J Comput Aided Mol Des (2013) 27:419–426
DOI 10.1007/s10822-013-9654-6

IS-Dom: a dataset of independent structural domains automatically delineated from protein structures

Teppei Ebina · Yuki Umezawa · Yutaka Kuroda

Received: 27 September 2012 / Accepted: 7 May 2013 / Published online: 29 May 2013

Abstract Protein domains that can fold in isolation are significant targets in diverse area of proteomics research as they are often readily analyzed by high-throughput methods. Here, we report IS-Dom, a dataset of Independent Structural Domains (ISDs) that are most likely to fold in isolation. IS-Dom was constructed by filtering domains from SCOP, CATH, and DomainParser using quantitative structural measures, which were calculated by estimating inter-domain hydrophobic clusters and hydrogen bonds from the full length protein's atomic coordinates. The ISD detection protocol is fully automated, and all of the computed interactions are stored in the server which enables rapid update of IS-Dom. We also prepared a standard IS-Dom using parameters optimized by maximizing the Youden's index. The standard IS-Dom, contained 54,860 ISDs, of which 25.5 % had high sequence identity and termini overlap with a Protein Data Bank (PDB) cataloged sequence and are thus experimentally shown to fold in isolation [coined autonomously folded domain (AFDs)]. Furthermore, our ISD detection protocol missed less than 10 % of the AFDs, which corroborated our protocol's ability to define structural domains that are able to fold independently. IS-Dom is available through the web server (http://domserv.lab.tuat.ac.jp/IS-Dom.html), and users can either, download the standard IS-Dom dataset, construct their own IS-Dom by interactively varying the parameters, or assess the structural independence of newly defined putative domains.

Keywords Structural domains · Autonomously folded · SCOP · CATH · DomainParser · Domain database · Protein dissection · High-throughput · Proteomics · Domain prediction

Abbreviation

ISD	Independent Structural Domain: Domains that fulfill the inter-domain interaction criteria calculated from the atomic coordinates of the full length protein, and are therefore likely to fold in isolation (or independently)
AFD	Autonomously folded domain: Domains that fulfill the sequence identity and termini overlap criteria with sequences listed in the PDB, and are therefore experimentally or "nearly" experimentally demonstrated to fold in isolation
CATH	Class, Architecture, Topology and Homologous superfamily
MC	Main-chain
SC	Side-chain
PDB	Protein Data Bank

Teppei Ebina and Yuki Umezawa contributed equally to this work.

Availability: IS-Dom is available at http://domserv.lab.tuat.ac.jp/IS-Dom.html and http://domserv.lab.tuat.ac.jp/

Electronic supplementary material The online version of this article (doi:10.1007/s10822-013-9654-6) contains supplementary material, which is available to authorized users.

T. Ebina (✉) · Y. Umezawa · Y. Kuroda (✉)
Department of Biotechnology and Life Science, Tokyo University of Agriculture and Technology, 12-24-16 Nakamachi, Koganei-shi, Tokyo 184-8588, Japan
e-mail: teppei-ebina@brain.riken.jp

Y. Kuroda
e-mail: ykuroda@cc.tuat.ac.jp

Present Address:
T. Ebina
Brain Science Institute, RIKEN, 2-1 Hirosawa, Wako-shi, Saitama 351-0198, Japan

SCOP Structural Classification of Proteins
SEM Standard Error of the Mean

Introduction

Most large proteins consist of multiple structural domains that fold and sometimes exert function in isolation [1]. Unlike their large, multi-domain protein counterparts from which they are isolated, stably folded structural domains are readily analyzable by high-throughput biochemical methods [2–5]. Their identifications and classifications into datasets are thus gaining importance in proteomics research, especially, as a mean for classifying fundamental structural units [6–8] or as training datasets for constructing domain predictors [9, 10]. Most approaches for delineating structural domains from the structure of the whole protein are based on the hypothesis that they contain more intra- than inter-domain interactions enabling them to fold independently in a native form. For example, several methods identify structural domains based on the detection of densely clustered amino acid regions from the protein's atomic coordinates [11–15].

SCOP [6] and CATH [7] are two popular, well established domain databases. Their domain definitions are widely accepted, and used for investigating the sequence and structure patterns of proteins and their domains. Nevertheless, the consensus between the SCOP and CATH domain definition remains moderate [16]. Two major reasons for the inconsistencies are: First a lack of a quantitative measure for objectively defining a structural domain; and second a lack of experimental assessment for the foldability of structural domains. Indeed, domains in SCOP are mostly detected by visual inspections combined with a search for conserved amino acid sequences [6]. Similarly, domains in CATH are detected by using sequence similarity information, computational structural domain prediction methods, such as DOMAK [17], and manual procedures [7].

In this study, we report IS-Dom, a dataset of independent structural domains (ISDs) that would have the ability to fold independently. Our definition of an ISD puts a special emphasis on the ability of structural domains to fold in isolation, because this can be used as an objective and straightforward criterion to assess the suitability of the definition (see Methods section). In order to construct IS-Dom, we developed an automatic protocol for determining ISDs based on a domain's putative foldability predicted using its atomic coordinates. We defined ISDs by selecting domains from SCOP [6], CATH [7] and DomainParser [15] that form little or no inter-domain interactions, and we assessed that ISDs were likely to fold independently by sequence identity and sequence overlap with the entire length of sequences listed in the PBD. All of the ISDs and the full list of inter-residue interactions were stored in our IS-Dom server (http://domserv.lab.tuat.ac.jp/IS-Dom.html) enabling a rapid retrieval of ISD datasets computed by varying the interaction parameters.

Methods

AFD definition from sequence identity to PDB cataloged sequences

We defined an AFD as a SCOP/CATH or DomainParser domain having an amino acid sequence identity >30 % [18], within a 10-residue termini margin, over the full length of a protein sequence cataloged in the PDB (Fig. S-1). A "PDB cataloged sequence" is a sequence listed in the PDB and retrieved from the pdbaa dataset (ftp:/ftp.ncbi.nih.gov/blast/db/FASTA/pdbaa.gz), and residues with and without atomic coordinates were both included. Sequence similarities were searched using BLAST with default parameters. Sequence alignment between a domain and a PDB cataloged was obtained from BLAST search, and the sequence identity was calculated as the number of the identical residues ("Identities" values in the BLAST result) divided by the number of residues in the query sequence. Although the AFD's foldability is not directly assessed, this criterion ensures that AFDs are very much likely to fold into their native structures. This AFD's definition is essentially the same as that previously used to assess the foldability of domains (previously coined AFU: Autonomously Folded Units [19]).

We constructed a control dataset of AFDs and non-AFDs (domains not classified as AFD) using 16,601 and 28,060 multi-domain proteins listed in SCOP release 1.75A (http:/scop.mrc-lmb.cam.ac.uk/scop/) and CATH release 3.5.0 (http:/www.cathdb.info/), respectively (NMR structures were not used). This resulted in 100,269 (37,037 SCOP and 63,232 CATH) continuous domains, of which 16,462 (6,372 SCOP and 10,090 CATH) were AFDs. Discontinuous domains were not included in this dataset as they are unlikely to fold in isolation and our definition of AFD was originally motivated by the practical need of detecting independently foldable domains for rapid structural and functional proteomics analysis [3, 19]. The AFDs and non-AFDs were clustered using a single linkage algorithm with a threshold of 30 % sequence identity (percentage calculated using the number of residues in the shorter sequence). Representatives were selected by choosing sequences with the highest mean sequence identity calculated across the cluster members. This yielded 460 and 3,574 representative AFDs and non-AFDs

respectively, which we used as control datasets for testing the ability of domains to fold in isolation.

ISD detection from the full length protein structure

We assumed that domains with little inter-domain interactions are able to fold independently and detected ISDs by calculating inter-domain interactions from the atomic coordinates of the full length protein structures. A domain was defined as an ISD when it formed less inter-domain hydrophobic clusters and H-bonds than a given set of threshold values, and as non-ISDs, otherwise.

A hydrophobic cluster was assumed when the minimum distance between the side-chain carbon atoms of three hydrophobic residues (Ala, Val, Leu, Ile, Phe, Met, Trp and Tyr) were within a distance threshold [20]. The distance threshold for hydrophobic cluster was optimized by using the p values of Wilcoxon rank-sum tests comparing the numbers of inter-domain hydrophobic clusters formed by AFDs and non-AFDs. We selected a distance threshold value with a p-value <0.05 as statistically significant differences between the numbers of AFDs and non-AFDs. H-bonds were detected by HBPLUS [21], and classified into main-chain main-chain hydrogen bonds (MC–MC H-bonds), main-chain side-chain hydrogen bonds (MC–SC H-bonds) and side-chain side-chain hydrogen bonds (SC–SC H-bonds).

IS-Dom construction

We constructed IS-Dom, a dataset of ISDs, using all of the 213,010 PDB chains as of January 15th 2013 (Fig. 1) which contained 16,601 SCOP, 28,030 CATH, and 37,753 DomainParser multi-domain proteins. We calculated ISDs for 21 different threshold numbers (hydrophobic cluster, MC–MC, MC–SC and SC–SC H-bonds; from 0 to 20) and 18 distance thresholds for the hydrophobic cluster (from 1.0 to 10.0 by step of 0.5Å). This yielded $21 \times 21 \times 21 \times 21 \times 19 = 3{,}695{,}139$ ISD datasets in IS-Dom corresponding to all possible combinations of interaction numbers and distance thresholds. ISDs with both N- and C-termini within 10 residues were considered identical, and the largest ISD was listed in IS-Dom.

Standard IS-Dom and optimization of the interaction numbers

From the 3,695,139 datasets, we defined a standard IS-Dom containing ISDs that formed less inter-domain interactions than an optimized set of threshold numbers (11 inter-domain MC–MC, 9 MC–SC, 6 SC–SC H-bonds and 7 hydrophobic clusters calculated with a 5.0 Å distance threshold). The parameters were optimized by searching exhaustively the combination that maximized the Youden's index (YI, [22]). The YI was calculated as R_{TP}-R_{FP}, where R_{TP} and R_{FP} are, respectively, the ratio of AFDs and non-AFDs in ISDs to those in the control dataset. ISD representatives were selected in a way identical to the AFDs, but sequences were clustered by using a 30 % sequence identity threshold and a 10-residue termini margin.

Results and discussion

Control AFD dataset and its assessment

We constructed a control dataset of AFDs using the SCOP and CATH domain definitions and according to the sequence identity and 10-residue termini margin based definition (see Methods;see also Figure S-1 in the supplemental material). AFDs were searched from 100,269 continuous domains cataloged either in SCOP (37,037) or CATH (63,232). The control AFD dataset contained 16,462 (6,372 SCOP and 10,090 CATH) AFDs which were clustered into 460 representatives (3,574 non-AFDs representatives).

63.0 % (290) of the AFDs were similar to a monomeric protein, i.e., its PDB file contained a single chain. Moreover, another 34.0 % (139) of the AFDs formed homo-oligomer and only 7 % were hetero-oligomers. These results indicate that most AFDs can fold in isolation and without a need for an additional heterologous polypeptide chain.

We tested the robustness of our AFDs definition by using sequence identities of 50 and 70 % (Table 1). These definitions yielded 338 and 296 AFD representatives, of which 59.1 % (200 of 338) and 57.8 % (174 of 296) were similar to a monomer, and 29.9 and 30.0 % of the domains formed homo-oligomers. These results suggest that the 30 % sequence identity is suitable for defining AFDs, not only because it yields a large dataset as expected, but also because the fraction of AFDs similar to monomeric proteins was the highest.

Assessment of inter-domain interactions using AFDs

We first investigated values for the distance threshold that best distinguished AFDs from non-AFDs. Independently of threshold distance used to calculate the hydrophobic clusters, AFDs formed less inter-domain hydrophobic clusters than non-AFDs ($p < 0.01$, Wilcoxon rank-sum test, Fig. 2a). The difference became statistically significant when the distance threshold was larger than 4.0 Å (Fig. 2b), which is close to the 4.663 Å adopted in [23]. Here, we chose 5.0 Å as the distance threshold for detecting the hydrophobic clusters, in line with our previous definition [20, 24]. According to this definition, AFDs

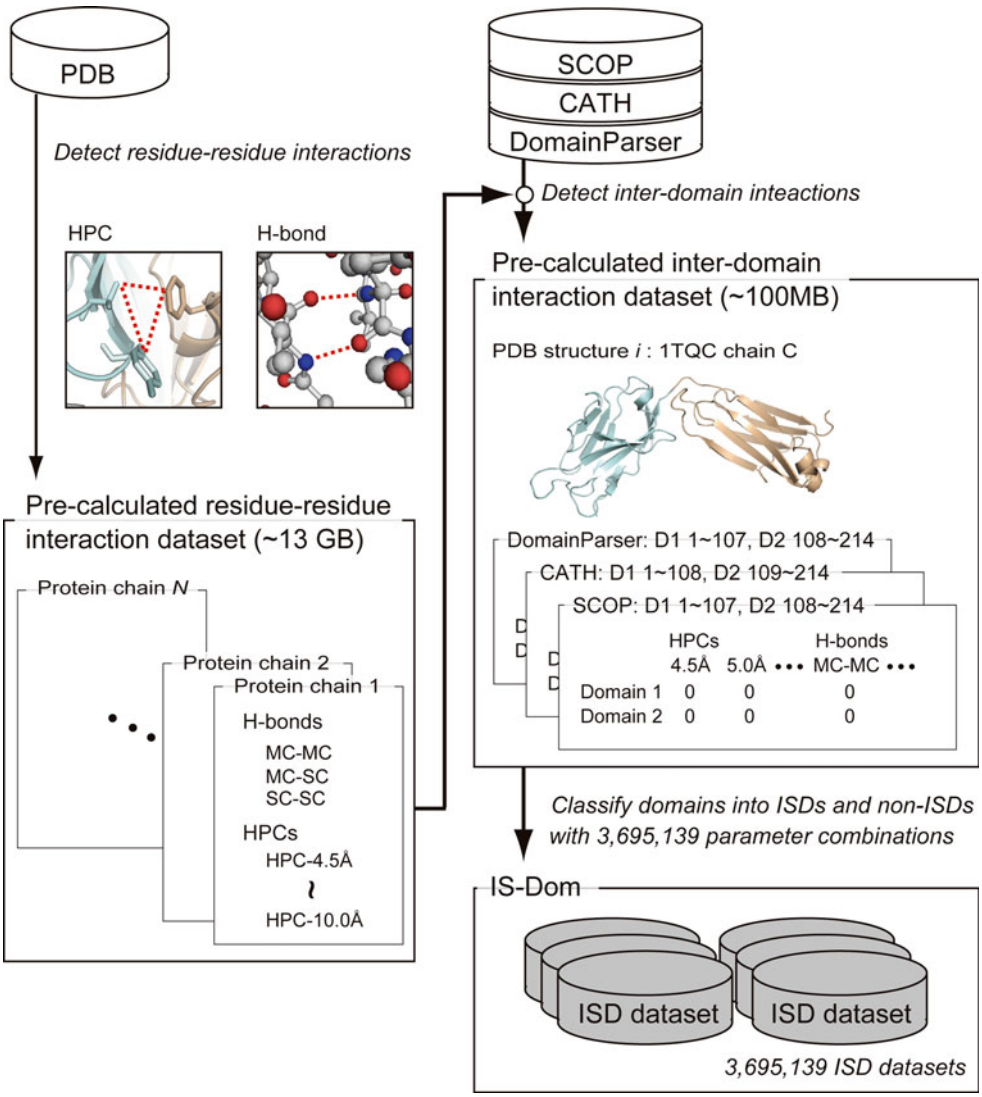

Fig. 1 Schematic of IS-Dom construction. We first detected H-bonds and hydrophobic clusters in 213,010 protein chains from 83,795 PDB files (left). In this calculation, hydrophobic clusters were detected using threshold distances of 1.0–10.0 Å with 0.5 Å step. The panel shows three side-chain carbon atoms of Val (left top, cyan), Phe(right, orange) and Tyr(left bottom, cyan) that are located within 5.0 Å, and form a hydrophobic cluster according to a 5.0 Å distance threshold. We then computed inter-domain interactions for SCOP, CATH and DomainParser defined domains, and identified domains that fulfilled the above ISD criteria (right). For example, 1TQC (chain C) is defined as a two-domain protein in SCOP with residues 1–107 and 108–214 corresponding to domain 1 and 2, respectively, and both domains qualify as ISD when calculated with the optimized parameters (see 'Methods' section and Supplemental Fig.S-1). The current version of IS-Dom includes 3,695,139 ISD datasets corresponding to the various parameter settings

Table 1 Number of AFDs defined with different identity threshold

Identity Threshold (%)[a]	Representative	Monomers (fraction)[b]	Homo-oligomer[c]	Hetero-oligomer[d]
30	460	290 (63.0 %)	139 (30.3 %)	31 (6.7 %)
50	338	200 (59.1 %)	101 (29.9 %)	37 (11 %)
70	296	174 (58.7 %)	89 (30.0 %)	33 (11.3 %)

[a] The AFDs were identified from SCOP 1.75A and CATH 3.5.0 multi-domain datasets with more than 30, 50 or 70 % sequence identity. [b] Monomers were defined as an AFD with >30, >50 or >70 % sequence identity to a monomeric protein having a single chain in the PDB file. Similarly, [c] homo- and [d] hetero-oligomer AFDs stands for domains that are similar to sequences with PDB files containing two or more chains (either same or different types)

formed 4.02 ± 0.20 (mean ± SEM.) inter-domain hydrophobic clusters, whereas it was 10.91 ± 0.42 for non-AFDs.

Similarly, AFDs formed less inter-domain H-bonds than non-AFDs. AFDs formed 1.49 ± 0.10 inter-domain MC–MC, 2.12 ± 0.13 MC–SC and 2.00 ± 0.18 SC–SC H-bonds, whereas, for non-AFDs, the corresponding values were 2.33 ± 0.05, 4.34 ± 0.08 and 3.75 ± 0.07 ($p < 0.01$, Wilcoxon rank-sum tests).

Finally, we also calculated the numbers of inter-domain disulfide bonds by DSSP [25] but did eventually not take them into account, because very few of them were detected. Namely, AFDs and non-AFDs formed 0.05 ± 0.02 and 0.04 ± 0.004 inter-domain disulfide bonds, respectively,

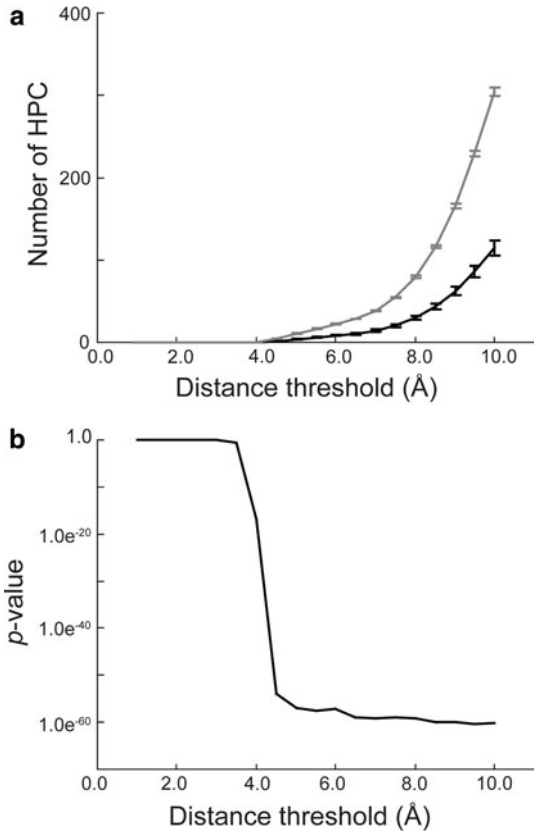

Fig. 2 Distance threshold for hydrophobic clusters. The number of inter-domain hydrophobic clusters (**a**) and the statistical significance of the differences (**b**) are shown as a function of the threshold distances. Black and gray lines indicate the number of interactions in AFDs and non-AFDs, respectively

which was not statistically significant ($p > 0.33$, Wilcoxon rank-sum test).

Optimization of interaction numbers and construction of a standard IS-Dom

In order to define a standard set of ISDs, we first analyzed the effects of the threshold numbers of inter-domain interactions on the ISD detection. The threshold inter-domain interaction numbers were optimized by maximizing the Youden's index (*YI* [22], see Methods), and yielded 7 hydrophobic clusters, 11 MC–MC, 9 MC–SC, and 6 SC–SC H-bonds (Fig. 3). The *YI*, R_{TP} and R_{FP} with the threshold numbers were 0.37, 0.83 and 0.46, respectively. The standard IS-Dom, defined using the above optimal parameters (with hydrophobic cluster distance threshold fixed to 5.0A) contained 54,860 ISDs (3,163 representatives) of which 14,017 (25.5 %) were AFDs. In the standard IS-Dom, 3,880, 8,651 and 17,972 ISDs were, respectively, selected from SCOP, CATH and DomainParser, and the remaining 24,357 were common to two or three datasets (Fig. 4).

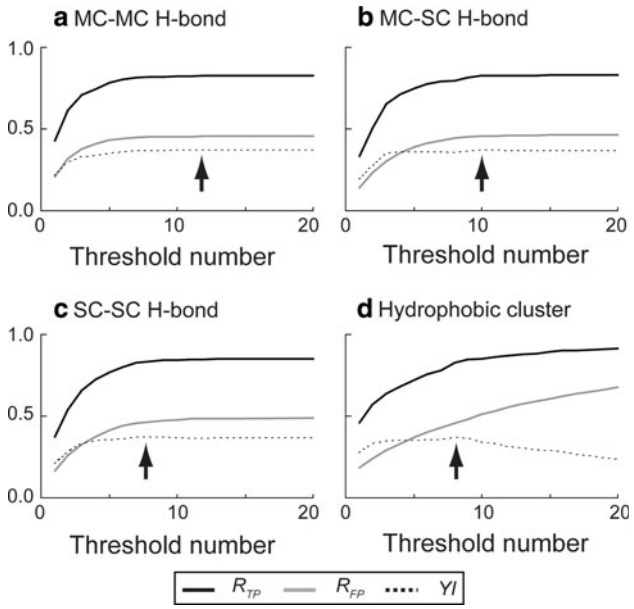

Fig. 3 Dependency of the R_{TP}, R_{FP} and *YI* on the threshold numbers. R_{TP}, R_{FP} and *YI* are shown as a function of the threshold numbers of inter-domain MC–MC (**a**), MC–SC (**b**), SC–SC H-bonds (**c**) and hydrophobic cluster (**d**). In each panel, the values were calculated by fixing the remaining parameters to their optimal values. R_{TP}, R_{FP} and *YI* are shown, respectively, in black, gray and dashed lines and the optimal threshold numbers are indicated by an arrow

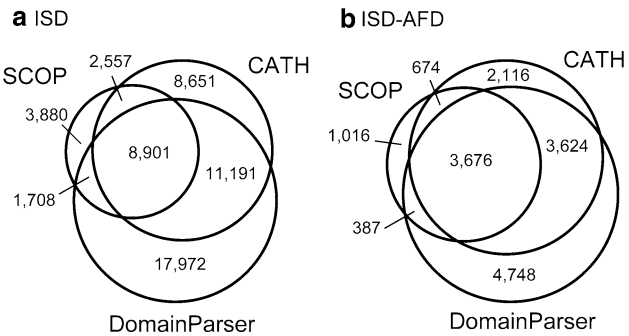

Fig. 4 Number of domains in the standard IS-Dom. **a** The Venn diagram represents the numbers of ISDs selected from SCOP, CATH and DomainParser. ISDs were defined using the optimized parameters. ISDs overlapping with 10-residue termini error were considered as identical. The total number of ISDs is 54,860. **b** Venn diagram for ISDs defined as AFDs. The conventions are the same as in **a**. Total number of ISDs defined as AFDs is 14,017

We assessed the robustness of our ISD's detection protocol and of the standard IS-Dom using a fivefold cross validation test and varying the interaction numbers (Table 2). The calculations were repeated five times by changing the testing and training dataset. Both the threshold numbers and the maximum *YI* values distributed around the mean values, as well as the optimized values calculated using all of the AFDs and non-AFDs (Table 2). The R_{TP} and R_{FP} ranged respectively, between 0.70–0.87

Table 2 Cross validation analysis of the ISD definition

Trial	Optimal threshold numbers				Indexes		
	MC–MC	MC–SC	SC–SC	HPC	*YI*	R_{TP}	R_{FP}
1	11	9	6	7	0.38	0.79	0.46
2	7	3	12	8	0.37	0.73	0.39
3	11	9	7	7	0.38	0.80	0.46
4	11	9	6	7	0.36	0.86	0.44
5	11	9	6	7	0.38	0.80	0.46
Mean ±SEM	10.2 ± 0.8	7.8 ± 1.2	7.4 ± 1.2	7.2 ± 0.2	0.37 ± 0.004	0.80 ± 0.02	0.44 ± 0.01

The optimal parameters, maximum *YI* values, R_{TP} and R_{FP} were calculated by a fivefold cross validation test using the control AFD and non-AFD datasets. HPC were calculated with a distance threshold of 5.0 Å

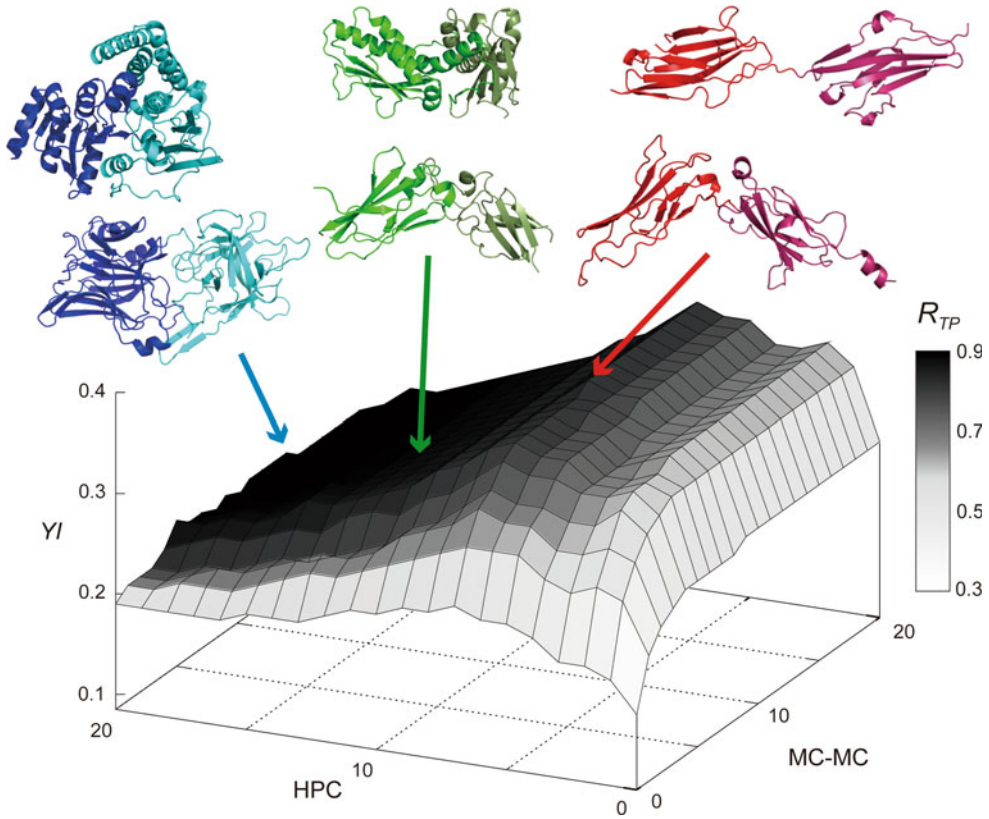

Fig. 5 3D Schematic illustration of the *YI* dependency on interaction numbers. The x and y axes indicate, respectively, the threshold numbers of hydrophobic cluster and MC–MC H-bond, and the z axis represents the *YI*. An equivalue plane calculated with the optimal values of MC–SC = 9 and SC–SC = 6 is shown. The plane is colored according to the R_{TP} values. Ribbon models of typical domains are shown on the top. Red domains fulfill the ISD definition using the optimal parameters (indicated by red yielded arrow) and are also AFD. Green and blue domains are defined, respectively, as non-AFD and non-ISD, according to the optimal parameters. However, they would be defined as ISD with non-optimal parameters indicated by the green ($YI = 0.29$ and $R_{TP} = 0.88$; HPC < 13, MC–MC < 9) and blue ($YI = 0.20$ and $R_{TP} = 0.95$; HPC < 16, MC–MC < 21) arrows. The red domains, which fulfill stringent conditions, are clearly far apart from each other unlike the domains colored blue

and 0.40–0.49 (Table 2). Furthermore, 3 of 5 cross validation trials yielded exactly the optimized interaction numbers. These results emphasize the robustness of the protocol and indicate that the optimal interaction numbers, but also nearly optimal ones, can be used for detecting ISDs, as illustrated in Fig. 5.

We evaluated our ISD detection protocol as well as the standard IS-Dom by calculating the fraction and the total number of AFDs in SCOP, CATH and DomainParser after and before the ISD selection. 90 % of AFDs contained in SCOP, CATH and DomainParser were retained by selecting ISD with the optimal threshold numbers (11 inter-

Table 3 Numbers and fraction of AFDs in the unfiltered and filtered datasets

	All domains[a]				ISDs[d]			
	SCOP	CATH	DP[b]	Total[c]	SCOP	CATH	DP[b]	IS-Dom[f]
Number of domains	37,037	63,232	85,182	128,522	17,046	31,300	39,772	54,860
Number of AFDs	6,372	10,090	12,435	16,175	5,753	9,139	11,258	14,017
(ISD & AFD)/AFD[e]	–	–	–	–	90.3 %	90.3 %	90.5 %	86.7 %
Fraction of AFDs	17.2 %	16.0 %	14.6 %	12.6 %	33.6 %	29.2 %	28.3 %	25.5 %

[a] All continuous domains in multi-domain proteins listed in SCOP, CATH or predicted using DomainParser

[b] *DP* DomainParser

[c] Total was constructed by removing sequence redundancy among the unfiltered databases with single linkage clustering

[d] ISDs were identified from the corresponding datasets using the optimized parameters (see Results and Discussion)

[e] (ISD & AFD) stands for ISDs that were identified as AFDs

[f] IS-Dom was constructed by removing sequence redundancy among the filtered datasets with single linkage clustering (see Methods section)

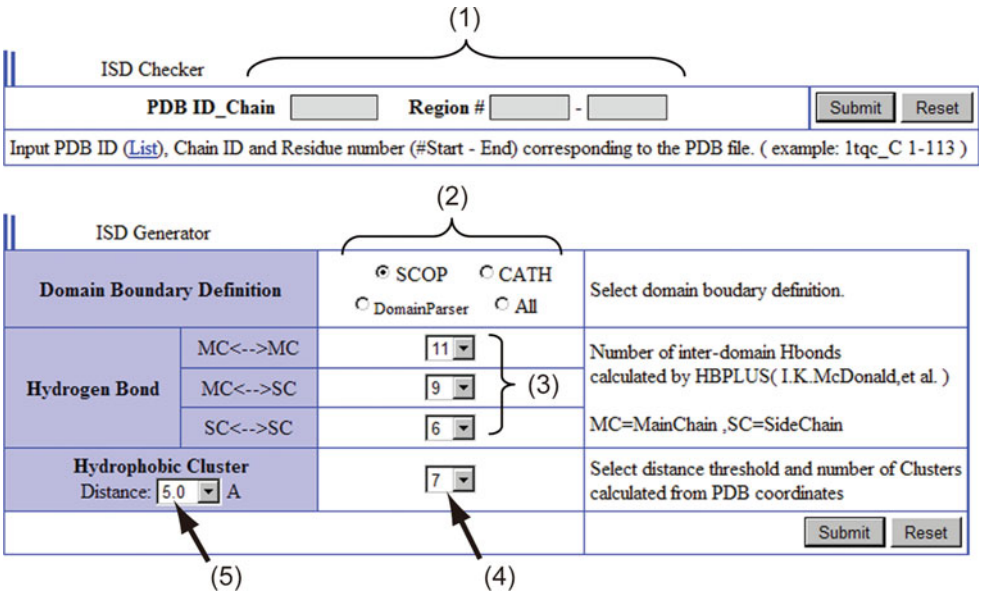

Fig. 6 Web interface of the IS-Dom server. The user can either assess the structural independence of a user's defined protein region and assess its suitability as an ISDs (*1*); or retrieve a set of ISDs by interactively adjusting the threshold parameters. The adjustable parameters are the datasets from which ISDs are derived (*2*), the threshold number of inter-domain H-bonds (*3*) and hydrophobic clusters (*4*). Users can also freely choose the distance threshold for detecting hydrophobic clusters (*5*)

domain MC–MC, 9 MC–SC, 6 SC–SC H-bonds and 7 hydrophobic clusters). Furthermore the fraction of AFDs in the filtered datasets increased by 12 % compared to the unfiltered domain databases (Table 3). Finally, ISD and AFD size distribution were very similar indicating that our ISD's detection protocol has no size bias (Fig. S-2). Overall, these results corroborate the efficiency and robustness of our algorithm for evaluating the domain's independent foldability from atomic coordinates of the full length protein.

IS-Dom server

We stored the full list of interactions (~13 GB) as well as the list of ISDs defined from SCOP, CATH and DomainParser for a wide range of interaction parameters (~100 MB) in our server. The user can either retrieve the standard IS-Dom dataset from the server, retrieve an IS-Dom dataset defined using the user's adjusted parameters, or test his own domain definition (Fig. 6). The adjustable parameters are the hydrophobic cluster's distance threshold, and the numbers of inter-domain H-bonds and hydrophobic clusters (Fig. 6). The current IS-Dom includes 3,695,139 ISD datasets (see Methods), which were constructed using different parameter combinations and each dataset is retrievable from the website.

The protocol for calculating ISDs is simple, automated, and computationally fast enough for calculating the structural independence and consequently estimating the foldability of a vast number of domains. The computational time for detecting residue–residue interactions (hydrophobic cluster and H-bonds) is about 2 s, and that to evaluate a

domain's foldability from a pre-calculated list of interactions is less than 3 s for a protein containing 700 amino acid residues. The fast ISD detection enabled to significantly increase the number of ISDs in IS-Dom from our previous dataset [20, 24] (containing 452 SCOP derived structural domains), not only because SCOP was updated, but also because CATH, and particularly DomainParser domain definitions were included in the construction of IS-Dom.

Conclusion

In conclusion, we developed a protocol for predicting independently foldable structural domains by quantitatively evaluating inter-domain interactions from the full length protein structure. The protocol's robustness and efficiency was corroborated by sequence comparison with proteins cataloged in the PDB. The application range of our protocol is not limited to filtering domains from existing domain databases and can be used for evaluating the foldability of domains in novel proteins. To this respect, IS-Dom, our dataset of ISDs, is expected to provide a consensus definition for a "structural domain" as independently foldable units [19], and enable their automatic identification based on quantitative structural criteria calculated from the full length protein's atomic coordinates. Future directions in IS-Dom updates could investigate possible improvements by including robust methods for detecting hydrophobic patches at the domain–domain interaction interfaces (such as those described in [26]).

Acknowledgments We thank Mr. Yuta Kumagai, Takao Arai, Tomohiro Furuyama, Shun Iwasaki and Ryotaro Tsuji (TUAT, Kuroda Lab) for their help with dataset construction. This work was funded by a Grant-in-aid from the Japanese Society for the Promotion of Science to Y.K. (JSPS-18500225).

References

1. Brenner SE (2000) Nat Struct Biol 7(Suppl):967
2. Jacobs SA, Podell ER, Cech TR (2006) Nat Struct Mol Biol 13(3):218
3. Hondoh T, Kato A, Yokoyama S, Kuroda Y (2006) Protein Sci 15(4):871
4. Vastermark A, Almen MS, Simmen MW, Fredriksson R, Schioth HB (2011) BMC Evol Biol 11:123
5. Chikayama E, Kurotani A, Tanaka T, Yabuki T, Miyazaki S, Yokoyama S, Kuroda Y (2010) BMC Bioinformatics 11:113
6. Andreeva A, Howorth D, Chandonia JM, Brenner SE, Hubbard TJ, Chothia C, Murzin AG (2008) Nucleic Acids Res 36(Database issue):D419
7. Greene LH, Lewis TE, Addou S, Cuff A, Dallman T, Dibley M, Redfern O, Pearl F, Nambudiry R, Reid A, et al (2007) Nucleic Acids Res 35(Database issue):D291
8. Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, Pang N, Forslund K, Ceric G, Clements J, et al (2012) Nucleic Acids Res 40(Database issue):D290
9. Miyazaki S, Kuroda Y, Yokoyama S (2002) J Struct Funct Genomics 2(1):37
10. Miyazaki S, Kuroda Y, Yokoyama S (2006) BMC Bioinformatics 7:323
11. Taylor WR (1999) Protein Eng 12(3):203
12. Swindells MB (1995) Protein Sci 4(1):103
13. Xu Y, Xu D, Gabow HN (2000) Bioinformatics 16(12):1091
14. Zhou H, Xue B, Zhou Y (2007) Protein Sci 16(5):947
15. Guo JT, Xu D, Kim D, Xu Y (2003) Nucleic Acids Res 31(3):944
16. Dumontier M, Yao R, Feldman HJ, Hogue CW (2005) J Mol Biol 350(5):1061
17. Siddiqui AS, Barton GJ (1995) Protein Sci 4(5):872
18. Rost B (1999) Protein Eng 12(2):85
19. Kuroda Y, Tani K, Matsuo Y, Yokoyama S (2000) Protein Sci 9(12):2313
20. Ebina T, Toh H, Kuroda Y (2011) Bioinformatics 27(4):487
21. McDonald IK, Thornton JM (1994) J Mol Biol 238(5):777
22. Youden WJ (1950) Cancer 3(1):32
23. Tanaka T, Yokoyama S, Kuroda Y (2006) Biopolymers 84(2):161
24. Ebina T, Toh H, Kuroda Y (2009) Biopolymers 92(1):1
25. Kabsch W, Sander C (1983) Biopolymers 22(12):2577
26. Goncalves-Almeida VM, Pires DE, de Melo-Minardi RC, da Silveira CH, Meira W, Santoro MM (2012) Bioinformatics 28(3):342