Costas Iliopoulos, Jason T.L. Wang, A.Y. Zomaya, Pattern Recognition in Computational Molecular Biology: Techniques and Approaches, Wiley, 2015.
[34] T.D.C. Negri, W.A.L. Alves, P.H. Bugatti, P.T.M. Saito, D.S. Domingues, A.R. Paschoal, Pattern recognition analysis on long noncoding RNAs: a tool for prediction in plants, Brief. Bioinform. 20 (2019) 682–689.
[35] A. Via, P.F. Gherardini, E. Ferraro, G. Ausiello, G. Scalia Tomba, M. Helmer-Citterich, False occurrences of functional motifs in protein sequences highlight evolutionary constraints, BMC Bioinform. 8 (2007) 68.
[36] S.R. Eddy, Accelerated profile HMM searches, PLoS Comput. Biol. 7 (2011) e1002195.
[37] Sasikumar Ramaraj, K V, Profile hidden markov model for sequence alignment to cancer sequence, Glob. J. Pure Appl. Math. 11 (2015) 3665–3675.
[38] T.M. Poulsen, M. Frith, Variable-order sequence modeling improves bacterial strain discrimination for Ion Torrent DNA reads, BMC Bioinform. 18 (2017) 299.
[39] C.K. Vilas, L.E. Emery, E.L. Denchi, K.M. Miller, Caught with one's zinc fingers in the genome integrity cookie jar, Trends Genet. TIG 34 (2018) 313–325.
[40] K. Kluska, J. Adamczyk, A Krezel, Metal binding properties, stability and reactivity of zinc fingers, Coord. Chem. Rev. 367 (2018) 18–64.
[41] T. Hard, A. Rak, P. Allard, L. Kloo, M. Garber, The solution structure of ribosomal protein L36 from Thermus thermophilus reveals a zinc-ribbon-like fold, J. Mol. Biol. 296 (2000) 169–180.
[42] J. Mistry, S. Chuguransky, L. Williams, M. Qureshi, G.A. Salazar, E.L.L. Sonnhammer, S.C.E. Tosatto, L. Paladin, S. Raj, L.J. Richardson, et al., Pfam: the protein families database in 2021, Nucleic Acids Res. 49 (2021) D412–D419.
[43] S.C. Potter, A. Luciani, S.R. Eddy, Y. Park, R. Lopez, R.D. Finn, HMMER web server: 2018 update, Nucleic Acids Res. 46 (2018) W200–W204.
[44] R.Y. Birnbaum, G. Hayashi, I. Cohen, A. Poon, H. Chen, E.T. Lam, P.Y. Kwok, O.S. Birk, W. Liao, Association analysis identifies ZNF750 regulatory variants in psoriasis, BMC Med. Genet. 12 (2011) 167.
[45] C. UniProt, UniProt: the universal protein knowledgebase in 2021, Nucleic Acids Res. 49 (2021) D480–D489.
[46] C.J. Sigrist, E. de Castro, L. Cerutti, B.A. Cuche, N. Hulo, A. Bridge, L. Bougueleret, I. Xenarios, New and continuing developments at PROSITE, Nucleic Acids Res. 41 (2013) D344–D347.
[47] S. Lu, J. Wang, F. Chitsaz, M.K. Derbyshire, R.C. Geer, N.R. Gonzales, M. Gwadz, D.I. Hurwitz, G.H. Marchler, J.S. Song, et al., CDD/SPARCLE: the conserved domain database in 2020, Nucleic Acids Res. 48 (2020) D265–D268.
[48] M. Blum, H.Y. Chang, S. Chuguransky, T. Grego, S. Kandasaamy, A. Mitchell, G. Nuka, T. Paysan-Lafosse, M. Qureshi, S. Raj, et al., The InterPro protein families and domains database: 20 years on, Nucleic Acids Res. 49 (2021) D344–D354.
[49] Q. Zhang, S. Qi, M. Xu, L. Yu, Y. Tao, Z. Deng, W. Wu, J. Li, Z. Chen, J. Wong, Structure-function analysis reveals a novel mechanism for regulation of histone demethylase LSD2/AOF1/KDM1b, Cell Res. 23 (2013) 225–241.
[50] B. Noh, S.H. Lee, H.J. Kim, G. Yi, E.A. Shin, M. Lee, K.J. Jung, M.R. Doyle, R.M. Amasino, Y.S. Noh, Divergent roles of a pair of homologous jumonji/zinc-finger-class transcription factor proteins in the regulation of Arabidopsis flowering time, Plant Cell 16 (2004) 2601–2613.
[51] Z. Tian, X. Li, M. Li, W. Wu, M. Zhang, C. Tang, Z. Li, Y. Liu, Z. Chen, M. Yang, et al., Crystal structures of REF6 and its complex with DNA reveal diverse recognition mechanisms, Cell Discov. 6 (2020) 17.

Chakkarai Sathyaseelan is a research scholar at the Department of Biotechnology, IIT Hyderabad. His-research interests include development of web tools for biomolecular structure and sequence analysis.

L Ponoop Prasad Patro is a research scholar at the Department of Biotechnology, IIT Hyderabad. His-research interests include development of web tools for biomolecular structure modeling and sequence analysis.

Thenmalarchelvi Rathinavelan is an Associate professor at the Department of Biotechnology, IIT Hyderabad. Her research interests are to understand the biological phenomena using computational and experimental approaches.