Received: 28 June 2019 | Revised: 12 November 2019 | Accepted: 24 February 2020

DOI: 10.1002/prot.25886

PROTEINS WILEY

RESEARCH ARTICLE

FiRES: A computational method for the de novo identification of internal structure similarity in proteins

Claudia Alvarez-Carreño[1,2] | Gerardo Coello[3] | Marcelino Arciniega[1]

[1]Department of Bioquímica y Biología Estructural, Instituto de Fisiología Celular, Universidad Nacional Autónoma de México, Mexico City, Mexico

[2]School of Chemistry and Biochemistry, Georgia Institute of Technology, Atlanta, Georgia

[3]Unidad de Cómputo, Instituto de Fisiología Celular, Universidad Nacional Autónoma de México, Mexico City, Mexico

Correspondence
Claudia Alvarez-Carreño and Marcelino Arciniega, Department of Bioquímica y Biología Estructural, Instituto de Fisiología Celular, Universidad Nacional Autónoma de México, Address Circuito Exterior s/n, Ciudad Universitaria, Apartado Postal 70-243, Mexico City 04510, Mexico.
Email: calvarez@ifc.unam.mx (C.A.-C.) and marciniega@ifc.unam.mx (M.A.)

Funding information
Dirección General de Asuntos del Personal Académico, Universidad Nacional Autónoma de México, Grant/Award Number: PAPIIT-DGAPA-IN213320; Dirección General de Cómputo y de Tecnologías de Información y Comunicación, Universidad Nacional Autónoma de México, Grant/Award Number: LANCAD-UNAM-DGTIC-320; Universidad Nacional Autónoma de México, Grant/Award Number: IN213320

Peer Review
The peer review history for this article is available at https://publons.com/publon/10.1002/prot.25886.

Abstract

Internal structure similarity in proteins can be observed at the domain and subdomain levels. From an evolutionary perspective, structurally similar elements may arise divergently by gene duplication and fusion events but may also be the product of convergent evolution under physicochemical constraints. The characterization of proteins that contain repeated structural elements has implications for many fields of protein science including protein domain evolution, structure classification, structure prediction, and protein engineering. FiRES (Find Repeated Elements in Structure) is an algorithm that relies on a topology-independent structure alignment method to identify repeating elements in protein structure. FiRES was tested against two hand curated databases of protein repeats: MALIDUP, for very divergent duplicated domains; and RepeatsDB for short tandem repeats. The performance of FiRES was compared to that of lalign, RADAR, HHrepID, CE-symm, ReUPred, and Swelfe. FiRES was the method that most accurately detected proteins either with duplicated domains (accuracy = 0.86) or with multiple repeated units (accuracy = 0.92). FiRES is a new methodology for the discovery of proteins containing structurally similar elements. The FiRES web server is publicly available at http://fires.ifc.unam.mx. The scripts, results, and benchmarks from this study can be downloaded from https://github.com/Claualvarez/fires.

KEYWORDS
internal structure similarity, protein domain, protein repeats, structural motif

1 | INTRODUCTION

Protein repeats consist of non-overlapping copies of either subdomain elements or entire domains located within a single protein. These copies, called repeated units, can be arranged in tandem or interspersed throughout the sequence[1-3] and may fold into similar three-dimensional structures. Protein repeats can adopt a variety of native conformations such as intrinsic disorder,[4] globular domains and open structures. For instance, six out of the 10 most prevailing globular domains are formed by repeated units.[5] These units have been studied as remnants of hypothetical peptide-like predecessors of the first folded proteins.[5,6] Open solenoid domains are characteristically formed by a stack of multiple short tandem repeats.[2,7,8] Solenoid domains appear to be an evolutionary adaptation more commonly found in eukaryotes and are usually involved in protein-ligand and

[Correction added on 09 April 2020, after first online publication: Author name Alvarez-Carreño Claudia has been updated to Claudia Alvarez-Carreño.]

protein-protein interactions.[2,9] At the domain level, duplication, fusion and terminal losses, have played a major role in the evolution of the modern repertoire of protein structures and functions.[10-12] Thus, repeated units are functionally and structurally diverse, as are the evolutionary mechanisms that preserve them.

Duplicated sequences accumulate point mutations as well as insertions and deletions, which ultimately hide the trace of similarity between them.[3,6,9,13,14] Sequence divergence determines the amount of structural variation displayed by protein repeats. However, three-dimensional structure changes more slowly over evolutionary time than sequence.[15] For instance, distantly related domains maintain a distinctive core of secondary structural elements even in the absence of significant sequence similarity[16] and circularly permutated proteins frequently preserve the same overall three-dimensional disposition of Cα atoms.[17] In the case of tandem repeats, the disparity between sequence divergence and structure conservation can be extreme.[18] Thus, many studies have incorporated structure-based analysis to facilitate the precise identification of very divergent polypeptide chains. Non-sequential structure comparison algorithms, such as CLICK,[19] allow similarities between protein structures to be detected irrespective of the topological connectivity of their secondary structural elements.

Sequence- and structure-based methods have been developed to identify internal similarity in proteins (Table 1). Algorithms that detect similarity at the sequence level are particularly useful for the analysis of protein repeats because their results facilitate homology inference. Examples of such algorithms include lalign,[20] HHrepID,[21] RADAR,[22] TPRpred,[23] and TRUST.[24] On the other hand, structure-based algorithms can be advantageous for the discovery of remote homologs. Structure-based repeat-detection algorithms (reviewed by Pellegrini[25]) are broadly classified in two categories: (a) inference methods, such as ReUPred,[18] RepeatsDB-Lite,[26] and IRIS,[27] which are based on libraries of already well-characterized reference units; and (b) de novo identification methods, which do not rely on previous knowledge of the features defining a repeated unit. The main strategies employed by de novo identification methods are self-structure comparison, and detection of periodicities using one or more structural parameters. Examples of self-structure comparison methods include CE-symm,[28] DAVROS,[29] GANGSTA+,[30] and SymD.[31] Examples of pattern recognition algorithms include Swelfe,[32] which employs α-angles; ProSTRIP,[33] which calculates dihedral angles; ConSole,[8] which relies on contact maps; and TAPO,[34] which detects periodicities of atomic coordinates and other parameters. It should be noted, however, that structure similarity may be the product of convergent evolution under functional and structural constraints,[35-37] thus, its identification does not unambiguously imply homology. This is especially true at the subdomain level, where structural similarity may represent energetically favorable conformations of secondary structural elements.[36,38,39] Therefore, the present study focuses on *similar structural elements*, which include both homologous and convergently evolved regions of internal structure similarity within a protein.

Here, we present FiRES (Find Repeated Elements in Structure), a computational protocol for the de novo identification of tandem and non-tandem repetitive elements in protein structures. FiRES exploits a topology-independent structure alignment method in order to detect similar groups of elements. The performance of FiRES was assessed

TABLE 1 Algorithms for the study of protein repeats

Algorithm	Type of data	Description
lalign[20]	Sequence	de novo identification of repeats
RADAR[22]	Sequence	de novo identification of repeats
TRUST[24]	Sequence	de novo identification of repeats
TPRpred[23]	Sequence	TPRs, PPRs, and SEL1-like solenoid repeats
HHrepID[21]	Sequence	de novo identification of repeats
ARD2[9]	Sequence	de novo identification of α-solenoid repeats
HMMER[59]/ Pfam[51]	Sequence, Pfam database	Reference-based identification of repeats
DAVROS[29]	Structure	de novo identification of repeats
GANGSTA+[30]	Structure	de novo identification of repeats
REPETITA[60]	Structure	de novo identification of solenoid repeats
SymD[31]	Structure	Identification of internal structure symmetry
ProSTRIP[33]	Structure	de novo identification of repeats
RAPHAEL[61]	Structure	de novo identification of solenoid repeats
CE-symm[28]	Structure	Identification of internal structure symmetry and repeating units
ConSole[8]	Structure	de novo identification of solenoid repeats
REPRO[62]	Sequence	de novo identification of repeats
DeepSymmetry[63]	Structure	de novo identification of tandem repeats
TAPO[34]	Structure	de novo identification of tandem repeats
IRIS[27]	Sequence or structure	Reference-based identification of repeats
RepeatsDB-Lite[26]	Structure, RepeatsDB	Reference-based identification of short tandem repeats
ReUPred[18]	Structure, RepeatsDB	Reference-based identification of short tandem repeats
Swelfe[32]	Sequence or structure	de novo identification of repeats

Abbreviations: TPR, Tetratrico peptide repeat; PPR, Pentatrico peptide repeats.

on two types of data: proteins with short tandem repeats and proteins with very divergent internal domain duplications. Finally, we show that FiRES can be used for the discovery of proteins containing similar structural elements with very low sequence identity (<20%), where homology inference remains an open question.

2 | MATERIALS AND METHODS

The FiRES algorithm (Figure 1) searches internal structure similarity within a protein by an iterative self-alignment process, which includes a scoring system based on the template modeling score[40] (TM-score).

2.1 | Generation of query-target pairs

FiRES generates a series of alignments between a fragment of an input protein structure and the input protein structure itself. To avoid alignments at the diagonal, each fragment, referred to as query q_i, is aligned to a target region t_i defined as the complement of q_i over the protein P:

$$t_i = \{P \backslash q_i\} \tag{1}$$

During the first iteration, the size of each query q_i depends on the total number N of secondary structural elements (SSEs) in the protein P. To determine the number of SSEs, each residue in P is tagged with a secondary structure label, namely helix, strand or loop. These tags are assigned by the DSSP algorithm.[41] Residues presenting geometrical features that do not classify as helix or strand are treated as loops. Consecutive residues with the same secondary structure label are grouped into SSEs. The maximal number n of SSEs within each query is calculated as follows:

$$n = \begin{cases} round\left(\frac{N}{4}\right), N < 28 \\ 7, N \geq 28 \end{cases} \tag{2}$$

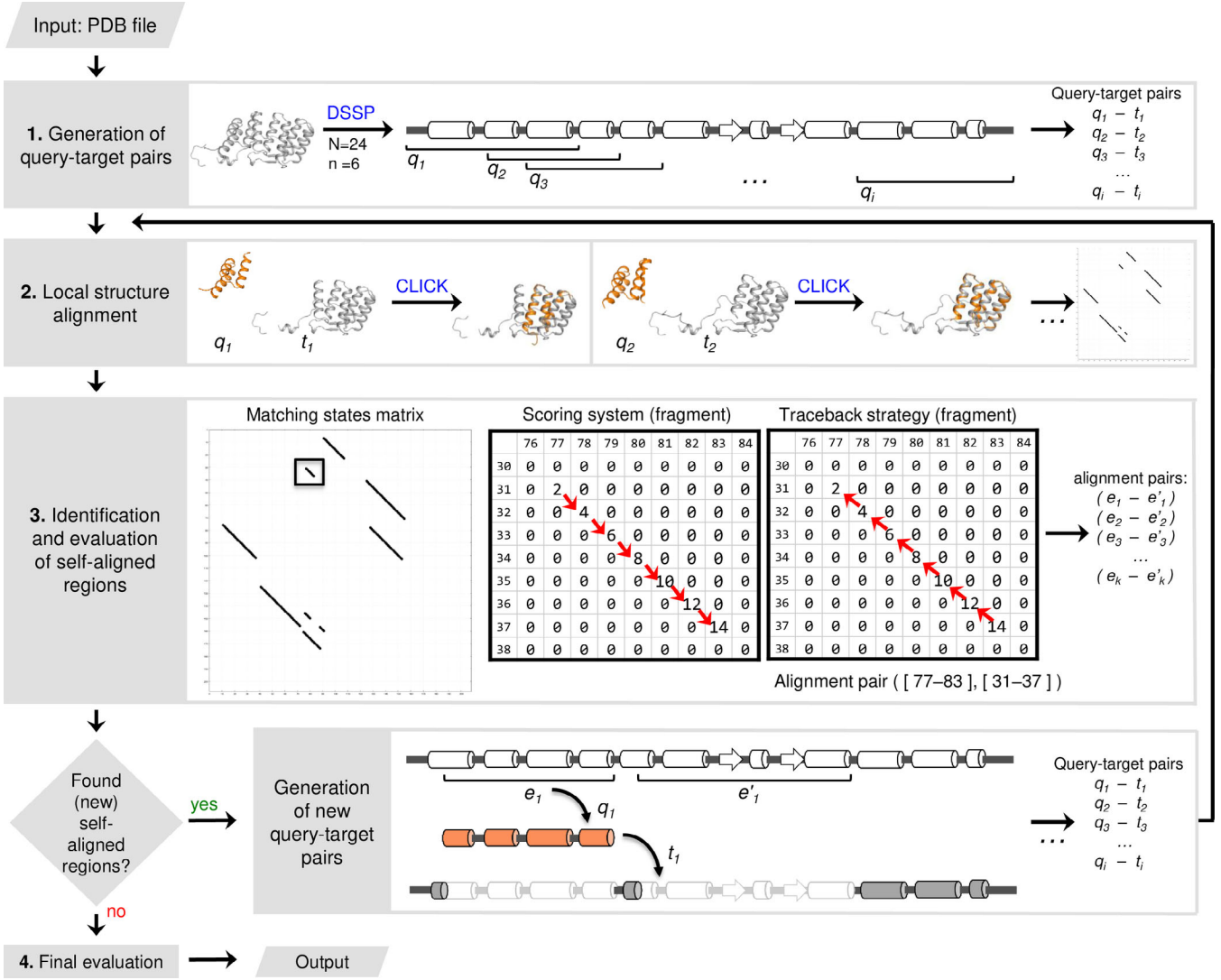


FIGURE 1 Flow diagram indicating the main steps of the FiRES algorithm [Color figure can be viewed at wileyonlinelibrary.com]

Where N is the total number of SSEs in P and n is the maximum number of SSEs in each query q_i. The set of i queries is generated by shifting the starting position of q_i to the next helix or strand (Figure 1, step 1).

2.2 | Local structure alignment

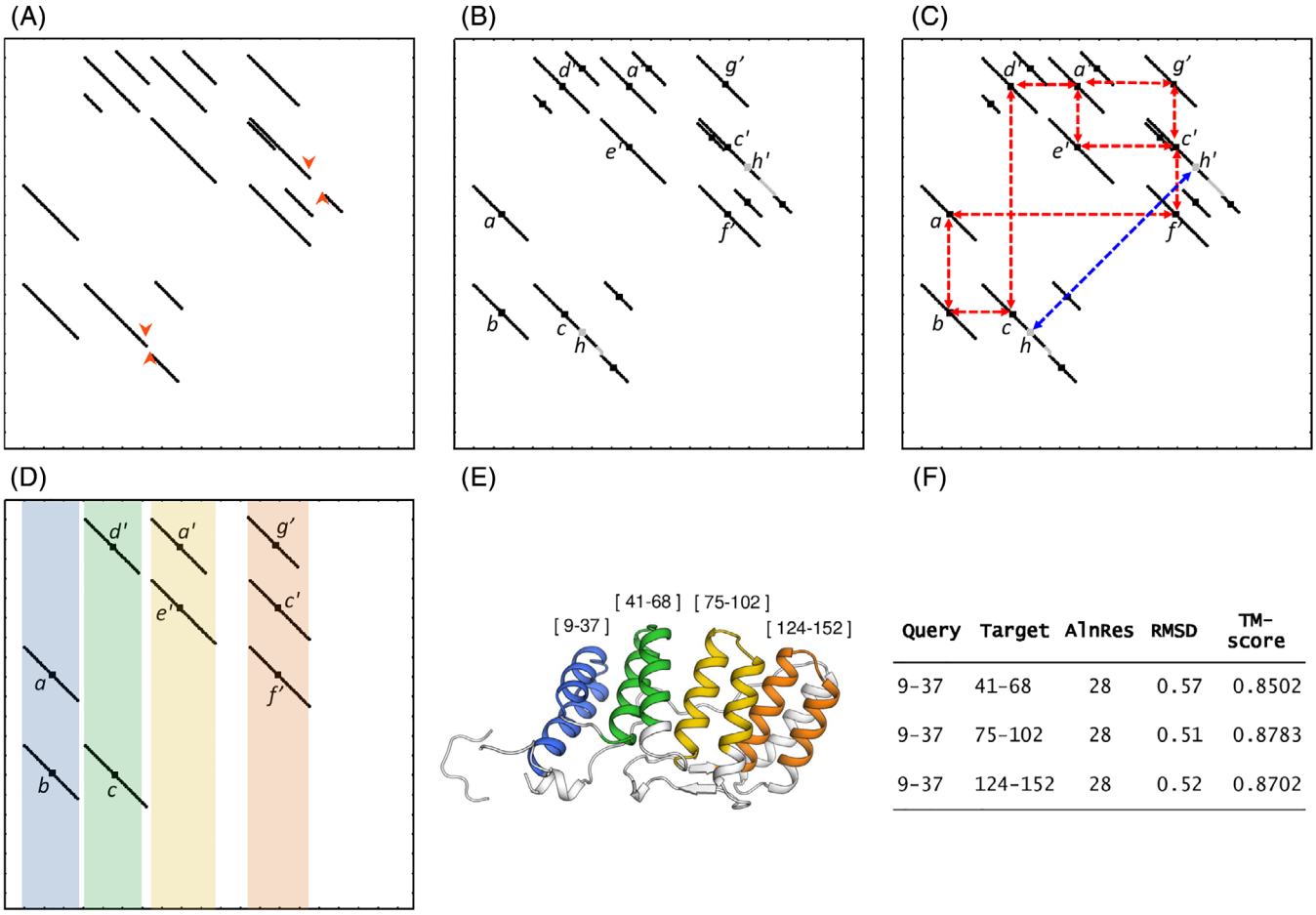
Query-target pairs are aligned using the CLICK algorithm with default parameters.[19] CLICK is a graph theory-based algorithm, which employs the Cartesian coordinates of Cα atoms of the query and target structures to form cliques. CLICK produces pairwise alignments by iteratively matching increasingly larger cliques of points of the query and target structures. For each query-target pair, CLICK returns pairs of Cα atoms that render a low RMSD alignment. When the whole set of query-target pairs has been aligned with CLICK, all pairs of matching residues are stored in a single two-dimensional matrix (Figure 1, step 2).

2.3 | Identification and evaluation of self-aligned regions

From the two-dimensional matrix containing matching states, local alignments are determined by a dynamic programming procedure, which uses the Smith-Waterman[42] scoring system and traceback strategy (Figure 1, step 3). The following parameters were used to generate the scoring matrix from which the local alignments are determined: match +2; gap-opening −1; and no gap-extensions. Only locally aligned regions with lengths over 10 residues are maintained. By the end of this step, the first element e_k of the aligned pair (e_k, e'_k) becomes a new query element, for which a new complementary target pair is generated by:

$$q_i = e_k$$

$$t_i = \{ P \backslash e_k \cup e'_k \} \quad (3)$$



Query	Target	AlnRes	RMSD	TM-score
9-37	41-68	28	0.57	0.8502
9-37	75-102	28	0.51	0.8783
9-37	124-152	28	0.52	0.8702

FIGURE 2 Schematic representation of the final evaluation and scoring step of FiRES. Internal structure similarity in the N-terminal TPR domain of p67phox. A, Dot plot of the local structure alignments generated once the iterations over steps 2 and 3 converge. Orange arrowheads indicate the terminal sites of the gap-extension process (∧ start, and ∨ end positions). B, Local structure alignment pairs that have passed the size filter. The middle position of gapped and ungapped pairs is indicated by yellow and red squares, respectively. C, Clustering by transitivity. Two different clusters are indicated. The first cluster includes pairs a, a', b, c, d', e', f', and g' (red arrows), the second cluster includes pairs h and h' (blue arrows). D, Similar pairs of elements within the first cluster. The length of individual elements along the x-axis is indicated by colored stripes. E, Visualization of the individual elements of the first cluster on the tree-dimensional model (color code as in C). F, Table of results for 1W5M_A

New query-target pairs enter a loop over steps 2 and 3, until no new matched residue pairs are found within a local alignment.

2.4 | Final evaluation step and scoring

Once the number and length of candidate elements remain constant through the iteration of steps 2 and 3, a gap extension process is initialized (Figure 2). The goal of this process is to retrieve pairs of elements that fold into similar three-dimensional structures, but that may be formed by non-sequential SSEs. A gap is extended between two fragment pairs if these are separated by <55 residues in sequence (Figure 2A). Gapped and ungapped pairs over 20 residues length (Figure 2B) undergo a final structure alignment, using CLICK. The TM-score of the CLICK-generated superimpositions is calculated to assess the similarity between candidate pairs of elements. Finally, elements are clustered by transitivity. Elements are considered equivalent upon transitivity if the middle position of any of the elements in a pair is at most two positions away on its sequence representation from the middle position of an element in any other pair (Figure 2B, C). Similar elements grouped by transitivity are displayed with the same initial reference in the output (Figure 2E).

2.5 | Protein repeats datasets

Two databases were selected in order to evaluate the ability of the methods to identify proteins with different types of repeats: MALIDUP for proteins with duplicated domains and RepeatsDB for proteins with multiple tandem repeats. To facilitate comparisons to other sequence and structure-based methods, two data sets were assembled from MALIDUP and RepeatsDB with entries fulfilling the following requirements: (a) at least one of the reference units should be longer than 15 AMino acids; (b) the protein should contain only one type of repeat; and (c) repeat units should not have circular permutations. The first data set contains 137 proteins with duplicated domains from the MALIDUP database.[43] The second data set includes 3522 proteins with short-tandem repeats retrieved from RepeatsDB.[44]

PDB files were obtained from the PDB database using the script get_pdb.py from Rosetta (www.rosettacommons.org). FASTA sequences were obtained from UniProt[45] using the corresponding PDB code and chain.

2.6 | Database of no repeats

A database of proteins without internal sequence and structure similarity was generated using three sequential filters. First, a database of non-symmetrical protein structures was constructed. To this end, a subset of the PDB composed of 28 337 non-redundant chains, as determined by BLASTClust at 30% sequence identity, was evaluated by SymD.[31] From a total of 6088 structures that were considered non-symmetrical (Z-score < 4), 3300 were randomly selected to continue to the next stage. Then, two random groups were assembled with 300 and 3000 proteins to resemble the sizes of MALIDUP and RepeatsDB, respectively. The second filter consisted in excluding proteins that presented repeated sequence signatures according to InterPro.[46] These signatures include annotations from CATH,[47] Gene3D,[48] CDD,[49] PANTHER,[50] Pfam,[51] ProDom,[52] PROSITE,[53] SMART,[54] SUPERFAMILY,[55] and TIGRFAMs.[56] The remaining proteins were evaluated by all algorithms tested in this study (see performance evaluation). Pairwise structure alignments of the predictions made by any of the tested algorithms were performed with TM-align. Predicted repeat pairs with >15 aligned residues and a TM-score higher than 0.5 were visually inspected to evaluate the content, connectivity and orientation of their SSEs. This inspection revealed the presence of 81 proteins with internal structure similarity, which were removed from the final control sets (Suppl. Table S1). Finally, two control sets, integrated by 132 and 2125 proteins with low probability of containing repeated elements, were established as negative controls for MALIDUP and RepeatsDB, respectively.

2.7 | Performance evaluation

The ability of FiRES to identify repeated structural elements was compared to that of three sequence-based and three structure-based methods. Sequence-based algorithms consisted of lalign from FASTA version 36,[20] RADAR[22] version 1.3, and HHrepID[21] version 1.0.0. Structure-based algorithms consisted of ReUPred,[18] Swelfe,[32] and CE-symm[28] version 2.0. In all cases, standalone versions of the software were used. To evaluate the performance of HHrepid, lalign, and FiRES, which output multiple answers, only their best-scoring results were considered. The results of lalign, HHrepID, and FiRES were ranked based on *E*-value, *P*-value, and TM-score, respectively. For sequence-based methods, only predictions made on parts of the protein for which a structural model is available were evaluated. To confirm structural similarity, the predicted units were evaluated with TM-align.[23] Alignments involving at least 50% of each unit and rendering a TM-score higher than 0.5 were considered correct.

The methods were evaluated both at the protein and at the unit level. At the protein level, a prediction was considered true positive if at least half of the predicted units had a correct alignment to at least one other predicted unit. This test evaluates if the methods can differentiate between proteins with and without internal similarity regardless of whether the predicted units correspond to the database definitions or not. At the unit level, a predicted unit was considered true positive if it had a correct alignment with at least one of the reference units as reported by MALIDUP or RepeatsDB. At both levels, the performance was evaluated by their true positive rate (TPR). Additionally, at the protein level, true negative rate (TNR) and accuracy were calculated based on the predictions of the methods in the control sets.

3 | RESULTS

3.1 | Benchmark test

The ability of FiRES, ReUPred, Swelfe, CE-symm, HHrepID, lalign, and RADAR to detect proteins with very divergent duplicated domains

was tested using a subset of MALIDUP. All methods obtained similar TPRs at the protein and at the unit level (Table 2). At both levels, FiRES obtained the highest TPR within the MALIDUP dataset. FiRES correctly identified repeats in 102 out of 137 proteins, which is more than twice the number of proteins identified by RADAR, HHrepID or laling (Table 2). CE-symm, which correctly identified 64 proteins, displayed the second best TPR. Both FiRES and CE-symm rely on structural alignments to detect internal similarity, which, in general, makes them better suited to identify larger repeated units.[28] However, CE-symm detects similar units only when the interface between units is conserved.[28] Swelfe identified only 26 proteins in the MALIDUP dataset and was clearly outperformed by their sequence-based counterparts. ReUPred detected six proteins containing duplicated domains. These six proteins are themselves formed by the repetition of supersecondary structure elements, namely αα- and ββ-hairpins, and βαβ-elements. ReUPred is a method that was specifically designed for the identification and classification of solenoid proteins.[18] Thus, it is not surprising that ReUPred performs poorly on a set, which exclusively includes duplicated domains.

Although protein structure tends to be more conserved than sequence, structural divergence of the units and of the interface between the units constitute a mayor challenge for structure-based repeats identification methods. Collectively, all methods predicted only 110 out of 137 proteins in MALIDUP. The 27 proteins that remained undetected contain very divergent pairs of domains that share <20% identical residues or that render an average TM-score below 0.5 (Suppl. Table S2).

The RepeatsDB dataset tests the ability of the methods to identify proteins containing units that are repeated multiple times. Proteins in RepeatsDB have on average 7.7 repeated units. All algorithms obtained better results at the protein level on the RepeatsDB dataset than on the MALIDUP dataset (Table 2). However, at the unit level the TPR of FiRES (0.18), Swelfe (0.03) and lalign (0.07) was very low, compared to their TPR at the protein level. The TPR at the unit level indirectly evaluates the ability of the different methods to assign boundaries to the units. At this level, CE-symm achieved the highest TPR (0.70), followed by HHrepID (0.49) and RADAR (0.39). FiRES tends to output units that are longer than the reference units in RepeatsDB. ReUPred, which was designed as a classification algorithm for solenoid domains, has a very similar TPR both at the protein and at the unit levels.

At the protein level, FiRES identified 91% of proteins in this set. The TPR of the rest of the methods at the protein level remained below 70% (Table 2). To gain further insights on the strengths and weaknesses of the FiRES algorithm, the results were broken down based on the classification of protein repeats (Table 3). The set of structures within RepeatsDB that was analyzed here included repeats within classes II, III, IV, and V (Table 3). FiRES displayed a high TPR (above 80%) over a broad spectrum of repeats. The most notable exceptions to this trend were sub-categories II.2) α helical coiled coil; III.4) β-trefoil/β hairpins; IV.3) β-trefoil; and V.3) α/β-Beads. In fact, none of the algorithms achieved adequate results for subclasses II.2) α helical coiled coil; III.4) β Trefoil/β Hairpins; and IV.3) β-trefoil. In contrast, some types of repeats turned out to be relatively easy to identify. For instance, almost all methods, apart from ReUPred, achieved TPR above 60% for IV.6) α-Barrel; IV.7) α/β Barrel; and IV.9) α/β Trefoil (Table 3).

The ability of the methods to differentiate proteins with repeated elements from proteins without internal similarity was tested on a new database called *database of no repeats*. This database was curated such that repeats of all types (short, long, tandem, and non-tandem) are filtered out (see Methods). The *database of no repeats* contains a total of 2257 non-redundant structures from the PDB.

As expected, structure-based methods produced a high true negative rate (Table 4). Structure-based methods directly evaluate structure, as opposed to sequence-based methods, for which structure similarity is a prediction. Remarkably, HHrepID, a sequence-based method, accomplished a true negative rate of 0.98 for both control sets (Table 4). In contrast, two thirds of the predictions made by RADAR turned out to be false positive results. Overall, FiRES was the method that most accurately differentiated between proteins with and without internal structure similarity, followed by CE-symm and

TABLE 2 True positive rate of the detection methods

	Structure-based methods				Sequence-based methods		
	FIRES	**ReUPred**	**Swelfe**	**CE-symm**	**HHrepID**	**lalign**	**RADAR**
MALIDUP							
Protein (n = 137)	102 (0.74)	6 (0.04)	26 (0.19)	64 (0.47)	34 (0.25)	39 (0.28)	28 (0.20)
Unit (n = 274)	185 (0.68)	7 (0.03)	55 (0.20)	136 (0.50)	80 (0.29)	65 (0.24)	55 (0.20)
RepeatsDB							
Protein (n = 3522)	3225 (0.92)	669 (0.19)	1524 (0.43)	2421 (0.69)	1704 (0.48)	1608 (0.46)	1748 (0.50)
Unit (n = 26 980)	4852 (0.18)	4960 (0.18)	906 (0.03)	18 755 (0.70)	13 189 (0.49)	1891 (0.07)	10 626 (0.39)

TABLE 3 Number of correctly detected proteins with short tandem repeats in RepeatsDB

	FiRES	ReUPred	Swelfe	CE-symm	HHrepID	lalign	RADAR
II.2 (*n* = 9)	**4 (0.44)**	3 (0.33)	0	0	0	0	0
III.1 (*N* = 321)	**258 (0.80)**	80 (0.25)	113 (0.35)	72 (0.22)	110 (0.34)	89 (0.28)	93 (0.29)
III.2 (*N* = 322)	**308 (0.96)**	83 (0.26)	247 (0.77)	231 (0.72)	190 (0.59)	254 (0.79)	212 (0.66)
III.3 (*N* = 863)	**804 (0.93)**	323 (0.37)	552 (0.64)	747 (0.87)	582 (0.67)	509 (0.59)	566 (0.66)
III.4 (*N* = 49)	**29 (0.59)**	2 (0.04)	14 (0.29)	14 (0.29)	10 (0.20)	14 (0.29)	10 (0.20)
III.5 (*N* = 57)	**54 (0.95)**	2 (0.04)	31 (0.54)	15 (0.26)	20 (0.35)	17 (0.30)	17 (0.30)
IV.1 (*N* = 523)	**471 (0.90)**	47 (0.09)	3 (0.01)	232 (0.44)	10 (0.02)	29 (0.05)	60 (0.11)
IV.2 (*N* = 77)	**68 (0.88)**	4 (0.05)	12 (0.16)	25 (0.32)	14 (0.18)	8 (0.10)	18 (0.23)
IV.3 (*N* = 24)	10 (0.42)	0	9 (0.38)	**13 (0.54)**	0	0	0
IV.4 (*N* = 780)	**755 (0.97)**	108 (0.14)	360 (0.46)	712 (0.91)	446 (0.57)	415 (0.53)	478 (0.61)
IV.5 (*N* = 177)	**176 (0.99)**	2 (0.01)	68 (0.38)	**176 (0.99)**	175 (0.99)	135 (0.76)	151 (0.85)
IV.6 (*N* = 5)	**4 (0.80)**	0	**4 (0.80)**	**4 (0.80)**	**4 (0.80)**	**4 (0.80)**	**4 (0.80)**
IV.7 (*N* = 5)	**5 (1.00)**	0	3 (0.60)	**5 (1.00)**	3 (0.60)	3 (0.60)	3 (0.60)
IV.8 (*N* = 102)	**100 (0.98)**	6 (0.06)	35 (0.34)	79 (0.77)	47 (0.46)	58 (0.57)	50 (0.49)
IV.9 (*N* = 15)	**14 (0.93)**	0	10 (0.67)	11 (0.73)	13 (0.87)	9 (0.60)	9 (0.60)
IV.10 (*N* = 45)	38 (0.84)	4 (0.09)	0	**45 (1.00)**	15 (0.33)	6 (0.13)	7 (0.16)
V.1 (*N* = 13)	**13 (1.00)**	2 (0.15)	9 (0.70)	10 (0.77)	8 (0.62)	9 (0.69)	7 (0.54)
V.2 (*N* = 37)	**32 (0.87)**	0	24 (0.65)	5 (0.14)	28 (0.76)	10 (0.27)	30 (0.81)
V.3 (*N* = 14)	6 (0.43)	0	2 (0.14)	**11 (0.79)**	2 (0.14)	2 (0.14)	3 (0.21)
V.4 (*N* = 41)	**35 (0.85)**	1 (0.02)	17 (0.41)	8 (0.20)	8 (0.20)	19 (0.46)	15 (0.37)
V.5 (*N* = 43)	**41 (0.95)**	2 (0.05)	11 (0.26)	6 (0.14)	19 (0.44)	18 (0.42)	15 (0.35)

Note: The highest true positive rate for each protein fold is highlighted in bold. II.1) α helical coiled coil; III.1) Solenoid; III.2) α/β Solenoid; III.3) α-Solenoid; III.4) β Trefoil/β Hairpins; III.5) Anti-parallel β Layer/β Hairpins. IV.1) TIM-Barrel; IV.2) β-Barrel/β-Hairpins; IV.3) β-Trefoil; IV.4) β-Propeller; IV.5) α/β Prism; IV.6) α-Barrel; IV.7) α/β Barrel; IV.8) α/β Propeller; IV.9) α/β Trefoil; IV.10) Aligned prism. V.1) α-Beads; V.2) β-Beads; V.3) α/β-Beads; V.4) β Sandwich beads; V.5) α/β Sandwich.
Abbreviation: *N*, number of cases.

TABLE 4 True negative rate of the detection methods

	Structure-based methods				Sequence-based methods		
	FIRES	ReUPred	Swelfe	CE-symm	HHrepID	lalign	RADAR
Control 1 [*N* = 132]	TN: 129 (97.7%)	TN: 128 (99.7%)	TN: 132 (100%)	TN: 131 (99.2%)	TN: 129 (97.7%)	TN: 113 (85.6%)	TN: 48 (36.3%)
Control 2 [*N* = 2125]	TN: 2009 (94.5%)	TN: 2070 (97.4%)	TN: 2125 (100%)	TN: 2112 (99.4%)	TN: 2078 (97.7%)	TN: 1749 (82.3%)	TN: 867 (40%)

Note: Negative control sets for MALIDUP (control 1) and RepeatsDB (control 2).

HHrepID (Table 5). FiRES is a powerful method for the identification of proteins that contain domain-size or subdomain-size similar structural elements.

3.2 | Detecting similar elements with very low sequence identity

During the construction of the *database of no repeats*, the algorithms identified a total of 81 proteins with internal structure similarity. Internal similarity within these proteins was not documented in InterPro. More than half of these cases were identified only by FiRES, whereas 21 were identified by FiRES and another method. Only 16 cases were identified by a method different from FiRES (Suppl. Table S1). From the 44 results that were exclusive to FiRES, two examples were selected to illustrate the use of FiRES to detect hidden evolutionary relationships between structurally similar elements (Figures 3 and 4 and Suppl. Methods). In both examples, the discrimination between homology and analogy required a combination of structure- and sequence-based methods.

	Structure-based methods				Sequence-based methods		
	FIRES	ReUPred	Swelfe	CE-symm	HHrepID	lalign	RADAR
Benchmark 1 [*N* = 269]	0.86	0.50	0.59	0.72	0.61	0.57	0.28
Benchmark 2 [*N* = 5647]	0.93	0.49	0.65	0.80	0.67	0.59	0.46

Note: Benchmark 1: union of MALIDUP and control set 1; Benchmark 2: union of RepeatsDB and control set 2.

TABLE 5 Accuracy of the detection methods at the protein level

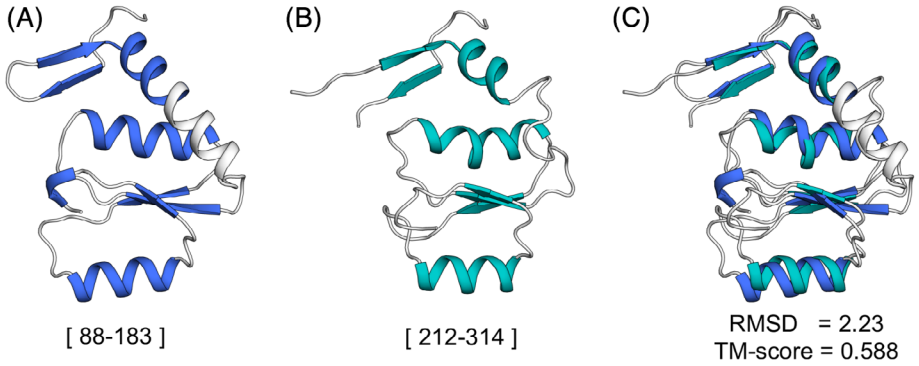


FIGURE 3 Structurally similar elements identified by FiRES in D1GLU5 from *Streptomyces lydicus* (PDB code: 4GR4_A). The ribbon representation of elements 1 (blue) and 2 (cyan) is shown. A, Element 1 found within the central AMP-binding domain. B, Element 2 also located within the central AMP-binding domain. C, Structural alignment of elements 1 and 2 [Color figure can be viewed at wileyonlinelibrary.com]

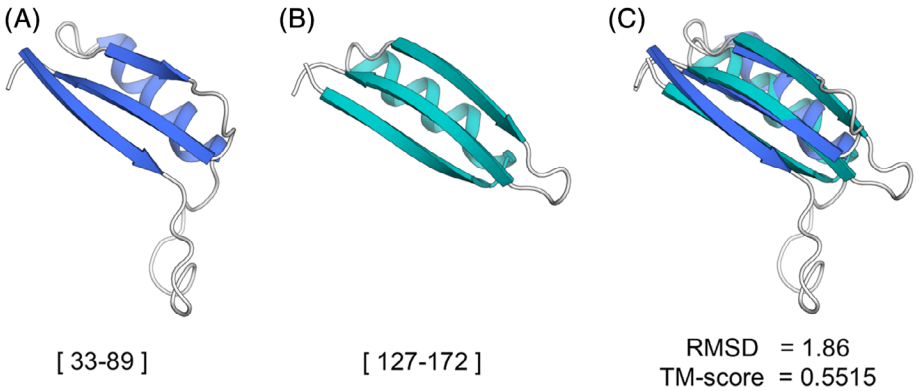


FIGURE 4 Structurally similar elements identified by FiRES in a P-loop NTPase domain (PDB code: 3E3X_A). The ribbon representation of elements 1 (blue) and 2 (cyan) is shown. A, Element 1. B, Element 2. C, Structural alignment of elements 1 and 2 [Color figure can be viewed at wileyonlinelibrary.com]

3.2.1 | Non-ribosomal peptide synthetase (D1GLU5)

The non-ribosomal peptide synthetase from *Streptomyces lydicus* is formed by an N-terminal MbtH domain (PF03621), a central AMP-binding domain (PF00501) and a C-terminal AMP-binding_C domain (PF13193). FiRES identified a repeated motif within the central AMP-binding domain, with a TM-score of 0.59 (Figure 3). However, between these two elements there are only 14% identical residues. These elements were analyzed using sequence-based methods, which confirmed sequence relationships between element 1 and element 2 (Supp. Methods). Put together, these arguments suggest that the AMP-binding domain originated from a duplicated motif.

3.2.2 | Elongation factor Tu GTP binding domain of BipA (Q9L5X8)

FiRES detected a repeated element within the GTP-binding domain (IPR005225) of the protein BipA from *Vibrio parahaemolyticus*. This domain belongs to the P-loop NTPase superfamily. The structural alignment of these elements produces an RMSD of 1.68 Å and a TM-score of 0.55. The pair of similar elements within the GTP-binding domain shares 10% identical residues, according to their structure-based sequence alignment (Figure 4). However, sequence-based analyses showed evidence to support that these structural motifs have a common evolutionary origin (Supp. Methods).

4 | DISCUSSION

We tested FiRES and six other repeat identification methods on two hand-curated databases: MALIDUP, for duplicated domains, and RepeatsDB for short tandem repeats. MALIDUP and RepeatsDB contain proteins with very divergent repeated units. In both cases, FiRES obtained outstanding results at the protein level in terms of TPR, TNR and accuracy. Furthermore, FiRES demonstrated to provide more consistent results than lalign, RADAR, CE-symm, HHrepID, ReUPred, and Swelfe to detect proteins with different types of repeats. FiRES was designed as a tool to detect internal structure similarity in proteins

where these similarities remain unnoticed. FiRES is not intended for classification of repetitive elements. However, the self-structure comparison strategy employed by FiRES may lead to the development of new methods for protein repeats classification.

Three key features were implemented on FiRES to produce highly accurate results. First, the comparison of non-sequential structural elements increases the sensibility of FiRES because it enables the identification of incomplete units, as well as of units with insertions, deletions, circular permutations and other types of fold change. Second, the iteration of the identification process makes it possible to detect multiple types of repeated elements within the same protein structure. Third, the last step of the algorithm is a time-consuming exhaustive structural comparison of each candidate pair, which renders the FiRES algorithm highly specific.

The identification of structurally similar units that lack sequence similarity can help elucidate remote homologous relationships. Here, we presented two examples where FiRES, in combination with state-of-the-art sequence similarity detection methods, provides new insights into the evolution of protein domain structure. Besides being of interest for evolutionary studies, protein repeats have been employed to construct well-folded and stable protein chimeras.[7,57] Individual elements contribute to the functional properties of the chimera protein, making it possible to target specific biological activities throughout a design process.[58] Computational tools allowing the identification of internal similarity within a protein structure can aid the design of new protein functions by recombination of specific domains or subdomains. Consequently, FiRES may have a positive impact on many fields of protein science.

ACKNOWLEDGMENTS

M.A. acknowledges to Programa de Apoyo a Proyectos de Investigación e Innovación Tecnológica to Dirección General de Cómputo y de Tecnologías de Información y Comunicación from the Universidad Nacional Autónoma de México for supporting this study through the grants PAPIIT-DGAPA-IN213320 and LANCAD-UNAM-DGTIC-320. G.C. acknowledges the computing facilities at the Cell Physiology Institute of the Universidad Nacional Autónoma de México. C.A.-C. was supported by a Fulbright-García Robles Fellowship.

CONFLICT OF INTERESTS

The authors declare that they have no conflicts of interest with the content of this article.

ORCID

Claudia Alvarez-Carreño https://orcid.org/0000-0002-1827-8946
Marcelino Arciniega https://orcid.org/0000-0002-7526-6941

REFERENCES

1. Andrade MA, Perez-Iratxeta C, Ponting CP. Protein repeats: structures, functions, and evolution. *J Struct Biol*. 2001;134(2–3):117-131. https://doi.org/10.1006/jsbi.2001.4392.
2. Marcotte EM, Pellegrini M, Yeates TO, Eisenberg D. A census of protein repeats. *J Mol Biol*. 1999;293(1):151-160. https://doi.org/10.1006/jmbi.1999.3136.
3. Rajathei DM, Selvaraj S. Analysis of sequence repeats of proteins in the PDB. *Comput Biol Chem*. 2013;47:156-166. https://doi.org/10.1016/j.compbiolchem.2013.09.001.
4. Jorda J, Xue B, Uversky VN, Kajava AV. Protein tandem repeats—the more perfect, the less structured. *FEBS J*. 2010;277(12):2673-2682. https://doi.org/10.1111/j.1742-4658.2010.07684.x.
5. Söding J, Lupas AN. More than the sum of their parts: on the evolution of proteins from peptides. *Bioessays*. 2003;25(9):837-846. https://doi.org/10.1002/bies.10321.
6. Broom A, Doxey AC, Lobsanov YD, et al. Modular evolution and the origins of symmetry: reconstruction of a three-fold symmetric globular protein. *Structure*. 2012;20(1):161-171. https://doi.org/10.1016/j.str.2011.10.021.
7. Main ER, Lowe AR, Mochrie SG, Jackson SE, Regan L. A recurring theme in protein engineering: the design, stability and folding of repeat proteins. *Curr Opin Struct Biol*. 2005;15(4):464-471. https://doi.org/10.1016/j.sbi.2005.07.003.
8. Hrabe T, Godzik A. ConSole: using modularity of contact maps to locate solenoid domains in protein structures. *BMC Bioinformatics*. 2014;15(119). https://doi.org/10.1186/1471-2105-15-119.
9. Fournier D, Palidwor GA, Shcherbinin S, et al. Functional and genomic analyses of alpha-solenoid proteins. *PLoS One*. 2013;8(11), e79894. https://doi.org/10.1371/journal.pone.0079894.
10. Bornberg-Bauer E, Beaussart F, Kummerfeld SK, Teichmann SA, Weiner J. The evolution of domain arrangements in proteins and interaction networks. *Cell Mol Life Sci*. 2005;62(4):435-445. https://doi.org/10.1007/s00018-004-4416-1.
11. Moore AD, Björklund ÅK, Ekman D, Bornberg-Bauer E, Elofsson A. Arrangements in the modular evolution of proteins. *Trends Biochem Sci*. 2008;33(9):444-451. https://doi.org/10.1016/j.tibs.2008.05.008.
12. Nacher JC, Hayashida M, Akutsu T. The role of internal duplication in the evolution of multi-domain proteins. *Biosystems*. 2010;101(2):127-135. https://doi.org/10.1016/j.biosystems.2010.05.005.
13. Grishin NV. Fold change in evolution of protein structures. *J Struct Biol*. 2001;134(2–3):167-185. https://doi.org/10.1006/jsbi.2001.4335.
14. Russell RB, Ponting CP. Protein fold irregularities that hinder sequence analysis. *Curr Opin Struct Biol*. 1998;8(3):364-371. https://doi.org/10.1016/S0959-440X(98)80071-7.
15. Illergård K, Ardell DH, Elofsson A. Structure is three to ten times more conserved than sequence—a study of structural response in protein cores. *Proteins Struct Funct Bioinform*. 2009;77(3):499-508. https://doi.org/10.1002/prot.22458.
16. Chothia C, Lesk AM. The relation between the divergence of sequence and structure in proteins. *EMBO J*. 1986;5(4):823-826. https://doi.org/10.1002/j.1460-2075.1986.tb04288.x.
17. Uliel S, Fliess A, Amir A, Unger R. A simple algorithm for detecting circular permutations in proteins. *Bioinformatics*. 1999;15(11):930-936. https://doi.org/10.1093/bioinformatics/15.11.930.
18. Hirsh L, Piovesan D, Paladin L, Tosatto SCE. Identification of repetitive units in protein structures with ReUPred. *Amino Acids*. 2016;48(6):1391-1400. https://doi.org/10.1007/s00726-016-2187-2.
19. Nguyen MN, Madhusudhan MS. Biological insights from topology independent comparison of protein 3D structures. *Nucleic Acids Res*. 2011;39(14), e94. https://doi.org/10.1093/nar/gkr348.
20. Pearson WR, Lipmant DJ. Improved tools for biological sequence comparison. *Proc Natl Acad Sci USA*. 1988;85(April):2444-2448.
21. Biegert A, Söding J. De novo identification of highly diverged protein repeats by probabilistic consistency. *Bioinformatics*. 2008;24(6):807-814. https://doi.org/10.1093/bioinformatics/btn039.
22. Heger A, Holm L. Rapid automatic detection and alignment of repeats in protein sequences. *Proteins Struct Funct Genet*. 2000;41(2):224-237.

23. Karpenahalli MR, Lupas AN, Söding J. TPRpred: a tool for prediction of TPR-, PPR- and SEL1-like repeats from protein sequences. *BMC Bioinformatics*. 2007;8(2). https://doi.org/10.1186/1471-2105-8-2.
24. Szklarczyk R, Heringa J. Tracking repeats using significance and transitivity. *Bioinformatics*. 2004;20(suppl 1):311-317. https://doi.org/10.1093/bioinformatics/bth911.
25. Pellegrini M. Tandem repeats in proteins: prediction algorithms and biological role. *Front Bioeng Biotechnol*. 2015;3:143 (September). https://doi.org/10.3389/fbioe.2015.00143.
26. Hirsh L, Paladin L, Piovesan D, Tosatto SCE. RepeatsDB-lite: a web server for unit annotation of tandem repeat proteins. *Nucleic Acids Res*. 2018;46(W1):W402-W407. https://doi.org/10.1093/nar/gky360.
27. Kao HY, Shih TH, Pai TW, Da Lu M, Hsu HH. A comprehensive system for identifying internal repeat substructures of proteins. Paper presented at: *CISIS* 2010—4th International Conference on Complex, Intelligent and Software Intensive Systems; 2010:689–693. doi: https://doi.org/10.1109/CISIS.2010.92
28. Bliven SE, Lafita A, Rose PW, Capitani G, Prlić A, Bourne PE. Analyzing the symmetrical arrangement of structural repeats in proteins with CE-Symm. *PLoS Comput Biol*. 2019;15(4):e1006842. https://doi.org/10.1371/journal.pcbi.1006842.
29. Murray KB, Taylor WR, Thornton JM. Toward the detection and validation of repeats in protein structure. *Proteins Struct Funct Bioinforma*. 2004;380(June):365-380. https://doi.org/10.1002/prot.20202.
30. Guerler A, Knapp E-W. Novel protein folds and their nonsequential structural analogs. *Protein Sci*. 2008;17(8):1374-1382. https://doi.org/10.1110/ps.035469.108.
31. Jha A, Flurchick KM, Bikdash M, Kc DB. Parallel-SymD: a parallel approach to detect internal symmetry in protein domains. *Biomed Res Int*. 2016;2016:4628592.
32. Abraham A, Rocha EPC, Pothier J. Swelfe: a detector of internal repeats in sequences and structures. *Bioinformatics*. 2008;24(13):1536-1537. https://doi.org/10.1093/bioinformatics/btn234.
33. Sabarinathan R, Basu R, Sekar K. ProSTRIP: a method to find similar structural repeats in three-dimensional protein structures. *Comput Biol Chem*. 2010;34(2):126-130. https://doi.org/10.1016/j.compbiolchem.2010.03.006.
34. Do Viet P, Roche DB, Kajava AV. TAPO: a combined method for the identification of tandem repeats in protein structures. *FEBS Lett*. 2015;589(19):2611-2619. https://doi.org/10.1016/j.febslet.2015.08.025.
35. Russell RB, Saqi MAS, Sayle RA, Bates PA, Sternberg MJE. Recognition of analogous and homologous protein folds: analysis of sequence and structure conservation. *J Mol Biol*. 1997;269:423-439. https://doi.org/10.1006/jmbi.1997.1019.
36. Krishna SS, Grishin NV. Structurally analogous proteins do exist! *Structure*. 2004;12(7):1125-1127. https://doi.org/10.1016/j.str.2004.06.004.
37. Jung J, Lee B. Circularly permuted proteins in the protein structure database. *Protein Sci*. 2001;10:1881-1886. https://doi.org/10.1101/ps.05801.Results.
38. Salem GM, Hutchinson EG, Orengo CA, Thornton JM. Correlation of observed fold frequency with the occurrence of local structural motifs. *J Mol Biol*. 1999;287(5):969-981. https://doi.org/10.1006/jmbi.1999.2642.
39. Cheng H, Kim B, Grishin NV. Discrimination between distant homologs and structural analogs: lessons from manually constructed, reliable data sets. *J Mol Biol*. 2008;377(4):1265-1278. https://doi.org/10.1016/j.jmb.2007.12.076.
40. Zhang Y, Skolnick J. Scoring function for automated assessment of protein structure template quality. *Proteins Struct Funct Genet*. 2004;57(4):702-710. https://doi.org/10.1002/prot.20264.
41. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*. 1983;22(12):2577-2637. https://doi.org/10.1002/bip.360221211.
42. Smith T, Waterman MS. Identification of common molecular subsequences. *J Mol Evol*. 1981;147:195-197.
43. Cheng H, Kim B, Grishin NV. MALIDUP: a database of manually constructed structure alignments for duplicated domain pairs. *Proteins Struct Funct Bioinform*. 2007;70(4):1162-1166. https://doi.org/10.1002/prot.21783.
44. Paladin L, Hirsh L, Piovesan D, et al. RepeatsDB 2.0: improved annotation, classification, search and visualization of repeat protein structures. 2017;45(November 2016):308-312. https://doi.org/10.1093/nar/gkw1136.
45. Bateman A. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res*. 2019;47(D1):D506-D515. https://doi.org/10.1093/nar/gky1049.
46. Mitchell AL, Attwood TK, Babbitt PC, et al. InterPro in 2019: improving coverage, classification and access to protein sequence annotations. *Nucleic Acids Res*. 2019;47(November 2018):351-360. https://doi.org/10.1093/nar/gky1100.
47. Dawson NL, Lewis TE, Das S, et al. CATH: an expanded resource to predict protein function through structure and sequence. *Nucleic Acids Res*. 2017;45(Database issue):D289-D295. https://doi.org/10.1093/nar/gkw1098.
48. Lewis TE, Sillitoe I, Dawson N, et al. Gene3D: extensive prediction of globular domains in proteins. *Nucleic Acids Res*. 2018;46(Database issue):D435-D439. https://doi.org/10.1093/nar/gkx1069.
49. Marchler-bauer A, Bo Y, Han L, et al. CDD/SPARCLE: functional classification of proteins via subfamily domain architectures. *Nucleic Acids Res*. 2017;45(Database issue):D200-D203. https://doi.org/10.1093/nar/gkw1129.
50. Mi H, Huang X, Muruganujan A, et al. PANTHER version 11: expanded annotation data from gene ontology and Reactome pathways, and data analysis tool enhancements. *Nucleic Acids Res*. 2017;45(Database issue):D183-D189. https://doi.org/10.1093/nar/gkw1138.
51. El-gebali S, Mistry J, Bateman A, et al. The Pfam protein families database in 2019. *Nucleic Acids Res*. 2019;47(October 2018):D427-D432. https://doi.org/10.1093/nar/gky995.
52. Servant F, Bru C, Peyruc D, Kahn D. ProDom: automated clustering of homologous domains. *Brief Bioinform*. 2002;3(3):246-251.
53. Sigrist CJA, de Castro E, Cerutti L, et al. New and continuing developments at PROSITE. *Nucleic Acids Res*. 2013;41(November 2012):D344-D347. https://doi.org/10.1093/nar/gks1067.
54. Letunic I, Bork P. 20 years of the SMART protein domain annotation. *Nucleic Acids Res*. 2018;46(October 2017):493-496. https://doi.org/10.1093/nar/gkx922.
55. Wilson D, Pethica R, Zhou Y, et al. SUPERFAMILY—sophisticated comparative genomics, data mining, visualization and phylogeny. *Nucleic Acids Res*. 2009;37(November 2008):D380-D386. https://doi.org/10.1093/nar/gkn762.
56. Haft DH, Loftus BJ, Richardson DL, et al. TIGRFAMs: a protein family resource for the functional identification of proteins. *Nucleic Acids Res*. 2001;29(1):41-43.
57. Eisenbeis S, Höcker B. Evolutionary mechanism as a template for protein engineering. *J Pept Sci*. 2010;16(10):538-544. https://doi.org/10.1002/psc.1233.
58. Rico JAF, Höcker B. Design of chimeric proteins by combination of subdomain-sized fragments. *Methods Enzymol*. 2013;523:389-405. https://doi.org/10.1016/B978-0-12-394292-0.00018-7.
59. Eddy SR. Accelerated profile HMM searches. *PLoS Comput Biol*. 2011;7(10):e1002195. https://doi.org/10.1371/journal.pcbi.1002195.
60. Marsella L, Sirocco F, Trovato A, Seno F, Tosatto SCE. REPETITA: detection and discrimination of the periodicity of protein solenoid repeats by discrete Fourier transform. *Bioinformatics*. 2009;25:289-295. https://doi.org/10.1093/bioinformatics/btp232.

61. Walsh I, Sirocco FG, Minervini G, Di Domenico T, Ferrari C, Tosatto SCE. RAPHAEL: recognition, periodicity and insertion assignment of solenoid protein structures. *Bioinformatics*. 2012;28(24):3257-3264. https://doi.org/10.1093/bioinformatics/bts550.
62. Heringa J, Argos P. A method to recognize distant repeats in protein sequences. *Proteins Struct Funct Bioinform*. 1993;17(4):391-411. https://doi.org/10.1002/prot.340170407.
63. Pagès G, Grudinin S. DeepSymmetry: using 3D convolutional networks for identification of tandem repeats and internal symmetries in protein structures. *Bioinformatics*. 2019;35(24):5113-5120. https://doi.org/10.1093/bioinformatics/btz454.

SUPPORTING INFORMATION

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

How to cite this article: Alvarez-Carreño C, Coello G, Arciniega M. FiRES: A computational method for the de novo identification of internal structure similarity in proteins. *Proteins*. 2020;88:1169–1179. https://doi.org/10.1002/prot.25886