Automated search of natively folded protein fragments for high-throughput structure determination in structural genomics

YUTAKA KURODA,¹ KAZUTOSHI TANI,¹ YO MATSUO,¹ AND SHIGEYUKI YOKOYAMA^{1,2}

¹Protein Research Group, Genomic Sciences Center (GSC), The Institute of Physical and Chemical Research (RIKEN), 1-7-22, Suehiro-cho, Tsurumi, Yokohama, Kanagawa 230-0045, Japan

²Department of Biophysics and Biochemistry, Graduate School of Science, University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan

(RECEIVED July 18, 2000; FINAL REVISION September 8, 2000; ACCEPTED September 8, 2000)

Abstract

Structural genomic projects envision almost routine protein structure determinations, which are currently imaginable only for small proteins with molecular weights below 25,000 Da. For larger proteins, structural insight can be obtained by breaking them into small segments of amino acid sequences that can fold into native structures, even when isolated from the rest of the protein. Such segments are autonomously folding units (AFU) and have sizes suitable for fast structural analyses. Here, we propose to expand an intuitive procedure often employed for identifying biologically important domains to an automatic method for detecting putative folded protein fragments. The procedure is based on the recognition that large proteins can be regarded as a combination of independent domains conserved among diverse organisms. We thus have developed a program that reorganizes the output of BLAST searches and detects regions with a large number of similar sequences. To automate the detection process, it is reduced to a simple geometrical problem of recognizing rectangular shaped elevations in a graph that plots the number of similar sequences at each residue of a query sequence. We used our program to quantitatively corroborate the premise that segments with conserved sequences correspond to domains that fold into native structures. We applied our program to a test data set composed of 99 amino acid sequences containing 150 segments with structures listed in the Protein Data Bank, and thus known to fold into native structures. Overall, the fragments identified by our program have an almost 50% probability of forming a native structure, and comparable results are observed with sequences containing domain linkers classified in SCOP. Furthermore, we verified that our program identifies AFU in libraries from various organisms, and we found a significant number of AFU candidates for structural analysis, covering an estimated 5 to 20% of the genomic databases. Altogether, these results argue that methods based on sequence similarity can be useful for dissecting large proteins into small autonomously folding domains, and such methods may provide an efficient support to structural genomics projects.

Keywords: automated procedure; computing tool; PASS; protein structure determination; sequence similarity; stable protein fragment, structural genomics

Among the various approaches intended to exploit the results of large scale sequencing projects, ambitious structural genomic efforts are now being programmed (Gaasterland, 1998; Kim, 1998; Shapiro & Lima, 1998; Brenner et al., 1999; Burley et al., 1999; Elofsson & Sonnhammer, 1999; Mallick et al., 2000; Yokoyama et al., 2000). Structural genomic projects require the ability to determine protein structures in a routine manner, ultimately in the way DNA sequences are determined. For relatively small proteins, recent pilot experiments have assessed the feasibility of such projects (Kim, 1998; Shapiro & Lima, 1998; Burley et al., 1999; Hoseki

et al., 1999; Nakagawa et al., 1999). On the other hand, the structure determination of large and often biologically interesting proteins remains a time consuming and tedious process.

A simple strategy to overcome this issue and to obtain structural insight for large proteins at a rate similar to that already attained for smaller ones is to cleave the entire protein in small units that can fold into their initial structures in isolation (Campbell & Downing, 1994). This strategy has often been used for structures determined by NMR, as this technique has a molecular weight limit of ~25,000 Da (Clare & Gronenborn, 1997; Salzmann et al., 1998; Wüthrich, 1998). In addition to this limitation, technical difficulties can also arise during the overexpression and purification of large proteins, especially when refolding steps are required. The average number of residues in the proteins registered in the Brook-

Reprint requests to: Yutaka Kuroda, Protein Research Group, Genomic Sciences Center, RIKEN, 1-7-22, Suehiro-cho, Tsurumi, Yokohama, Kanagawa 230-0045, Japan; e-mail: kuroda@gsc.riken.go.jp.

haven Protein Data Bank (PDB) is only 215 residues, which reflects the technical challenges met when solving large protein structures by both NMR and X-ray crystallography.

Structural genomics projects that include the analysis of large proteins require a computing tool to automatically detect polypeptide segments that can fold in isolation, or autonomously folding units (AFU). Cleavage sites that produce stable protein fragments that fold in isolation may be experimentally determined by a limited proteolysis analysis (see Hubbard, 1998, for a review). On the other hand, stable fragments can also be anticipated when a portion of the amino acid sequence has many similar sequences, while the rest of the sequence has no or few similar sequences. This educated guess relies on the modular nature of large proteins and the conservation of biologically important domains among various organisms. Protein dissections based on such educated guesses have yielded many NMR structure of biologically significant protein domains (e.g., Baron et al., 1990; Yu et al., 1992; Kohda et al., 1993, 1996; Koshiba et al., 1999; Lugovskoy et al., 1999).

To develop an automated AFU prediction tool based on sequence similarities, and capable of genome-wide searches, we need to investigate three issues. First, the domains identified based on sequence similarities usually relate to functional units (Sonnhammer & Kahn, 1994; Sonnhammer et al., 1997; Corpet et al., 1998), rather than structural entities. Thus, we need to establish whether a prediction method based on sequence similarities can indeed identify AFUs. Then, we have to quantify the ability of this automated tool to identify putative AFUs. Last, we need to determine the conditions (parameters) under which the identification of AFUs is most efficiently automated.

We examined these issues by developing a program that screens the results of BLAST searches (Altschul et al., 1990, 1997) and determines segments that have large numbers of similar sequences. It is a simple and fast computing tool, and it is essentially free of adjustable parameters. A test calculation using 99 test protein sequences, containing 150 subsequences with solved structures, indicated that segments with high numbers of similar sequences correlate well with segments with solved structures. Finally, we applied our program to various sequence databases to confirm its ability to identify candidates for structural studies.

Results

Test with PDB sequences and determination of optimal parameters

Our program (PASS for Prediction of AFUs based on Sequence Similarities) has only two adjustable parameters (Fig. 1). These are the BLAST E-value cutoff and the number of similar sequences at each residue position required for detection (threshold). We optimized these two parameters for maximum sensitivity and reliability of the prediction. Thus, we first maximized the number of predicted AFUs to cover a large fraction of the genomic data (true positive). Second, we minimized the number of false positives, i.e., regions that are predicted to fold in isolation but do not.

We determined the fraction of correct predictions and of detected sequences for E-value cutoffs of 1.0, 10^{-7} , and 10^{-20} , and thresholds of 5, 10, and 20 sequences (Table 1). A natural tradeoff between reliability and sensitivity is observed (Fig. 2A). For example, a loose cutoff of 1.0 with a threshold of five sequences allows the detection of a large proportion of AFUs (37%), but with a low reliability (30%). A reasonable balance between these two

Table 1. Identification of segments similar to a PDB registered sequence^a

Threshold ^b (cutoff)	Nb of Frg ^c	Percent correct ^d	Percent detected ^e	Average Frg length ^f
5 (1.0)	205	29.7	36.6	160
10 (1.0)	178	30.1	34.0	156
20 (1.0)	138	31.2	26.0	151
5 (10^{-7})	146	34.9	32.7	160
10 (10^{-7})	114	37.7	28	164
20 (10^{-7})	80	45	23.3	168
5 (10^{-20})	113	42.5	30.7	170
10 (10^{-20})	79	44.4	20.7	184
20 (10^{-20})	58	46.7	17.3	184

^aThe calculations were carried out using 99 sequences containing 150 segments similar (E-value 10^{-20} , identity 40%) to PDB registered sequences and shorter than 300 residues. The average length of the analyzed sequences was 632 amino acids.

^bMinimum number of similar sequences required for fragment identification. E-value cutoff used for BLAST searches are indicated into parenthesis.

^cNumber of predicted fragments.

^dPercentage of predicted fragments that correspond to PDB-similar sequences with termini error of 20 amino acids.

^ePercentage of PDB sequences detected using the specified parameters.

^fAverage length of the predicted fragments.

numbers can be reached with E-values smaller than 10^{-7} , with 5 or 10 similar sequences used for predictions of AFUs (28% of test sequences detected with a 38% success rate). For sequences with more than 50 similar sequences with E-values better than 10^{-7} , the chances that they are AFUs are over 50% (Fig. 2A).

Overall, our calculations indicate that up to a third of the 150 test PDB sequences are detected (Fig. 2). When our method did not detect a PDB sequence, it was usually because the number of similar sequences was insufficient. In a few other cases, similar sequences were found, but our program (Fig. 3) did not recognize the distorted shapes of the elevations. On the other hand, our method also predicted AFUs that do not have a structure listed in the PDB. However, the predicted fragments might nonetheless be structured and their structures solved in the future. Thus, predicted AFUs without PDB structures are not necessarily "false positives" but rather "unknown positives," and the actual number of false positives is probably smaller than that listed in Table 1.

Accuracy of the predicted fragment termini and consideration for primer design

From an experimental standpoint, it is particularly important to estimate the error window of the prediction. Experimentally, when a fragment is determined with termini errors of ± 20 residues, its ability to fold can be roughly tested with three DNA constructs starting -20 , 0 , and $+20$ residues apart from the predicted start or end. Combination of three constructs for each terminus will result in nine plasmids for each predicted fragment, which yields a manageable number of samples. When more than 20 similar sequences with E-values better than 10^{-7} are required for fragment identification, over 40% of the predicted segments correspond to a PDB fragment with a termini error of ± 10 residues (Fig. 2B). This

Fig. 1. Flowchart of the PASS program suite. Step zero is a BLAST comparison of a query sequence with the NCBI nonredundant database (Benson et al., 2000). We then compute the number of similar sequences at each position of the query sequence. A graphical representation shows that the plot has a typical rectangular shape at conserved polypeptide segments. Step two uses this characteristic and identifies delimiters at positions where large relative increase or decrease occur. Similarity start and end delimiters are indicated by green and red vertical bars, respectively. Orange bars indicate overlapping start and end delimiters. In step three, fragments are identified. Segments delimited by both a start and end delimiters are indicated in gray. Light blue is used for regions detected either between two start or end delimiters. Finally, the resulting fragments are listed according to the number of changes of similar sequences at the delimiter site. Fragments with changes larger than a given threshold are selected as candidate AFUs.

Fig. 2. A: Fraction of PDB-similar segments detected by our program (PASS) vs. the minimum number of similar sequences required for detection (broken lines). Continuous line indicates the fraction of fragments identified by our program that correspond to PDB-similar segments. Fragments termini errors are set to ± 20 residues. Fractions are computed using similar sequences with E-values better than 1.0 (red), 10^{-7} (green), and 10^{-20} (blue). Sequences listed in PDB were excluded from the calculations. **B:** Same fractions as in A but plotted as a function of the termini error. The minimum number of similar sequences required for detection is fixed to 20 sequences.

percentage increases slightly when the termini error is increased to 20 residues (Fig. 2B; Table 1).

Test with SCOP domains

The application of PASS to the sequences listed in SCOP provides an estimate for the presence of false positive predictions, since the structure of the entire sequence is known. Additionally, a test with

SCOP sequences also assesses the ability of our program to detect small AFUs. Altogether, 15 to 20% of the SCOP domain boundaries are detected by the present method, and the predicted linkers correspond to domain boundaries with a probability of over 50% (Fig. 4; Table 2). Analogous to the results obtained with the PDB data set, we observe a tradeoff between the fraction of SCOP domain boundaries that are identified, and the fraction of correctly predicted linkers. The result of our test run with the SCOP sequences (Table 2) implies that the fraction of false positives is indeed lower than the values found with the PDB data set (Table 1). However, the fraction of SCOP domain boundaries predicted by our method remains modest: 10 to 20%, i.e., only 8 to 17 linkers, out of 75, are predicted within an error of ± 20 residues (Fig. 4A).

Test with sequences from various organisms

It is generally recognized that chimeric proteins are more likely to be found in higher order organisms. Because our method relies on the shuffling of independent domains, we wanted to assess whether it could detect a sufficient number of foldable fragments in sequences from lower order organisms. We thus applied PASS to sequences from *BmNPV* (Gomi et al., 1999), and to randomly chosen SWISS PROT (Bairoch & Apweiler, 1996) sequences from *Escherichia coli*, *Thermus thermophilus*, *Caenorhabditis elegans*, mouse (*Musculus*), and human (*Sapiens*). Although more AFUs are detected in higher order organisms (around 20% of the sequences covered in mouse and human), a significant number (5 to 10% in *E. coli*) of AFUs are also detected in lower order organisms (Table 3).

Discussion

Protein dissection has often been used to determine the structures of biologically significant domains (see Campbell & Downing, 1994, for a review). However, the determination of a sequence segment that forms a natively structured protein fragment has remained an art. Often, the structured fragments have been identified based on the premise that conserved sequences form domains. To extend this intuitive idea to a tool capable of analyzing whole genome data, it needed to be refined and quantitatively substantiated. Especially because residues are conserved for both functional and structural grounds, one could imagine that many, if not most, domains identified based on sequence similarity would not fold in isolation. Thus, our calculation with sequences containing polypeptide segments related to PDB structures statistically confirms the correlation between segments with a large number of similar sequences and segments that fold in isolation. For the purpose of discussion, we refer to these domains as autonomously folding units (AFU).

We originally assumed that domain shuffling would create a difference in the number of similar sequences along the query polypeptide chain. Indeed, some regions along the amino acid sequence have many similar sequences, whereas others have few or none (Fig. 3). However, early in our study, we noticed that segments between two end or start delimiters (indicated in blue in Fig. 3) are less likely to have a solved structure than those defined by a start and an end delimiter. For example, all of the PDB sequences shown in Figure 3 correspond to segments defined by a start and an end delimiter. None of them are identified by blue area. The reason for this discrepancy is not clear. One hypothesis is that regions identified by either two start or two end delimiters are less reliably iden-

Fig. 3. Three examples of similarity plots for sequences containing segments similar to PDB sequences: (A) glucose specific III ABC compound (gi:131493), (B) FPK-type peptidyl-prolyl *cis-trans* isomerase (gi:3979950), and (C) promyelocytic leukemia zinc finger protein (gi:1172536). Color codes are the same as in Figure 1. E-value cutoff for sequence similarity plots are set to 1.0, 10^{-7} , and 10^{-20} from top to bottom in each panel. Segments with sequences similar to PDB registered sequences are indicated by blue horizontal bars (E-value below 10^{-20} and over 40% identity over the entire PDB sequence length). **A:** Segment 539–699 of gi:131493 has a structure registered in the PDB (1GPR); PASS detects an AFU corresponding to this segment and extending from residue 549 to 694. No structure has yet been reported for a sequence similar to segment 441–517. However, this fragment is not detected when the cutoff value is raised to 10^{-7} and is thus subject to caution. **B:** The C-terminal fragment of gi:3979950 has a 58% identity (E-value of 10^{-54}) with the PDB sequence 1BKF. Segment 80–191 is another AFU candidate, but it is not detected when the cutoff is raised to 10^{-7} . **C:** A candidate AFU is detected at the N-terminus of gi:1172536 (residues 1–97). It corresponds to a PDB similar segment extending from residue 1–126, which is similar to the BTB/POZ domain of PLZP (57% identity with 1BUO, E-value = 10^{-54}). This AFU is identified despite the small number of similar sequences detected at the N-terminal. Another AFU is identified at the C terminus (401–653), but the closest segment is a PDB-similar segment extending from 495 to 600. This segment is similar to 1UBD, a zinc finger protein (identity 59% but E-value of 10^{-19}). Though this similarity is probably meaningful, our method is unable to identify this PDB fragment within an error window of ± 30 residues. Note that the profile at the C-terminus is broad and has not a typical shape. This region is barely recognized by PASS because there is no flat plateau at the top of the elevation.

tified, as they necessarily have a smaller number of similar sequences than regions between a start and an end delimiter. For example, while gray areas are constantly identified in the sequence of prolyl *cis-trans* isomerase for E-values cutoffs between 10^{-20} and 1.0, the identification of a blue area depends on the cut-off E-value used (Fig. 2B). Thus, as a first approximation, we decided to account only for regions delimited by one start and one end.

By changing the cut-off E-value, we also observed that including remote similarities decreases the percentage of successful detection without dramatically improving the coverage rate (Table 1). AFUs are more reliably identified with a cutoff E-value of 10^{-7} or 10^{-20} than with a loose cutoff of 1.0. Presumably, the regions detected by including remote similarities identify subsegments of AFUs that are conserved for functional reasons but are not able to fold in isolation.

Our choice of PDB registered structures for AFUs implies that they are not necessarily minimum folding units, as PDB structures may contain several domains. Hence, the AFUs predicted by PASS may include multiple smaller domains, as defined by SCOP (Hub-

bard et al., 1999), FSSP (Holm & Sander, 1998), or CATH (Orengo et al., 1994, 1997, 1999). However, relating AFUs to PDB sequences guarantees that the isolated sequence will fold into a native protein structure, which is small enough for convenient structure determination; the average length of the predicted AFUs is around 165 residues (Table 1).

Nonetheless, because small protein structures are easier to determine, we wanted to know whether our program can detect the small domains listed in SCOP (Murzin et al., 1995; Hubbard et al., 1999). Approximately 50% of the predicted linkers are actual linkers defined in SCOP (Table 2), and this percentage rises to 75% for a threshold of 50 sequences (Fig. 4). On the other hand, our program failed to detect several SCOP linkers (Table 2). In addition, many of the correctly predicted linkers are detected by small variations in the number of similar sequences (Fig. 5A–C), which occur on top of a large number of similar sequences covering a wider portion of the query sequence (Fig. 5A). Altogether, PASS appears to detect AFUs with sizes appropriate for structure determination by NMR, but the inclusion of information from, for ex-

Fig. 4. A: Fraction of detected domain linkers (total of 75 SCOP domain boundaries in 52 nonredundant sequences; broken lines); and fraction of predicted delimiters found in SCOP (continuous lines) as a function of the number of similar sequences required for detection (threshold). Sequences with similarity better than E-values of 1.0 (red), 10^{-7} (green), and 10^{-20} (blue) are used for calculating the delimiters. PDB sequences are not included in this calculation. The fragment termini error is fixed to ± 20 residues. **B:** Fraction of identified SCOP domain boundary (broken line), and of predicted delimiters corresponding to a SCOP domain linker (continuous lines) as a function of the error allowed between the delimiters and the SCOP boundary. Sequences with similarity better than E-values of 1.0 (red), 10^{-7} (green), and 10^{-20} (blue) were used for calculating the delimiters. The minimum number of similar sequences required for detection is fixed to 20 sequences. PDB sequences are not used in the calculation.

ample, multiple-sequence alignment might improve the prediction reliability for small SCOP-size domains.

In summary, we developed and analyzed a program, which searches AFU candidates for structural analyses from whole genome sequence libraries. Our method is much simpler and com-

Table 2. Identification of SCOP domain linkers^a

Threshold (cutoff) ^b	Nb of predicted linkers ^c	Percent correct ^d	Percent detected ^e
5 (1.0)	65	41.3	21.3
10 (1.0)	56	42.8	18.7
20 (1.0)	42	52.5	14.4
5 (10^{-7})	42	35.7	18.7
10 (10^{-7})	31	48.4	17.3
20 (10^{-7})	22	50.0	12.0
5 (10^{-20})	21	47.6	12.0
10 (10^{-20})	18	50.0	10.7
20 (10^{-20})	14	57.1	9.3

^aThe average length of the 52 analyzed sequence is 405 amino acids.

^bMinimum number of similar sequences required for fragment identification. The E-value cutoff used for BLAST search is indicated into parenthesis.

^cTotal number of predicted delimiters.

^dPercent of predicted domain linkers that correspond to a SCOP domain boundary within an error of ± 20 amino acids.

^ePercentage of the 75 SCOP domain linkers detected by the present method.

Table 3. Fragment detection in sequences from various species

Species ^a	Nb of sequence (average length) ^b	Threshold (cutoff) ^c	Nb of Frg (PDB) ^d	Average Frg length ^e	% covered ^f
<i>BmNPV</i>	49 (483)	20 (1.0)	4 (1)	218	3.7
		5 (10^{-7})	2 (1)	248	2.1
		20 (10^{-7})	2 (1)	248	2.1
<i>E. coli</i>	112 (478)	20 (1.0)	23 (7)	189	7.1
		5 (10^{-7})	35 (7)	222	11.6
		20 (10^{-7})	14 (7)	210	5.1
<i>C. elegans</i>	154 (603)	20 (1.0)	87 (5)	166	10.5
		5 (10^{-7})	83 (4)	203	12.7
		20 (10^{-7})	52 (3)	197	8.5
<i>Musculus</i>	144 (655)	20 (1.0)	166 (19)	173	19.1
		5 (10^{-7})	123 (15)	203	20.0
		20 (10^{-7})	92 (12)	203	16.1
<i>Sapiens</i>	153 (643)	20 (1.0)	165 (29)	169	17.8
		5 (10^{-7})	147 (32)	198	19.5
		20 (10^{-7})	105 (29)	198	14.7

^aSequences from *E. coli*, *C. elegans*, mouse, and *Homo sapiens* were randomly selected from the SWISS-PROT data bank (Bairoch & Apweiler, 1996). Sequences from *Bombix mori* nucleopolyhedrovirus were from Gomi et al. (1999).

^bNumber of analyzed sequences with their average length indicated in parenthesis.

^cMinimum number of similar sequences and E-values used for fragment detection.

^dNumber of AFU candidates identified using the respective E-values and threshold values. The numbers of fragments similar to a sequence with known structures (registered in PDB) are indicated into parenthesis.

^eAverage length of the predicted AFUs.

^fPercent of amino acid contained in predicted fragments vs. the total number of amino acids in the database.

Fig. 5. Examples of SCOP domain boundaries detection. Schematic drawings (MOLSCRIPT; Kraulis, 1991) of the molecules are shown on the left side of the panel. SCOP domain boundaries are indicated in blue. Domains as defined by SCOP are differentiated by increasing intensities of red. Domain delimiters detected by our program are indicated in yellow. A similarity plot with a BLAST E-value cutoff of 10^{-7} of the corresponding sequence is shown at the right of the panel. The color code is the same as in Figure 1. Regions corresponding to SCOP domain linkers are shown by blue horizontal bars at the bottom of each graph. **A:** Structure and domain boundary of Gamma-B crystallin protein (1AMM). A linker extending from residues 81–88 binds the two SCOP domains. Our program locates a domain boundary at residue 86. **B:** Cyclodextrin glucanotransferase (1PAM; B-chain) is classified as a four-domain protein with three domain linkers at 365–385, 493–499, and 582–585. The similarity plot detects only the second linker and predicts a linker at 473–480 (shown in yellow). **C:** Endoexocellulase (1TF4; B-chain) is a two-domain protein with a SCOP domain linker encompassing residue 445 to 462. The similarity plot reveals a linker at residues 433–450, which is located at the end of an α -helix, and at the beginning of the SCOP domain boundary loop.

putationally less demanding than previous programs aimed at finding protein domains on the basis of sequence similarity (Sonnhammer & Kahn, 1994; Sonnhammer et al., 1997; Corpet et al., 1998; Gracy & Argos, 1998; Nicodeme, 1998; Henikoff et al., 1999). Although our program is not suited for producing a comprehensive list of domains (Sonnhammer et al., 1997; Corpet et al., 1998; Gracy & Argos, 1998; Henikoff et al., 1999), it yields a sizable number of AFUs candidates for high-throughput structure determination. We believe that combined to experimental methods, this procedure will provide an efficient tool to dissect large proteins and enhance our understanding of their structures.

Methods

Similarity plot of the query sequence

Our program suite, PASS, first rearranges (gapped) BLAST outputs (Altschul et al., 1990, 1997) to plot the number of similar sequences at each position of the query sequence (Fig. 5, step 1). When we visually inspected several “similarity plot,” we noticed

that most increase (or decrease) in the number of similar sequences occur abruptly. After this sharp increase, the number of similar sequences usually remains relatively constant. This abrupt change translates in rectangular shaped elevation in the similarity plot. Positions at the start or end of elevations are identified as delimiters. Specifically, a residue is a delimiter when (1) the change in the number of similar sequences at this position is larger than 20% of the maximum of the corresponding elevation, and (2) the slope of the plateau following the change is less than 10% over 30 residues (Fig. 1, step 2).

In the next step, we identify AFUs candidates defined by the delimiters. Considering all possible combinations of delimiters would yield a large and unmanageable number of AFU candidates (Fig. 5, step 3). We have thus introduced two limitations. The first is to consider only fragments shorter than 300 amino acids. Second, when delimiters are separated by less than 15 residues (fixed parameter), the delimiter yielding the longest fragment is selected, provided the fragment is shorter than 250 residues.

Candidate AFUs are then classified according to the (absolute) numbers of changes observed at the delimiting position

(Fig. 1, step 4). This provides a second selection that is based on absolute change. Thus, both relative and absolute changes are considered in the selection process of AFUs. This allows a reliable and sensitive detection of fragments that is maintained even when one region has a number of similar sequences much larger than another region.

Test with PDB registered sequences

The PDB was checked for sequences that are part of a longer amino acid sequences. First, we clustered PDB sequences into 1,400 PDB nonredundant representative sequences (according to a BLAST E-value cutoff of 10^{-7}). Then, we compared the representative sequences to the SWISS-PROT database (Bairoch & Apweiler, 1996). We found 99 amino acid sequences containing one or more sequences similar to a PDB registered sequence (E-value smaller than 10^{-20} and sequence identity $>40\%$), and that fulfilled two conditions. First, the length of the entire protein is longer than twice the PDB sequence length. Second, the sequence similarity encompasses the entire PDB registered sequences (we required the similarity to start and end within 10 residues of the PDB sequence start and end, respectively). Following the above definition, 150 "PDB-similar sequences" contained in the 99 amino acid sequences were selected for the test run.

The 99 sequences were then compared to the NCBI nonredundant data set (Benson et al., 2000) using a gapped BLAST search. Using the result of this search, we identified delimiters and fragments as described in the previous section. PDB sequences were not included for calculating the delimiters to prevent any possible bias. A PDB sequence was accounted as detected by our program when both its start and end corresponded to, respectively, the start and end of a predicted fragment (within an error window).

Test with SCOP domains

To test for the detection of smaller fragments, we applied our program to protein sequences with domains listed in the SCOP database (Murzin et al., 1995). Many SCOP domains relate to evolutionary or functional domains that may or may not fold in isolation. We thus choose 52 nonredundant sequences containing 75 domain linkers that are most likely to separate domains that fold independently. The criteria used to select the sequences are that (1) the domain linker is a loop as defined by DSSP (Kabsch & Sander, 1983); and (2) the domains are globular, when inspected visually. Here too, PDB sequences were excluded from the calculations to prevent possible biases.

Acknowledgments

We thank Dr. H. Toh (BERI, Osaka) for discussion and advice, S. Miyazaki for visual inspection of SCOP linkers and discussion, and members of the Protein Research Group for discussion. PASS can be accessed at www.bio.gsc.riken.go.jp.

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol* 215:403–410.
Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res* 25:3389–3402.

Bairoch A, Apweiler R. 1996. The SWISS-PROT protein sequence data bank and its new supplement TREMBL. *Nucleic Acids Res* 24:21–25.
Baron M, Norman D, Willis A, Campbell ID. 1990. Structure of the fibronectin type 1 module. *Nature* 345:642–646.
Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Rapp BA, Wheeler DL. 2000. GenBank. *Nucleic Acids Res* 28:15–18.
Brenner SE, Barken D, Levitt M. 1999. The PRESAGE database for structural genomics. *Nucleic Acids Res* 27:251–253.
Burley SK, Almo SC, Bonanno JB, Capel M, Chance MR, Gaasterland T, Lin D, Sali A, Studier FW, Swaminathan S. 1999. Structural genomics: Beyond the human genome project. *Nat Genet* 23:151–157.
Campbell ID, Downing AK. 1994. Building protein structure and function from modular units. *Trends Biotechnol* 12:168–172.
Clare GM, Gronenborn AM. 1997. NMR structures of proteins and protein complexes beyond 20,000 M(r). *Nat Struct Biol Suppl* 4:849–853.
Corpet F, Gouzy J, Kahn D. 1998. The ProDom database of protein domain families. *Nucleic Acids Res* 26:323–326.
Elofsson A, Sonnhammer EL. 1999. A comparison of sequence and structure protein domain families as a basis for structural genomics. *Bioinformatics* 15:480–500.
Gaasterland T. 1998. Structural genomics: Bioinformatics in the driver's seat. *Nat Biotechnol* 16:625–627.
Gomi S, Majima K, Maeda S. 1999. Sequence analysis of the genome of *Bombyx mori* nucleopolyhedrovirus. *J Gen Virol* 80:1323–1337.
Gracy J, Argos P. 1998. Automated protein sequence database classification. I. Integration of compositional similarity search, local similarity search, and multiple sequence alignment. *Bioinformatics* 14:164–173.
Henikoff S, Henikoff JG, Pietrokovski S. 1999. Blocks+: A non-redundant database of protein alignment blocks derived from multiple compilations. *Bioinformatics* 15:471–479.
Holm L, Sander C. 1998. Touring protein fold space with Dali/FSSP. *Nucleic Acids Res* 26:316–319.
Hoseki J, Yano T, Koyama Y, Kuramitsu S, Kagamiyama H. 1999. Directed evolution of thermostable kanamycin-resistance gene: A convenient selection marker for *Thermus thermophilus*. *J Biochem (Tokyo)* 126:951–956.
Hubbard SJ. 1998. The structural aspects of limited proteolysis of native proteins. *Biochim Biophys Acta* 1382:191–206.
Hubbard TJ, Ailey B, Brenner SE, Murzin AG, Chothia C. 1999. SCOP: A structural classification of proteins database. *Nucleic Acids Res* 27:254–256.
Kabsch W, Sander C. 1983. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 22:2577–2637.
Kim SH. 1998. Shining a light on structural genomics. *Nat Struct Biol Suppl* 5:643–645.
Kohda D, Hatanaka H, Odaka M, Mandiyan V, Ullrich A, Schlessinger J, Inagaki F. 1993. Solution structure of the SH3 domain of phospholipase C-gamma. *Cell* 72:953–960.
Kohda D, Morton CJ, Parkar AA, Hatanaka H, Inagaki FM, Campbell ID, Day AJ. 1996. Solution structure of the link module: A hyaluronan-binding domain involved in extracellular matrix stability and cell migration. *Cell* 86:767–775.
Koshiba S, Kigawa T, Iwahara J, Kikuchi A, Yokoyama S. 1999. Solution structure of the Eps15 homology domain of a human POB1 (partner of RalBP1). *FEBS Lett* 442:138–142.
Kraulis PJ. 1991. MOLSCRIPT: A program to produce both detailed and schematic plots of protein structures. *J Appl Crystallogr* 1991:946–950.
Lugovskoy AA, Zhou P, Chou JJ, McCarty JS, Li P, Wagner G. 1999. Solution structure of the CIDE-N domain of CIDE-B and a model for CIDE-N/CIDE-N interactions in the DNA fragmentation pathway of apoptosis. *Cell* 99:747–755.
Mallick P, Goodwill KE, Fitz-Gibbon S, Miller JH, Eisenberg D. 2000. Selecting protein targets for structural genomics of *Pyrobaculum aerophilum*: Validating automated fold assignment methods by using binary hypothesis testing. *Proc Natl Acad Sci USA* 97:2450–2455.
Murzin AG, Brenner SE, Hubbard T, Chothia C. 1995. SCOP: A structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol* 247:536–540.
Nakagawa N, Sugahara M, Masui R, Kato R, Fukuyama K, Kuramitsu S. 1999. Crystal structure of *Thermus thermophilus* HB8 UvrB protein, a key enzyme of nucleotide excision repair. *J Biochem (Tokyo)* 126:986–990.
Nicodeme P. 1998. SSMAL: Similarity searching with alignment graphs. *Bioinformatics* 14:508–515.
Orengo CA, Jones DT, Thornton JM. 1994. Protein superfamilies and domain superfolds. *Nature* 372:631–634.
Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM. 1997. CATH—A hierarchic classification of protein domain structures. *Structure* 5:1093–1108.

- Orengo CA, Todd AE, Thornton JM. 1999. From protein structure to function. *Curr Opin Struct Biol* 9:374–382.
- Salzmann M, Pervushin K, Wider G, Senn H, Wüthrich K. 1998. TROSY in triple-resonance experiments: New perspectives for sequential NMR assignment of large proteins. *Proc Natl Acad Sci USA* 95:13585–13590.
- Shapiro L, Lima CD. 1998. The Argonne Structural Genomics Workshop: Lamaze class for the birth of a new science. *Structure* 6:265–267.
- Sonnhammer EL, Eddy SR, Durbin R. 1997. Pfam: A comprehensive database of protein domain families based on seed alignments. *Proteins* 28:405–420.
- Sonnhammer EL, Kahn D. 1994. Modular arrangement of proteins as inferred from analysis of homology. *Protein Sci* 3:482–492.
- Wüthrich K. 1998. The second decade—Into the third millennium. *Nat Struct Biol Suppl* 5:492–495.
- Yokoyama S, Hirota H, Kigawa T, Yabuki T, Shirouzu M, Terada T, Ito Y, Matsuo Y, Kuroda Y, Nishimura Y, et al. 2000. Structural genomics projects in Japan. *Nat Struct Biol* 7 Suppl:943–945.
- Yu H, Rosen MK, Shin TB, Seidel-Dugan C, Brugge JS, Schreiber SL. 1992. Solution structure of the SH3 domain of Src and identification of its ligand-binding site. *Science* 258:1665–1658.