PLOS ONE

RESEARCH ARTICLE

Extending Protein Domain Boundary Predictors to Detect Discontinuous Domains

Zhidong Xue[1]*, Richard Jang[1,2], Brandon Govindarajoo[2], Yichu Huang[1], Yan Wang[3]*

1 School of Software Engineering, Huazhong University of Science and Technology, Wuhan, Hubei, 430074, China, **2** Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI, 48109, United States of America, **3** School of Life Science and Technology, Huazhong University of Science and Technology, Wuhan, Hubei, 430074, China

* zdxue@hust.edu.cn (ZX); yanw@hust.edu.cn (YW)

OPEN ACCESS

Citation: Xue Z, Jang R, Govindarajoo B, Huang Y, Wang Y (2015) Extending Protein Domain Boundary Predictors to Detect Discontinuous Domains. PLoS ONE 10(10): e0141541. doi:10.1371/journal.pone.0141541

Editor: Ramanathan Sowdhamini, NCBS-TIFR, INDIA

Received: July 3, 2015

Accepted: October 10, 2015

Published: October 26, 2015

Data Availability Statement: All datasets are available through GitHub at http://github.com/xuezhidong/DomEx.

Funding: The project is supported in part by the National Natural Science Foundation of China (30700162, 61073095), the Fundamental Research Funds for the Central Universities of China (HUST:2014TS138 and HUST2015QN101) and the China Postdoctoral Science Foundation (2014M552043). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Abstract

A variety of protein domain predictors were developed to predict protein domain boundaries in recent years, but most of them cannot predict discontinuous domains. Considering nearly 40% of multidomain proteins contain one or more discontinuous domains, we have developed DomEx to enable domain boundary predictors to detect discontinuous domains by assembling the continuous domain segments. Discontinuous domains are predicted by matching the sequence profile of concatenated continuous domain segments with the profiles from a single-domain library derived from SCOP and CATH, and Pfam. Then the matches are filtered by similarity to library templates, a symmetric index score and a profile-profile alignment score. DomEx recalled 32.3% discontinuous domains with 86.5% precision when tested on 97 non-homologous protein chains containing 58 continuous and 99 discontinuous domains, in which the predicted domain segments are within ±20 residues of the boundary definitions in CATH 3.5. Compared with our recently developed predictor, ThreaDom, which is the state-of-the-art tool to detect discontinuous-domains, DomEx recalled 26.7% discontinuous domains with 72.7% precision in a benchmark with 29 discontinuous-domain chains, where ThreaDom failed to predict any discontinuous domains. Furthermore, combined with ThreaDom, the method ranked number one among 10 predictors. The source code and datasets are available at https://github.com/xuezhidong/DomEx.

Introduction

Proteins consist of one or several stable, compact, and autonomously folding substructures, which are referred to as domains. The identification of protein domains plays an important role in determining protein structures by experimental methods including Nuclear Magnetic Resonance (NMR) and X-ray crystallography[1,2]. Meanwhile, it is also a preliminary step in computational methods of protein structure prediction [3–5]. Moreover, detailed knowledge of domains is essential to advancing our understanding of protein function and evolution [6,7].

Although protein domains usually have a single continuous segment of protein chain, there are still many domains formed from two or more nonsequential segments, which are called

Competing Interests: The authors have declared that no competing interests exist.

"discontinuous domains"[8]. For example, 28, 279 out of the 181,356 domains (~15%) are discontinuous in the CATH3.5 library[9,10] and nearly 16,761 proteins (~18%) have at least one discontinuous domain based on the domain classifications by DomainParser2[11] in the PDB library.

Over the last three decades, a number of methods have been developed to identify protein domains, which are roughly classified into two categories according to their input data: structure or sequence. The structure-based methods can accurately identify continuous and discontinuous domains from the atomic coordinates of proteins [8,11–15]. The sequence-based methods predicting domains from sequences alone have obtained some progress in predicting continuous domains. Even including tertiary structure libraries like CATH[9,10], SCOP[16], SMART[17] that provide domain partitions of continuous and discontinuous domains, few sequence-based methods can predict the discontinuous domains. Then the discontinuous domain prediction is an open and challenging problem.

An accurate discontinuous domain prediction includes predicting the accurate domain boundaries and the number of segments within one discontinuous domain. Currently, the sequence-based methods mainly focused on domain number and boundary prediction. DGS [18] guesses the domain number and further infers domain boundaries by predicting the size and the segment number of domains. DomCut[19] predicts inter-domain linker regions based solely on amino acid sequence composition information. Pfam[20–22], EVEREST[23,24] ADDA[25], and FiefDom[26] focus on domain boundaries prediction based on homologous alignments. CHOPnet[27],Dompro[28], DomNet[29], PPRODO[30], DROP[31] and DOBO [32] use different machine learning methods to identify domain boundaries.

Some methods such as SnapDRAGON[33], RosettaDom[34] and OPUS-DOM[35] first constructed a 3D model and then extracted domain boundaries with structure-based domain partition tools such as DAIL[13], PDP[12] and DomainParser [11]. Although these methods can detect discontinuous domains, the success of the domain assignments relies on the correctness of the predicted models, which are applicable only to small proteins[4]. DomainDiscovery [36] was developed to predict discontinuous domains mainly based on the predicted inter-residue contact interaction values, while the accuracy of long-range contacts prediction from the sequence alone is very low[37]. ThreaDom[38] uses a template cluster method to detect discontinuous domain based on the meta-server threading program LOMETS[39]. However, it will fail to identify discontinuous domains if there is no available template deposited in the PDB. And it didn't use the domain information from the sequence domain libraries, such as Pfam [20–22] et al.

In this work, we present a new strategy, DomEx, to enable continuous domain boundary predictors to predict discontinuous domains based on the sequence segment assembly. A template similarity score, symmetric index score and a profile-profile alignment score were developed to detect the discontinuous domains through a comprehensive single-domain library collecting not only from the structure domain databases (SCOP[16] and CATH[9,10]) but also from the sequence domain database, Pfam-A[20–22]. We trained and tested this method on various large-scale datasets and further tested its effectiveness in extending protein domain boundary predictors to detect discontinuous domains through combining DomEX with several domain predictors.

Methods and Materials

Domain library

DomEx detects discontinuous-domains by comparing candidate domain sequences with the sequence of known protein domains in the domain library. The DomEx domain library is

constructed from three protein domain databases: CATH3.5 [40], SCOP1.75 [41] and Pfam-A. CATH and SCOP are 3D structure databases categorized semi-manually using structural alignment tools. Pfam-A is a curated sequence domain database derived from UniProtKB [42] and containing profile hidden Markov models for sequence search. A pairwise sequence identity cutoff ($\geq$90%) was used to filter out redundant entries from the initial DomEx library, resulting in 5,308,138 domains, where 24,368 domains are from CATH and SCOP and 55,283,770 from Pfam-A. Since the majority of the domains (~99%) are from the Pfam sequence database, the coverage is increased significantly over the structure-only library in ThreaDom.

Procedure to detect discontinuous domain

DomEx makes three assumptions: (a) Homologous protein domains can be detected by sequence-based profile-profile alignments; (b) Homologous domain pairs have approximately similar length; (c) The coverage and sequence similarity between the different segment pairs in the same homologous domain pairs are usually symmetric. Assumptions (a) and (b) are straightforward. For assumption (c), let some discontinuous domain A has two segments (A_1 and A_2) from N- to C-terminal, and it has a homologous domain partner B. The position of the last residue of segment A_1 is marked as n, then the alignment pair (A-B) could be divided into two segment pairs (A_1-B_1 and A_2-B_2) at the position between n and n+1. The coverage and sequence similarity of the segment pair A_1-B_1 should be close to that of the segment pair A_2-B_2. In other words, the coverage and sequence similarity between the two segment pairs are symmetric at the separated point.

Template Similarity Score, Symmetry Index score and Profile-Profile Alignment Score are designed to detect the discontinuous domain. DomEx uses a five-step procedure to assemble and detect the discontinuous domains:

Step 1: Predict the domain/segment boundary positions of a query protein sequence using ThreaDom [38] (or any other domain prediction software).

Step 2: Take all possible nonconsecutive segment pairs as putative discontinuous domains by concatenation.

Step 3: Search the DomEx domain library for hits to homologues templates of the putative domain sequence through a two stage profile alignment with PSI-BLAST.

Step 4: Evaluate the domain assemble score by TS-score, SI and length similarity.

Step 5: Filter the templates from step 4 that are found in Pfam using the profile-profile alignment (PPA) score.

Step 6: Detect conflicts and report the final result.

The entire flowchart of DomEx is shown in Fig 1. The pseudo code of the main procedure of DomEx is shown in Fig 2. The input consists of the query sequence X, the predicted boundaries B, and the segment number N. DomEx outputs the final detection result by calling FindHit as shown in Fig 3.

An assembled domain sequence Q_i is predicted as discontinuous, if there is at least one hit T_j with length error $e(Q_i, T_j) < 0.2$, Ts-score$(Q_i, T_j) > T_{TS}$ and $SI(Q_i, T_j) < T_{SI}$. The parameters T_{TS} and T_{SI} are the cutoffs of TS-score and Symmetric Index, and they satisfy the constraint function $T_{SI} = f(T_{TS}, b)$. This function can be decided by maximizing the Matthews Correlation Coefficient (MCC) value in the training datasets (see below).

If there are multiple candidates that pass through the decision tree, then the candidate with the lowest PPA-score is selected.

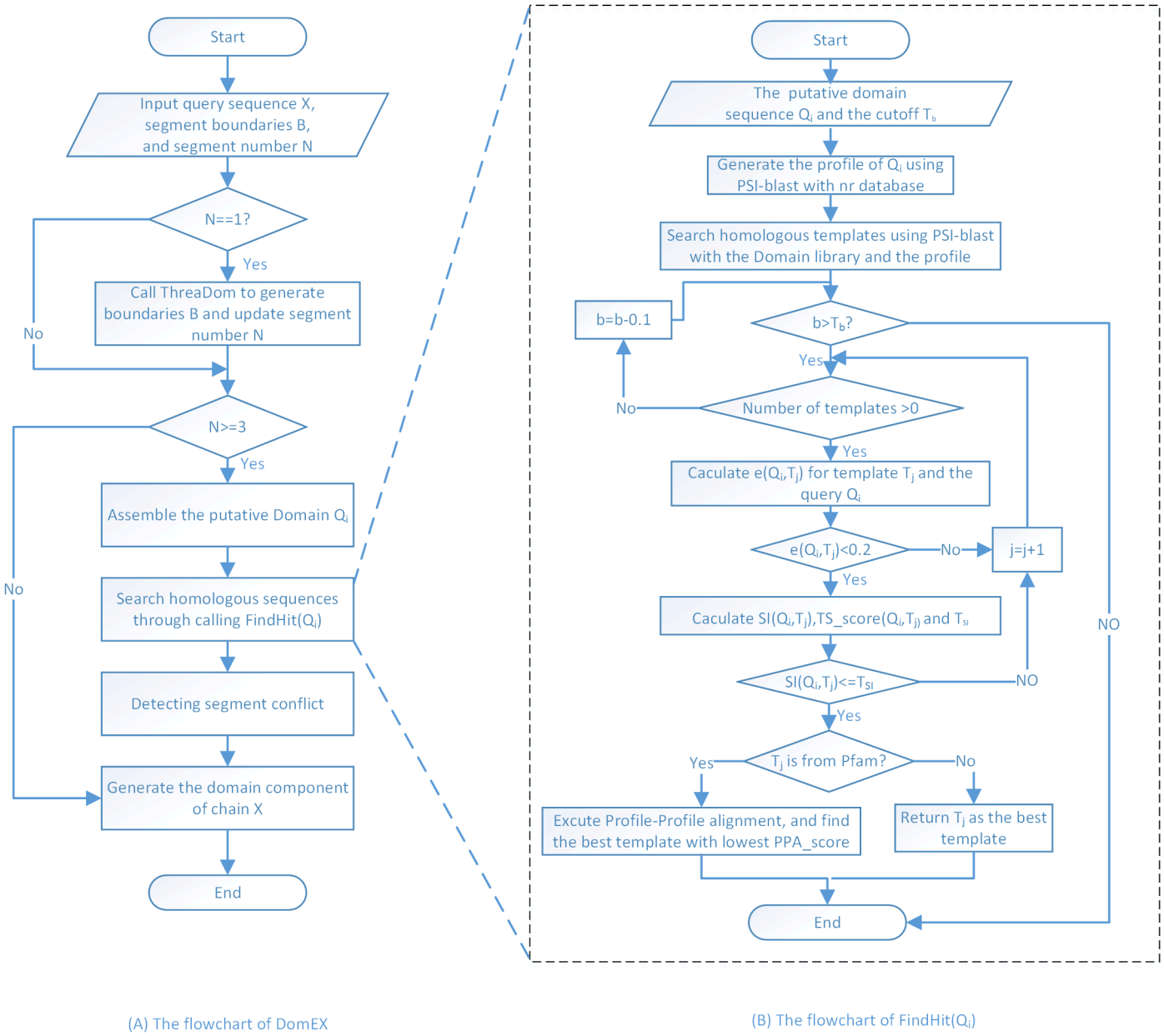

Fig 1. The flowchart of DomEx.

doi:10.1371/journal.pone.0141541.g001

Template similarity score and symmetry index

Consider a protein chain X which is divided into n+1 segments (S_1, . . ., S_{n+1}) by n boundary bars (B_1, . . ., B_n), which can be predicted by ThreaDom [38] or any other domain prediction tool. We select two nonadjacent segments S_p and S_q ($p \neq q$), and assemble them into a new putative domain $Q = (S_p \cup S_q)$. To examine the possibility that Q is a discontinuous domain, we first search the DomEx domain library for hits to some homologous template T of the putative domain sequence through a two stages profile alignment with PSI-BLAST[43], and then we use the Template Similarity score (TS-score) and Symmetric Index (SI) to screen the PSI-BLAST hits.

Algorithm 1 The Pseudo Code of DomEx

INPUT: The Query Sequence X;Protein Segment Boundaries B;Segment Number N.

OUTPUT: Protein Domain Prediction Results Y:

```
procedure DomEx(X, B, N)
    Y ← ∅
    Q ← ∅
    if N == 1 then
        Call ThreaDom to generate Boundaries B and Segment Number N
    end if
    if N ≥ 3 then
        Assemble the Putative Domain Q_i,i = 1,...,n.
        Q ← {Q_1, ..., Q_n}
    end if
    for all Q_i ∈ Q do
        HIT ← FindHit(Q_i)
    end for
    Detection segment conflict according to HIT
    Y ← generate the domain component of chain X
    return Y
end procedure
```

Fig 2. The Pseudo code of DomEx.

doi:10.1371/journal.pone.0141541.g002

The TS-score between the query and the template is defined as:

$$\text{TS} - \text{score} = s \times h \times l \tag{1}$$

where s is the sequence identity between Q and the template domain T after the alignment. h is the normalized E-value from the alignment, i.e. $h = min(E_0, -log_{10}E)/E_0$, where $E_0 = 10$ is the normalization parameter. For example, $h = 0.3$ if the E-value $E = 0.001$, and $h = 1.0$ if $E \leq 1E-10$. l is a factor associated with the alignment coverage (c):

$$l = \begin{cases} 0 & \text{if } c \leq 1/3 \\ \dfrac{1}{1 + [1/(3c-1)]^5\}} & \text{if } c > 1/3 \end{cases} \tag{2}$$

where c equals to the number of aligned residues divided by the length of Q.

To account for the symmetry of the component segments, we define a character vector $\vec{v}_k = [s_k\ c_k]^T$ for the kth segment, where s_k and c_k are the sequence identity and the alignment coverage between the segment of the putative domain and the segment of the template, respectively. A Symmetric Index (SI) between the two segment-pairs is defined as the Euclidean distance between the vectors $\vec{v}_p$ and $\vec{v}_q$.

$$\text{SI} = \|\vec{v}_p - \vec{v}_q\| = \sqrt{(s_p - s_q)^2 + (c_p - c_q)^2} \tag{3}$$

To measure the sequence length similarity between the query and template domains, we defined the length variation e between the putative query domain (Q) and the template domain

Algorithm 2 The Pseudo Code of FindHit

INPUT: The Putative Domain Sequence Q_i;The cutoff T_b of Parameter b.

OUTPUT: The Best Templates T

```
1: procedure FINDHIT(Q_i)
2:     P(Q_i) ← Generate the Profile of Q_i using PSI-blast with nr database
3:     Templates ← Search homologous sequences using PSI-blast with the domain library and P(Q_i)
4:     homologs ← the template number in Templates
5:     for b = 0.9 → T_b, step = 0.1 do
6:         if homoglos > 0 then
7:             for all T_j ∈ Templates do
8:                 calculate e(Q_i, T_j);
9:                 if e(Q_i, T_j) < 0.2 then
10:                    next
11:                end if
12:                calculate SI(Q_i, T_j),TS_score(Q_i, T_j)
13:                T_TS ← TS_score(Q_i, T_j)
14:                T_SI ← f(T_TS, b)(by Eq.6)
15:                if SI(Q_i, T_j) ≤ T_SI then
16:                    if T_j is from Pfam then
17:                        Excute Profile-Profile alignment
18:                        if PPA_score ≤ T_PPA then
19:                            if PPA_score < min_score then
20:                                min_score ← PPA_score
21:                                bestTemplate ← T_j
22:                            end if
23:                        end if
24:                    else
25:                        bestTemplate ← T_j
26:                        last
27:                    end if
28:                end if
29:            end for
30:        end if
31:    end for
32:    return bestTempate
33: end procedure
```

Fig 3. The Pseudo code of FindHit.

doi:10.1371/journal.pone.0141541.g003

(T) from the domain library:

$$e = \frac{|L_T - L_Q|}{L_Q} \tag{4}$$

The three parameters of TS-score, SI and *e* will be used to help DomEx find homologous templates from the single-domain library.

Profile-profile alignment score

The profile-profile alignment combined with the predicted secondary structure information is used to filter out spurious discontinuous domains whose homologous templates are from the Pfam library. The sequence profile and secondary structure prediction are constructed by PSI--BLAST and the consensus of PSSpred (http://zhanglab.ccmb.med.umich.edu/PSSpred) and PSIPRED[44], respectively. The score function is similar to the threading algorithm PPA-I in LOMETS[39]. The profile-profile alignment score is defined as the score of the best alignment from dynamic programming between the query and template.

Training, validation and testing datasets

We constructed three datasets including Training Dataset, Validation Dataset and Testing Dataset. Training Dataset and Validation Dataset are used to train and validate the parameters of DomEx. In the training procedure, holdout validation is employed. The Validation Dataset is independent of the Training Dataset. The Testing Dataset is used to compare DomEx with ThreaDom. Furthermore, the Testing Dataset is also used to test the performance of detecting the discontinuous domain when DomEx is combined with other domain predictors.

The "Positive" and the "Negative" domain samples in the Training and the Validation Datasets are derived from the known structure domain segments. A positive sample refers to the segment combination that constitutes a true structure domain, while a negative sample is the combination that does not constitute a structural domain. Fig 4 shows an example of protein chain consisting of four segments that form three domains: (A1A2)(B)(C). Segment A1 and A2 form one structure domain; B and C form the other two independent domains. Then the segment assembly (A1A2) is a "Positive" sample, while (A1C) and (BC) are treated as "Negative" samples. Combinations of adjacent segments combination such as (A1B), (BA2), (A2C) are ignored as they are neighboring in sequence. The reversed combinations from the C- to N-terminal, e.g. (BA1) and (CB), are also ignored here, but will be discussed in the discussion section. Only the discontinuous domains containing two segments were considered here, since discontinuous domains including more than three segments are very rare (<2% in CATH3.5), and the extension to three-segment domains is straightforward.

From the CATH3.5 library, we collected 481 non-homologous proteins, which have known domain structure and consist of at least three segments. Among them, 326 contain at least one

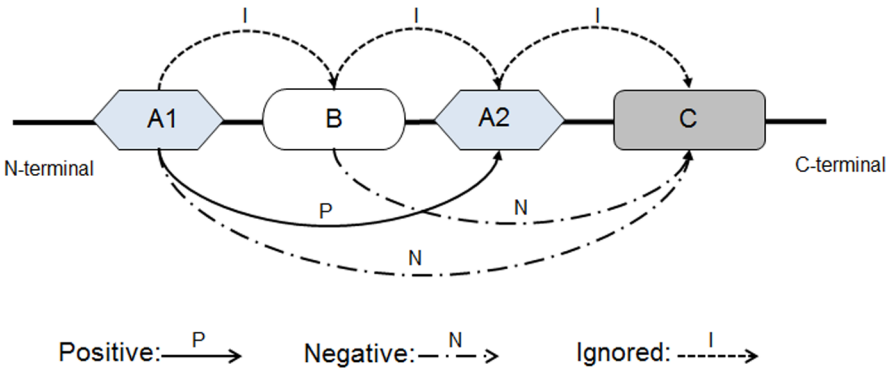

Fig 4. An illustration of the procedure to generate the samples. A 3-domain chain is defined as (A1A2)(B)(C). A1 and A2 form one structure domain, while B and C are independent domain, respectively. The (A1A1) is treated as "Positive" sample; (A1C) and (BC) as "Negative" and other combinations are ignored.

doi:10.1371/journal.pone.0141541.g004

discontinuous domain and 155 have three or more continuous domains. The pairwise sequence identity between the proteins is below 25%. From these proteins, we generated 344 positive discontinuous domains and 822 negative samples. The 822 negative samples have 273 from continuous multi-domain chains and 549 from incorrect discontinuous domain segment assemblies. Here, we only consider the cases that the segments have at least 40 residues, because most of protein domain predictors[28,31,32,38] consider a prediction to be "correct" if the predicted boundaries are within ±20 residues away from the true boundary. Then the maximum error of a correctly predicted segment is 40 residues and these domain predictors cannot report the domain boundary when the protein domain is less than 40 residues according to this criterion.

From the segment assemblies, we randomly selected 229 positive and 548 negative samples which are used as the training dataset to decide the parameters T_{TS} and T_{SI}; the others are used as the validation dataset to test the parameters obtained from training.

Our test set includes two subsets TEST-SET-I and TEST-SET-II. TEST-SET-I is used to test the robustness of DomEx by comparing it to ThreaDom alone on discontinuous domain detection. It contains 97 discontinuous domain protein chains, and all the boundaries predicted by ThreaDom are within ±20 residues to the annotated boundaries, and 80% of the boundary predictions have the error within ±5 residues.

TEST-SET-II is used to benchmark the domain predictors. It contains the same chains from which the training and the validation datasets were derived, but the boundaries will be predicted by different predictors.

Evaluation

The standard measurements of recall, precision and Matthews Correlation Coefficient (MCC) are employed to evaluate the performance of detecting the assembly of discontinuous domains from segments:

$$\begin{cases} recall = \dfrac{TP}{TP + FN} \\ precision = \dfrac{TP}{TP + FP} \\ \mathrm{MCC} = \dfrac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(FP + TN)(TN + FN)}} \end{cases} \quad (5)$$

where TP, FP, TN and FN denote the number of true positives, false positives, true negatives and false negatives, respectively.

NDO-score [45]is used to benchmark the different protein domain predictors. The NDO-score is defined as the normalized overlap rate of all predicted domain and linker regions with the true domain assignment in the native structure.

Results

Training and validation of DomEx

We trained DomEx using a 3-stage strategy: Exhaustive Search Training (EST), Equation Constraint Validation (ECV) and PPA Check Training (PCT). The EST procedure is used to train the cutoff T_{TS} of the TS-score and the cutoff T_{SI} of the SI with the Training Dataset. The ECV procedure is used for tuning the correlated cutoff parameter b in the constraint function $T_{SI} = f(T_{TS},b)$ based on the Training Dataset. PCT is used to train the PPA-score

cutoff T_{PPA} which is used to filter out incorrect templates from Pfam using the profile-profile alignment. All the parameters are validated on the Validation Dataset based on the holdout validation.

Exhaustive search training (EST). An exhaustive grid search of the 2-dimensional space of TS-score and SI was performed to find the best combination of cutoffs T_{TS} (for TS-score) and T_{SI} (for SI). Using a step size of 0.05, we searched for the optimal values of T_{TS} and T_{SI} in the range [0.1, 1] and [0.05, 0.5], respectively. The average MCC, recall and precision for the training dataset with different T_{TS} and T_{SI} are shown in Fig 5A–5C, respectively. It is easy to see that a high TS-score cutoff usually has high precision but low recall. As shown in Fig 5A to 5C, the boundaries of the adjacent regions are close to vertical lines when TS-score >0.5, which indicates that the MCC, recall and precision values are not sensitive to the Symmetric Index in the region of high TS-score.

The Symmetric Index becomes more important when TS-score <0.5. The highest MCC values (> 0.3) are in the region of TS-score ∈[0.1, 0.2] and SI∈[0.1, 0.2], where DomEx has a reasonable precision (>0.6) and maximum recall (>0.34).

Equation constraint validation (ECV). Since there is no unique cutoff that can achieve the best MCC, we designed a constraint relationship function $T_{SI} = f(T_{TS}, b)$ with a dynamic relationship between the cutoff T_{TS} and the cutoff T_{SI} controlled by a single

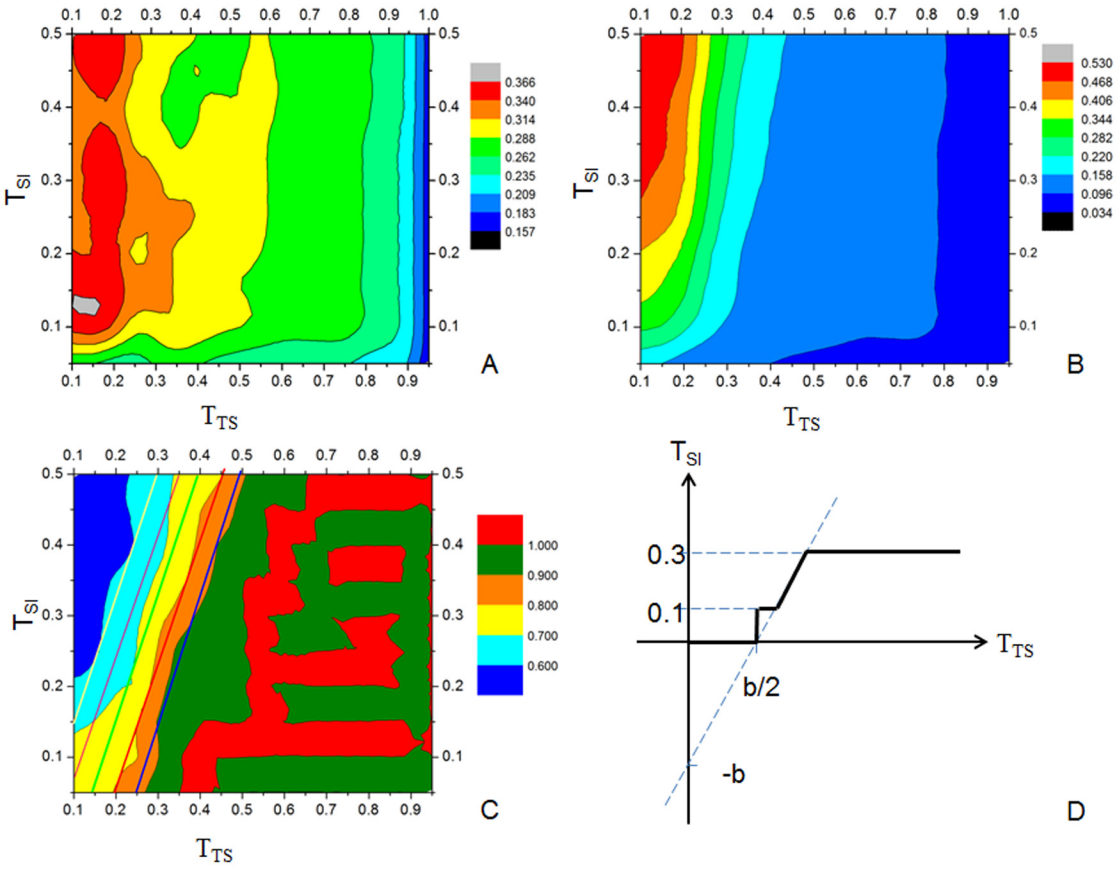

Fig 5. The recognition results of discontinuous domains at various TS-score and SI cutoffs. (A) MCC; (B) Recall; (C) Precision (5 parallel lines show the boundaries of the precision region at different b values). (D) The figure of Eq 6.

doi:10.1371/journal.pone.0141541.g005

parameter b, i.e.

$$T_{\mathrm{SI}} = \mathrm{f}(\mathrm{T_{TS}}, \mathrm{b}) = \begin{cases} 0.3, & 0.15 + \mathrm{b}/2 < T_{TS} \\ 2\mathrm{T}_{TS} - \mathrm{b}, & 0.05 + \mathrm{b}/2 < T_{TS} < 0.15 + \mathrm{b}/2 \\ 0.1, & \mathrm{b}/2 < T_{TS} < 0.05 + \mathrm{b}/2 \\ 0, & \mathrm{T}_{TS} < \mathrm{b}/2 \end{cases} \quad (6)$$

where different values of b correspond to different parallel lines that separate regions with similar precision values in Fig 5C. The curve of Eq 6 is shown in Fig 5D, which ensures that SI stays in the favorable region of [0.1, 0.3]. For example, given a parameter b = 0.4, a query sequence is aligned to some template with TS-score = 0.30 and SI = 0.15, then let TS-score$\geq T_{TS}$ = 0.30, and we get T_{SI} = 0.2 according to the Eq 6. We can infer that this query is a discontinuous domain because TS-score$\geq T_{SI}$ = 0.30 and SI$\leq T_{SI}$ = 0.2. Algorithm 2 (Fig 3) also shows this calculation procedure.

As confirmation, DomEx can achieve a reasonable MCC prediction above 0.3 for all the b values not only on the Training Dataset (Fig 6A), but also on the Validation Dataset (Fig 6B). It is more robust than that in EST step.

PPA check training (PCT). The training results were rechecked by Profile-Profile Alignment (PPA). The PPA-score gives the alignment quality between the query and template sequence. The score is usually negative. A low score means a good alignment. In the ECV training, there are 109 out of the 288 positive samples correctly detected as discontinuous, and 72 out of the 548 negative samples incorrectly detected as discontinuous when b>0.1. We use a total of 181 alignments as input to train the best PPA-score cutoff. We found the PPA can improve the MCC and precision when b<0.5 and PPA-score <-1.90 when the templates are collected from Pfam. Then the query is treated as a "Positive" detection, when the parameter b<0.5 and the PPA-score <-1.90.

We tested the method on 390 discontinuous-domain samples on the Validation Dataset, using b from 0.9 to 0.1 with step -0.1 and PPA-score <-1.90 when b<0.5. Similar to the tendency on the Training Dataset, the precision of DomEx in the Validation Dataset varies from 1.0 to 0.771. The MCC values are all higher than 0.30 with recall ranging from 0.183 to 0.339. For better balance between precision and recall, we selected b = 0.3 and PPA-score cutoff T_{PPA}<−1.90 as the default parameters in DomEx.

Test of DomEx

Discontinuous domain detection using accurately predicted boundaries. To detect the discontinuous domains, DomEx depends on the predicted boundaries from other predictors, such as ThreaDom. The test dataset TEST-SET-I contains 99 positive and 58 negative discontinuous domains which were derived from a non-redundant set of protein chains consisting of 97 discontinuous domains (Identity<25%, the length of shortest segments >40). Each boundary was predicted by ThreaDom within an error of ±20 residues to the boundary definitions in CATH 3.5.

In Table 1, we summarize the prediction results according to MCC, recall and precision and compare the results with the discontinuous domain detection of ThreaDom. DomEx recalled 32% of the discontinuous domains with 86% precision, which is similar to the results of using the annotated segment divisions from CATH 3.5. The discontinuous-domain detection method of ThreaDom is based on clustering the boundaries of the discontinuous domain templates. It is not surprising that it achieves a recall of 67.7% and precision of 87%, which is much higher than DomEx because ThreaDom is structured-based. But DomEx uses both the

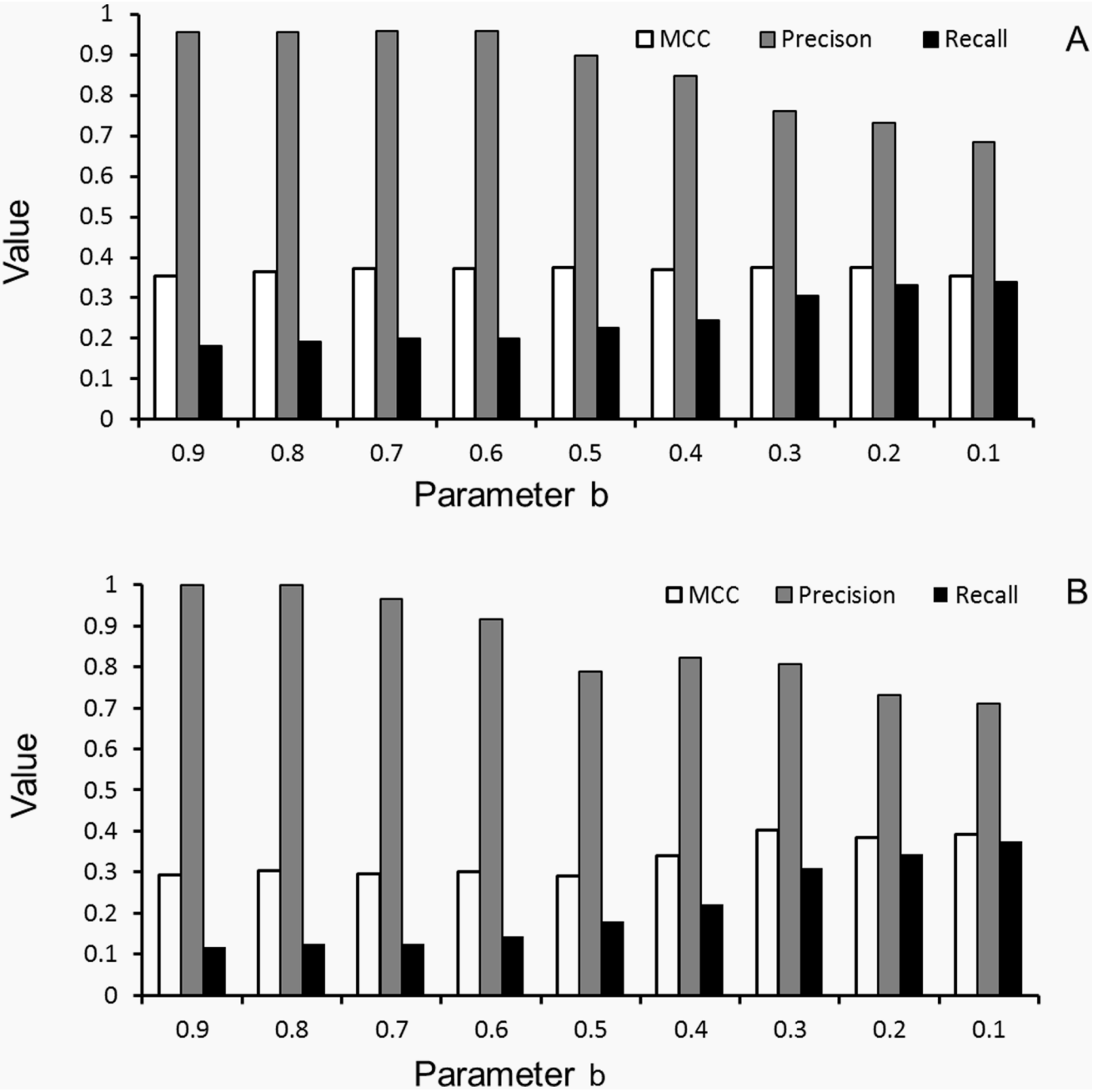

Fig 6. The training and validation results using $T_{SI} = f(T_{TS},b)$ constraints. The cutoff T_{TS} (for TS-score) and T_{SI} (for SI) in Fig 5 are constrained by Eq 6. (A) The results on the Training Dataset with parameter b from 0.9 to 0.1; (B) The results on the Validation Dataset with parameter b from 0.9 to 0.1, respectively.

doi:10.1371/journal.pone.0141541.g006

structure and sequence-based libraries, so it can handle the cases without 3D templates, where ThreaDom failed. We found that there are 29 chains where ThreaDom failed to detect the discontinuous domains (Group II in Table 1). However, for these chains, DomEx recalled 26.7% of the discontinuous domains with 72.7% precision. Half of the correct sequence templates came from Pfam, which demonstrates that DomEx works when there are no templates with known 3D structure.

Table 1. Discontinuous domain detections from predicted boundaries by ThreaDom.

Group	Prediction results			
	Method	Recall	Precision	MCC
[a]**Group I**	DomEx	0.348	0.923	0.271
Group I	ThreaDom	1.000	1.000	1.000
[b]**Group II**	DomEx	**0.267**	**0.727**	0.226
Group II	ThreaDom	—	—	—
ALL	DomEx	0.323	0.865	0.270
ALL	ThreaDom	0.677	0.870	0.568

DomEx predicted discontinuous domains with the boundaries predicted by ThreaDom boundary prediction method. ThreaDom predicted discontinuous domains with its own discontinuous domain detection method and ThreaDom boundary prediction method.

[a]Group I: The subset of TEST-SET I with 69 positive samples and 26 Negative samples from 68 protein domain chains that ThreaDom detects the discontinuous domain correctly.

[b]Group II: The subset of TEST-SET I with 30 positive samples and 32 Negative samples from 29 protein domain chains that ThreaDom fails to detect the discontinuous domain.

doi:10.1371/journal.pone.0141541.t001

Discontinuous domain prediction using TEST-SET-II. As a control, we employed five publicly available domain predictors, including ThreaDom[38], FIEFDom[26], DomPro [28], DROP[31] and PPRODO[30], which represent different types of homology- and machine-learning-based methods. Among them, ThreaDom can detect discontinuous domains, while the others cannot.

To test the situation where there are only weakly homologous templates, for ThreaDom, we excluded all templates that have a sequence identity >30% to the target protein, and we also excluded templates that are detectable by PSI-BLAST with an E-value<0.05.

For FIEFDom, we kept two group boundary prediction results, FIEFDom I and FIEFDom II. FIEFDom I excludes the templates if their sequence identity >30%; while FIEFDom II includes all the templates.

The dataset TEST-SET-II contains 481 multi-domain chains, which is the same size as the training and validation sets. It includes 326 discontinuous chains and 155 continuous domain chains. Each chain has at least 3 segments, and the length of each segment is not less than 40 residues. Fig 7 illustrates the NDO-score of the different methods. Here, we chose the best two boundary predictors, ThreaDom_Bdr and FIEFDom II to benchmark the performance of DomEx. They are denoted as ThrDm_Bdr+DomEx and FIEFDom II+ DomEx, respectively. We used ThreaDom_Bdr to represent the method which only predicts the boundaries without the boundary optimization and discontinuous domain detection option of ThreaDom. ThreaDom+DomEx uses DomEx to detect discontinuous domains when ThreaDom does not detect any discontinuous domains. In Fig 7, the dark bars represent the methods that support discontinuous-domain detection.

When detecting the discontinuous domains, both DomEx with ThreaDom boundaries (ThrDm_Bdr+ DomEx) and DomEx with FIEFDom II boundaries (FIEFDom II+DomEx) have a higher NDO-score than their boundary prediction without DomEx detection. And the NDO-score of ThrDm_Bdr+DomEx (0.70) is higher than all the predictors that do not support discontinuous-domain detection. ThreaDom+DomEx has the highest NDO-score of 0.759. The results demonstrate that DomEx can improve the discontinuous domain detection when combined with other boundary predictors because of the addition of a sequence-based domain library and the symmetric alignment score.

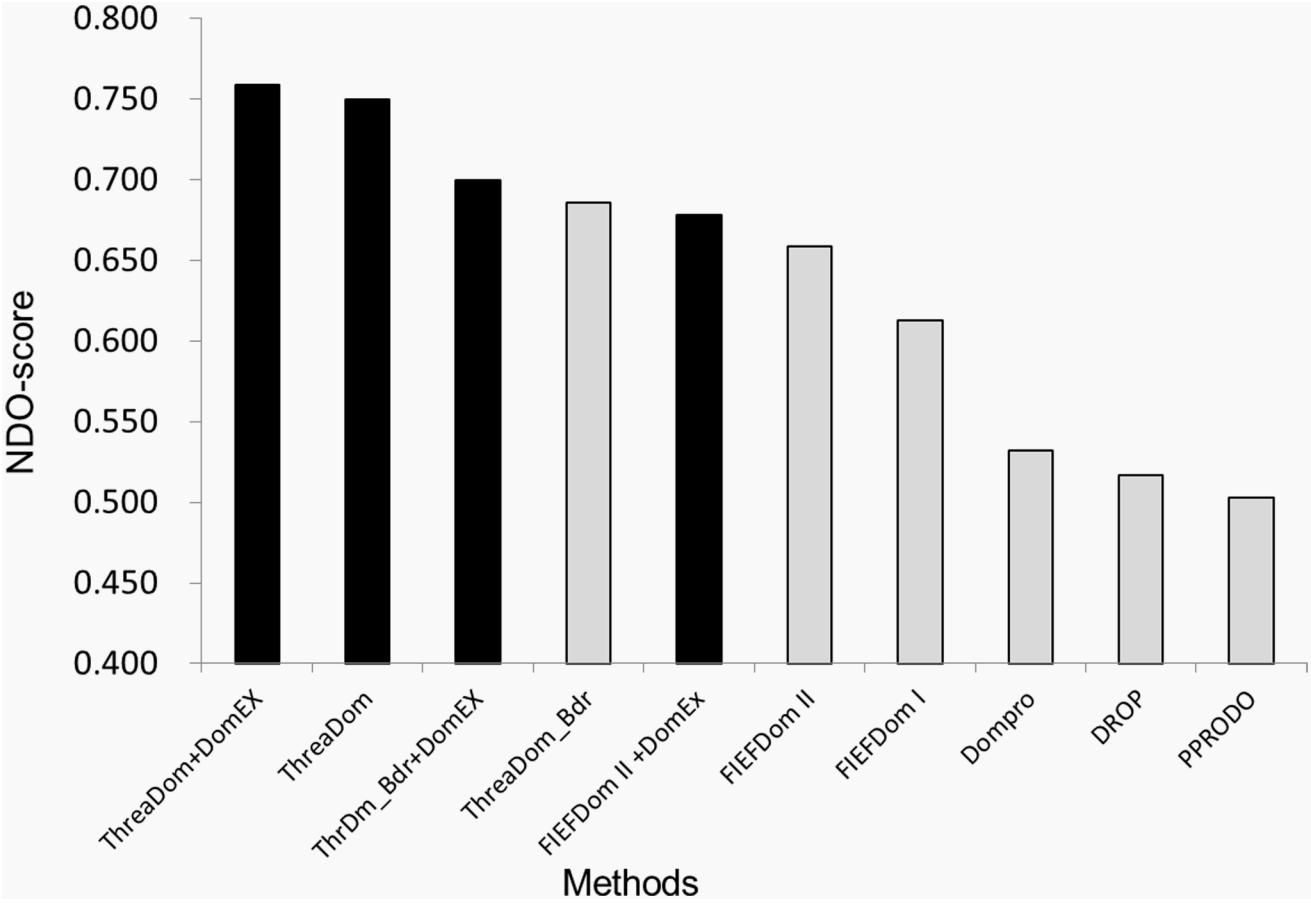

Fig 7. The benchmark results of DomEx with domain boundary predictors. The methods with discontinuous domain detection are shown as dark bars.

doi:10.1371/journal.pone.0141541.g007

Test DomEx with CASP targets. There are a total of 17 targets which have > = 3 continuous segments (length > 30 residues) from CASP8 to CASP10 experiments. Six contain multiple continuous domains, and eleven contain at least a discontinuous domain. A summary of the domain definitions from the CASP assessors is listed in S1 Table. To eliminate the negative effect of inaccurate boundary prediction, the assessor-based boundaries were used as input. The result showed that 36.7% discontinuous-domain proteins were correctly detected.

Discussion

Effect of templates from SCOP+CATH or Pfam

The domain sequence library of DomEx is based on CATH, SCOP and Pfam-A. The domain boundaries of SCOP and CATH are defined based on the 3D structure of the proteins. Pfam-A is based on the HMM classification of sequences from whole-genomes. It is observed that the accuracy of templates from the Pfam library is about 60~70% of that from CATH+SCOP. Fig 8A shows the comparison of the templates from CATH+SCOP, Pfam and CATH+SCOP +Pfam. Given a high b (for example, $b>0.8$), DomEx achieves a high precision and is independent of the template sources (CATH+SCOP or Pfam). When $b<0.8$, the accuracy of the Pfam-based prediction is significantly lower than CATH+SCOP. If the template comes from CATH or SCOP, DomEx has a high precision (>90%). If the templates come from Pfam, the precision declines to about 50% when b is less than 0.65.

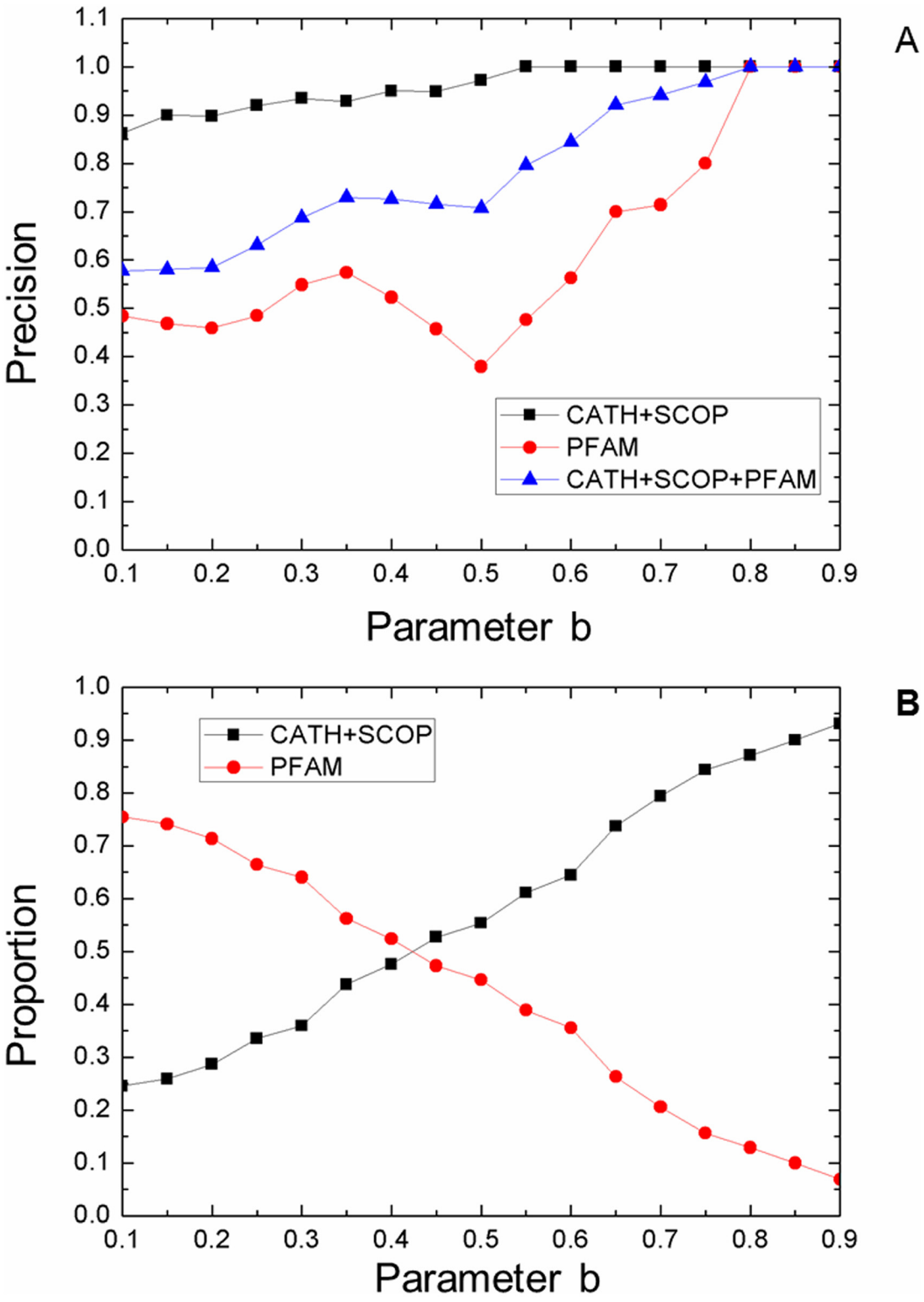

Fig 8. The comparison of the templates from CATH+SCOP, Pfam and CATH+SCOP+PFAM. (A) Precision comparison; (B) The proportion of templates coming from CATH+SCOP and Pfam as parameter b varies.

doi:10.1371/journal.pone.0141541.g008

Fig 8B gives the proportion of the templates from CATH+SCOP and Pfam across different values of *b*. For large *b*, the templates come mainly from CATH and SCOP, while for small *b*, much more templates are from Pfam. (We also note that the datasets are collected from CATH3.5, but the results are more convincing because the domain boundary definitions are based on known 3D structures) When $b = 0.2$, about 70% of the templates come from Pfam, and the recall of DomEx is about 44%. The recall is 100% higher than when $b = 0.5$, where 50% of the templates come from Pfam. The templates from Pfam increase the recall of DomEx, even though the precision of DomEx decreases when b decreases. The domain sequences from Pfam improve recall because of its large size relative to the other databases.

Effect of PPA score

In DomEx, the Profile-Profile Alignment combined with the predicted secondary structure information is used to help DomEx improve the low precision when the templates are from Pfam. Table 2 compares three PPA checking strategies: 1) with PPA check of only the templates from Pfam (Group I), 2) with PPA check of the templates from CATH+SCOP+Pfam (Group II), and 3) without PPA check (Group III). The PPA-score cutoff T_{PPA} in Table 2 is the one which has maximum MCC for some *b* when T_{PPA} varies from 0.0 to -5.0 with step -0.1. The average MCC and precision of Group I, which are 0.341 and 0.901, respectively, are higher than the other two. From the table, one can see that Group I has higher recall and MCC values for the cases when the precision values are similar to the other groups (underlined in table). For example, Group I has precision of 80.6% at $b = 0.3$, which is a bit higher than Group II at $b = 0.4$ and Group I at $b = 0.5$, while its recall is 31% and 61% higher than the other two, respectively.

PPA filtering is not as effective for Group II as it is for Group I. The main reason is the domain definitions from SCOP and CATH are more likely to be correct than Pfam since they are derived from known 3D structures. Therefore, we used PPA filtering only for templates from Pfam library.

Reversed assembly

In the segment assembly, we considered only the segment orderings from N- to C-termini, e.g. (A1A2) in Fig 4. We also examined the possibility of reversed segment assembly, e.g. (A2A1) in Fig 4, where the order of segments is reversed but the residue order within each segment

Table 2. Comparison of the Different PPA Checking Strategy.

b	With PPA check only to Pfam templates (Group I)				With PPA check to all type templates (Group II)				Without PPA check (Group III)		
	T_{PPA}	Recall	Precision	MCC_best	T_{PPA}	Recall	Precision	MCC_best	Recall	Precision	MCC
0.9	0.0	0.118	1.000	0.294	-0.15	0.118	1.000	0.294	0.122	1.000	0.299
0.8	0.0	0.127	1.000	0.305	-0.15	0.127	1.000	0.305	0.131	1.000	0.310
0.7	0.0	0.127	0.967	0.295	-0.15	0.131	0.968	0.301	0.135	0.969	0.306
0.6	-3.20	0.1354	1.0000	0.3153	-2.20	0.140	0.941	0.303	0.153	0.897	0.304
0.5	-3.30	0.1572	0.9730	0.3327	-1.50	0.179	0.820	0.302	0.192	0.786	0.300
0.4	-2.90	0.1921	0.9167	0.3500	-1.50	0.236	0.794	0.339	0.253	0.753	0.334
0.3	-1.90	**0.3100**	**0.8068**	**0.4014**	-1.50	0.332	0.768	0.396	0.358	0.739	0.398
0.2	-1.90	0.3450	0.7315	0.3849	-1.50	0.371	0.675	0.367	0.406	0.620	0.349
0.1	-1.90	0.3755	0.7107	0.3919	-1.50	0.424	0.660	0.387	0.476	0.602	0.372
Average	–	0.210	**0.901**	**0.341**	–	0.229	0.847	0.333	0.247	0.818	0.330

doi:10.1371/journal.pone.0141541.t002

sequence is unchanged. We found that DomEx reported only 3 discontinuous domains out of all the 344 positive samples in the training and validation dataset using the reversed order when b = 0.2, which is less than 0.9% of the positive samples. Therefore, we have ignored the reverse assembly in the default settings of DomEx to save CPU time (about twice as long).

Nevertheless, the reversed segment assembly can detect some interesting domain structures. All the three reversed cases are from proteins with segment-swapping domains (SSDs) [46]. There are two cases the templates coming from CATH and one case from Pfam. Fig 9 shows two examples with templates from CATH3.5 whose structure is known. In Fig 9A, the template has the PDB ID of 1axkB which is defined as (1–156|342–393)(157–341) in CATH3.5. The first

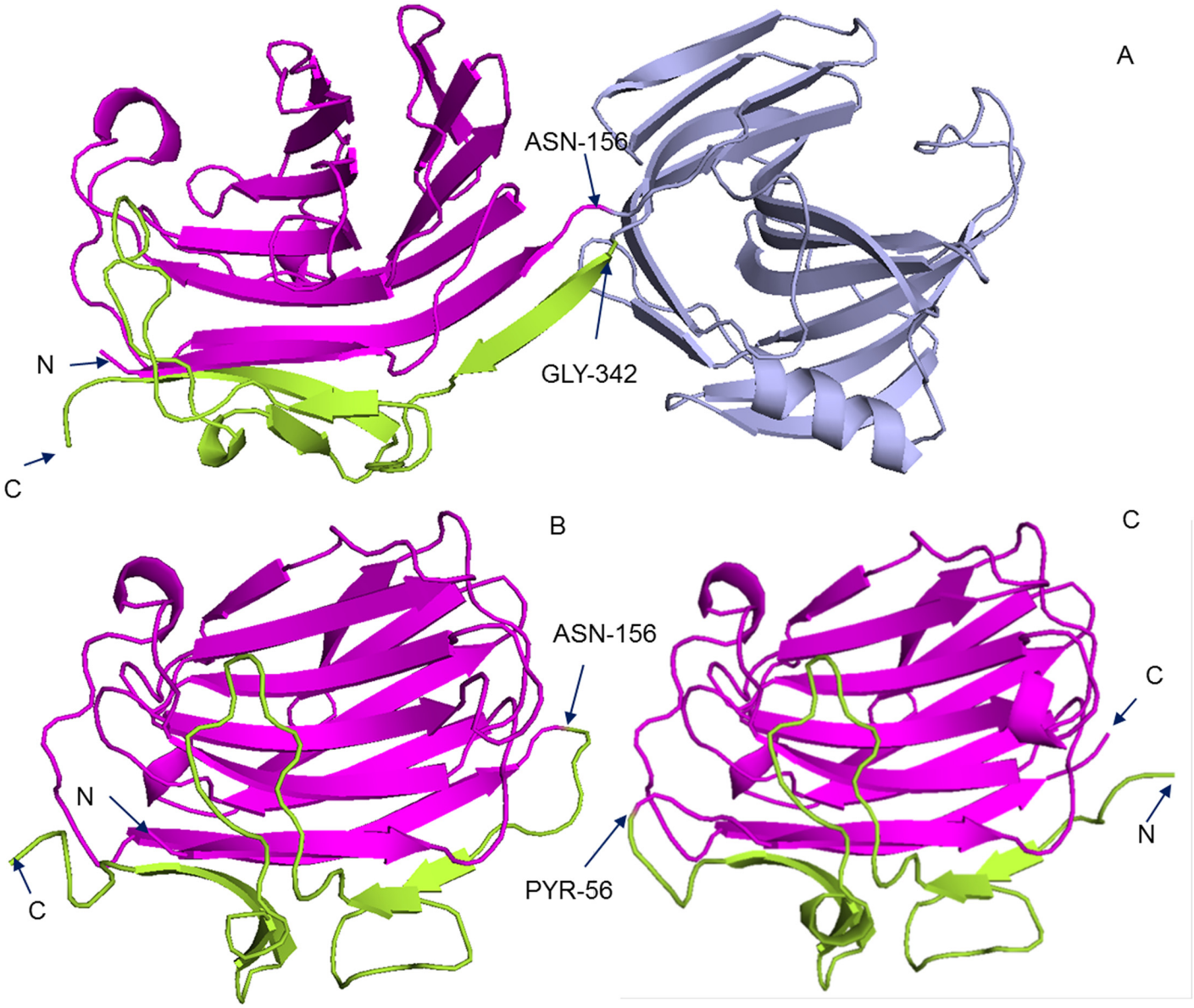

Fig 9. The cases of N- to C-termini assembly. (A) The 3D structure of PDB: 1axkB. The two segments of the discontinuous domain (1–156|342–393) are colored in magenta and lemon green, respectively. (B) The 3D structure of PDB:1u0aD. It is an AB type Segment-Swapping Domain. (C) The 3D structure of PDB:1cpmA. It is a BA type Segment-Swapping Domain.

doi:10.1371/journal.pone.0141541.g009

domain is discontinuous, and the two segments are colored in magenta and lemon green, respectively. Fig 9B and 9C show the templates (PDB ID: 1u0aD and 1cpmA) for the same sequence, which are detected by segment assembly using orders from N- to C-termini and from C- to N-termini, respectively. The two templates are both single-domain chains, where 1u0aD is an AB type SSD and 1cpmA is a BA type SSD. They both have similar 3D structure. In the DomEx package, we have included an option to turn on reversed segment assembly.

Conclusion

We have proposed a new strategy, DomEx, to extend the ability of domain boundary predictors to detect discontinuous domains. The method assembles and detects discontinuous domains from the sequence segments. DomEx incorporates template similarity, symmetry of segment pairs, profile-profile alignments, and structure-based and structure-free libraries.

Two test benchmarks showed that DomEx not only worked with the boundary predictors, but also was complementary to the discontinuous-domain detection method in ThreaDom. DomEx recalled 26.7% of the cases where ThreaDom failed. Half of these cases are attributed to templates from Pfam. When compared to other methods, DomEx plus ThreaDom gave the best NDO score, which further confirms that DomEx can detect discontinuous domains even without known 3D-structure templates. The main advantage of DomEx is that it searches for templates using domain-domain alignments rather than chain-chain alignments. Using domain-domain alignments improves recall because chain-chain alignments may miss templates where the domains match, but the rest of the chain does not match well. The benchmark results show that DomEx is an effective method, which opens the possibility of finding discontinuous domains in genome-wide studies. Currently, DomEx supports the detection of two-segment discontinuous domains. Further work will extend the model to detect discontinuous domains with more than 2 segments, and try to utilize the domain annotated information from CATH, SCOP and Pfam to enhance the performance of detecting discontinuous domains. The accuracy of boundary predictors and sequence alignment tools should also improve the detection results. The source code and datasets of DomEx are available at https://github.com/xuezhidong/DomEx.

Supporting Information

S1 Table. Domain definition of the 17 targets in CASP8, CASP9 and CASP10 to test DomEx.
(PDF)

Acknowledgments

The authors wish to thank Drs. J. Yang, D. Xu and J. Wang. The project is supported in part National Natural Science Foundation of China (30700162, 61073095), the Fundamental Research Funds for the Central Universities of China (HUST2014TS138 and HUST2015QN101)) and China Postdoctoral Science Foundation (2014M552043).

Author Contributions

Conceived and designed the experiments: ZX YW. Performed the experiments: ZX RJ BG YH YW. Analyzed the data: ZX RJ BG YH YW. Contributed reagents/materials/analysis tools: ZX RJ BG YH YW. Wrote the paper: ZX RJ BG YW.

References

1. Hondoh T, Kato A, Yokoyama S, Kuroda Y. Computer-aided NMR assay for detecting natively folded structural domains. Protein Sci. 2006; 15(4):871–83. doi: 10.1110/ps.051880406 PMID: 16522794; PubMed Central PMCID: PMC2242495.
2. Folkers GE, van Buuren BN, Kaptein R. Expression screening, protein purification and NMR analysis of human protein domains for structural genomics. Journal of structural and functional genomics. 2004; 5 (1–2):119–31. doi: 10.1023/B:JSFG.0000029200.66197.0c PMID: 15263851.
3. Contreras-Moreira B, Bates PA. Domain fishing: a first step in protein comparative modelling. Bioinformatics. 2002; 18(8):1141–2. PMID: 12176841.
4. Zhang Y. Progress and challenges in protein structure prediction. Curr Opin Struct Biol. 2008; 18 (3):342–8. Epub 2008/04/26. doi: 10.1016/j.sbi.2008.02.004 S0959-440X(08)00034-1 [pii]. PMID: 18436442; PubMed Central PMCID: PMC2680823.
5. Zhang Y. Interplay of I-TASSER and QUARK for template-based and ab initio protein structure prediction in CASP10. Proteins. 2014; 82 Suppl 2:175–87. doi: 10.1002/prot.24341 PMID: 23760925; PubMed Central PMCID: PMC4067246.
6. Reeves GA, Dallman TJ, Redfern OC, Akpor A, Orengo CA. Structural diversity of domain superfamilies in the CATH database. J Mol Biol. 2006; 360(3):725–41. doi: 10.1016/j.jmb.2006.05.035 PMID: 16780872.
7. Dessailly BH, Redfern OC, Cuff AL, Orengo CA. Detailed analysis of function divergence in a large and diverse domain superfamily: toward a refined protocol of function classification. Structure. 2010; 18 (11):1522–35. doi: 10.1016/j.str.2010.08.017 PMID: 21070951; PubMed Central PMCID: PMC3023962.
8. Siddiqui AS, Barton GJ. Continuous and discontinuous domains: an algorithm for the automatic generation of reliable protein domain definitions. Protein Sci. 1995; 4(5):872–84. doi: 10.1002/pro.5560040507 PMID: 7663343; PubMed Central PMCID: PMC2143117.
9. Cuff AL, Sillitoe I, Lewis T, Clegg AB, Rentzsch R, Furnham N, et al. Extending CATH: increasing coverage of the protein structure universe and linking structure with function. Nucleic Acids Res. 2011; 39 (Database issue):D420–6. Epub 2010/11/26. doi: 10.1093/nar/gkq1001 gkq1001 [pii]. PMID: 21097779; PubMed Central PMCID: PMC3013636.
10. Cuff A, Redfern OC, Greene L, Sillitoe I, Lewis T, Dibley M, et al. The CATH hierarchy revisited-structural divergence in domain superfamilies and the continuity of fold space. Structure. 2009; 17(8):1051–62. Epub 2009/08/15. doi: 10.1016/j.str.2009.06.015 S0969-2126(09)00258-5 [pii]. PMID: 19679085; PubMed Central PMCID: PMC2741583.
11. Xu Y, Xu D, Gabow HN. Protein domain decomposition using a graph-theoretic approach. Bioinformatics. 2000; 16(12):1091–104. PMID: 11159328
12. Alexandrov N, Shindyalov I. PDP: protein domain parser. Bioinformatics. 2003; 19(3):429–30. Epub 2003/02/14. PMID: 12584135.
13. Pugalenthi G, Archunan G, Sowdhamini R. DIAL: a web-based server for the automatic identification of structural domains in proteins. Nucleic Acids Res. 2005; 33(Web Server issue):W130–2. doi: 10.1093/nar/gki427 PMID: 15980441; PubMed Central PMCID: PMC1160188.
14. Taylor WR. Protein structural domain identification. Protein Eng. 1999; 12(3):203–16. PMID: 10235621.
15. Siddiqui AS, Dengler U, Barton GJ. 3Dee: a database of protein structural domains. Bioinformatics. 2001; 17(2):200–1. PMID: 11238081.
16. Andreeva A, Howorth D, Brenner SE, Hubbard TJ, Chothia C, Murzin AG. SCOP database in 2004: refinements integrate structure and sequence family data. Nucleic Acids Res. 2004; 32(Database issue):D226–9. Epub 2003/12/19. doi: 10.1093/nar/gkh039 32/suppl_1/D226 [pii]. PMID: 14681400; PubMed Central PMCID: PMC308773.
17. Ponting CP, Schultz J, Milpetz F, Bork P. SMART: identification and annotation of domains from signalling and extracellular protein sequences. Nucleic Acids Res. 1999; 27(1):229–32. Epub 1998/12/10. doi: gkc010 [pii]. PMID: 9847187; PubMed Central PMCID: PMC148142.
18. Wheelan S, Marchler-Bauer A, Bryant SH. Domain size distributions can predict domain boundaries. Bioinformatics. 2000; 16(7):613–8. PMID: 11038331
19. Suyama M, Ohara O. DomCut: prediction of inter-domain linker regions in amino acid sequences. Bioinformatics. 2003; 19(5):673–4. Epub 2003/03/26. PMID: 12651735.
20. Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, et al. The Pfam protein families database. Nucleic Acids Res. 2012; 40(Database issue):D290–301. Epub 2011/12/01. doi: 10.1093/nar/gkr1065 gkr1065 [pii]. PMID: 22127870; PubMed Central PMCID: PMC3245129.

21. Sonnhammer EL, Eddy SR, Birney E, Bateman A, Durbin R. Pfam: multiple sequence alignments and HMM-profiles of protein domains. Nucleic Acids Res. 1998; 26(1):320–2. Epub 1998/02/21. doi: gkb019 [pii]. PMID: 9399864; PubMed Central PMCID: PMC147209.

22. Sonnhammer EL, Eddy SR, Durbin R. Pfam: a comprehensive database of protein domain families based on seed alignments. Proteins. 1997; 28(3):405–20. Epub 1997/07/01. doi: 10.1002/(SICI)1097-0134(199707)28:3<405::AID-PROT10>3.0.CO;2-L [pii]. PMID: 9223186.

23. Portugaly E, Linial N, Linial M. EVEREST: a collection of evolutionary conserved protein domains. Nucleic Acids Res. 2007; 35(Database issue):D241–6. Epub 2006/11/14. doi: gkl850 [pii] doi: 10.1093/nar/gkl850 PMID: 17099230; PubMed Central PMCID: PMC1669739.

24. Portugaly E, Harel A, Linial N, Linial M. EVEREST: automatic identification and classification of protein domains in all protein sequences. BMC Bioinformatics. 2006; 7:277. Epub 2006/06/06. doi: 1471-2105-7-277 [pii] doi: 10.1186/1471-2105-7-277 PMID: 16749920; PubMed Central PMCID: PMC1533870.

25. Heger A, Wilton CA, Sivakumar A, Holm L. ADDA: a domain database with global coverage of the protein universe. Nucleic Acids Res. 2005; 33(Database issue):D188–91. Epub 2004/12/21. doi: 33/suppl_1/D188 [pii] doi: 10.1093/nar/gki096 PMID: 15608174; PubMed Central PMCID: PMC540050.

26. Bondugula R, Lee MS, Wallqvist A. FIEFDom: a transparent domain boundary recognition system using a fuzzy mean operator. Nucleic Acids Res. 2009; 37(2):452–62. Epub 2008/12/06. doi: 10.1093/nar/gkn944 gkn944 [pii]. PMID: 19056827; PubMed Central PMCID: PMC2632928.

27. Liu J, Rost B. Sequence-based prediction of protein domains. Nucleic acids research. 2004; 32 (12):3522–30. PMID: 15240828

28. Cheng J, Sweredoski MJ, Baldi P. DOMpro: protein domain prediction using profiles, secondary structure, relative solvent accessibility, and recursive neural networks. Data Mining and Knowledge Discovery. 2006; 13(1):1–10.

29. Yoo PD, Sikder AR, Taheri J, Zhou BB, Zomaya AY. DomNet: protein domain boundary prediction using enhanced general regression network and new profiles. IEEE Trans Nanobioscience. 2008; 7 (2):172–81. Epub 2008/06/17. doi: 10.1109/TNB.2008.2000747 PMID: 18556265.

30. Sim J, Kim SY, Lee J. PPRODO: prediction of protein domain boundaries using neural networks. Proteins. 2005; 59(3):627–32. Epub 2005/03/25. doi: 10.1002/prot.20442 PMID: 15789433.

31. Ebina T, Toh H, Kuroda Y. DROP: an SVM domain linker predictor trained with optimal features selected by random forest. Bioinformatics. 2011; 27(4):487–94. Epub 2010/12/21. doi: 10.1093/bioinformatics/btq700 btq700 [pii]. PMID: 21169376.

32. Eickholt J, Deng X, Cheng J. DoBo: Protein domain boundary prediction by integrating evolutionary signals and machine learning. BMC Bioinformatics. 2011; 12:43. Epub 2011/02/03. doi: 10.1186/1471-2105-12-43 1471-2105-12-43 [pii]. PMID: 21284866; PubMed Central PMCID: PMC3036623.

33. George RA, Heringa J. SnapDRAGON: a method to delineate protein structural domains from sequence data. J Mol Biol. 2002; 316(3):839–51. Epub 2002/02/28. doi: 10.1006/jmbi.2001.5387 S0022283601953872 [pii]. PMID: 11866536.

34. Kim DE, Chivian D, Malmstrom L, Baker D. Automated prediction of domain boundaries in CASP6 targets using Ginzu and RosettaDOM. Proteins. 2005; 61 Suppl 7:193–200. Epub 2005/09/28. doi: 10.1002/prot.20737 PMID: 16187362.

35. Wu Y, Dousis AD, Chen M, Li J, Ma J. OPUS-Dom: applying the folding-based method VECFOLD to determine protein domain boundaries. J Mol Biol. 2009; 385(4):1314–29. Epub 2008/11/26. doi: 10.1016/j.jmb.2008.10.093 S0022-2836(08)01407-1 [pii]. PMID: 19026662; PubMed Central PMCID: PMC2753268.

36. Sikder AR, Zomaya AY. Inferring boundary information of discontinuous-domain proteins. IEEE Trans Nanobioscience. 2008; 7(3):200–5. Epub 2008/09/10. doi: 10.1109/TNB.2008.2002283 PMID: 18779100.

37. Chen P, Li J. Prediction of protein long-range contacts using an ensemble of genetic algorithm classifiers with sequence profile centers. BMC structural biology. 2010; 10 Suppl 1:S2. doi: 10.1186/1472-6807-10-S1-S2 PMID: 20487509; PubMed Central PMCID: PMC2873825.

38. Xue Z, Xu D, Wang Y, Zhang Y. ThreaDom: extracting protein domain boundary information from multiple threading alignments. Bioinformatics. 2013; 29(13):i247–i56. Epub 2013/07/03. doi: 10.1093/bioinformatics/btt209 btt209 [pii]. PMID: 23812990; PubMed Central PMCID: PMC3694664.

39. Wu S, Zhang Y. LOMETS: a local meta-threading-server for protein structure prediction. Nucleic acids research. 2007; 35(10):3375–82. PMID: 17478507

40. Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM. CATH—a hierarchic classification of protein domain structures. Structure. 1997; 5(8):1093–108. Epub 1997/08/15. PMID: 9309224.

41. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol. 1995; 247(4):536–40. PMID: 7723011.

42. Magrane M, Consortium U. UniProt Knowledgebase: a hub of integrated protein data. Database (Oxford). 2011; 2011:bar009. Epub 2011/03/31. doi: bar009 [pii] doi: 10.1093/database/bar009 PMID: 21447597; PubMed Central PMCID: PMC3070428.

43. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997; 25 (17):3389–402. Epub 1997/09/01. doi: gka562 [pii]. PMID: 9254694; PubMed Central PMCID: PMC146917.

44. McGuffin LJ, Bryson K, Jones DT. The PSIPRED protein structure prediction server. Bioinformatics. 2000; 16(4):404–5. Epub 2000/06/27. PMID: 10869041.

45. Tai CH, Lee WJ, Vincent JJ, Lee B. Evaluation of domain prediction in CASP6. Proteins: Structure, Function, and Bioinformatics. 2005; 61(S7):183–92.

46. Szilagyi A, Zhang Y, Zavodszky P. Intra-chain 3D segment swapping spawns the evolution of new multidomain protein architectures. J Mol Biol. 2012; 415(1):221–35. Epub 2011/11/15. doi: 10.1016/j.jmb.2011.10.045 S0022-2836(11)01194-6 [pii]. PMID: 22079367; PubMed Central PMCID: PMC3249503.