

# CHOP Proteins Into Structural Domain-Like Fragments

Jinfeng Liu<sup>1,3,4\*</sup> and Burkhard Rost<sup>1,2,3\*</sup>

<sup>1</sup>CUBIC, Department of Biochemistry and Molecular Biophysics, Columbia University, New York, New York

<sup>2</sup>Columbia University Center for Computational Biology and Bioinformatics (C2B2), New York, New York

<sup>3</sup>North East Structural Genomics Consortium (NESG), Department of Biochemistry and Molecular Biophysics, Columbia University, New York, New York

<sup>4</sup>Department of Pharmacology, Columbia University, New York, New York

**ABSTRACT** We developed a method CHOP dissecting proteins into domain-like fragments. The basic idea was to cut proteins beginning from very reliable experimental information (PDB), proceeding to expert annotations of domain-like regions (Pfam-A), and completing through cuts based on termini of known proteins. In this way, CHOP dissected more than two thirds of all proteins from 62 proteomes. Analysis of our structural domain-like fragments revealed four surprising results. First, >70% of all dissected proteins contained more than one fragment. Second, most domains spanned on average over ~100 residues. This average was similar for eukaryotic and prokaryotic proteins, and it is also valid—although previously not described—for all proteins in the PDB. Third, single-domain proteins were significant longer than most domains in multidomain proteins. Fourth, three fourths of all domains appeared shorter than 210 residues. We believe that our CHOP fragments constituted an important resource for functional and structural genomics. Nevertheless, our main motivation to develop CHOP was that the single-linkage clustering method failed to adequately group full-length proteins. In contrast, CLUP—the simple clustering scheme CLUP introduced here—succeeded largely to group the CHOP fragments from 62 proteomes such that all members of one cluster shared a basic structural core. CLUP found >63,000 multi- and >118,000 single-member clusters. Although most fragments were restricted to a particular cluster, ~24% of the fragments were duplicated in at least two clusters. Our thresholds for grouping two fragments into the same cluster were rather conservative. Nevertheless, our results suggested that structural genomics initiatives have to target >30,000 fragments to at least cover the multi-member clusters in 62 proteomes. *Proteins* 2004; 55:678–688. © 2004 Wiley-Liss, Inc.

**Key words:** genome sequence analysis; protein domains; automatic sequence clustering; protein structure; structural genomics

## INTRODUCTION

### Domains Are the Structural Units of Proteins

Although <1% of the proteins in entirely sequenced archae and prokaryotes is longer than 1000 residues, >7% of the proteins in the six entirely sequenced eukaryotes (yeast, fly, worm, weed, human, and mouse) constitute such large macromolecules.<sup>9,10</sup> Undoubtedly, all these large proteins are built of more than one domain. The term “domain” is not well defined: many biologists refer to any consecutive segment, such as a coiled-coil helix or a nuclear localization signal as a domain. Structural biologists view domains as semi-independent three-dimensional (3D) subunits that are compact and may fold independently.<sup>11–18</sup> Some structural domains appear to constitute units that evolve independently.<sup>15,19–24</sup> Methods that automatically assign structural domains from 3D coordinates identify domains through their compactness.<sup>3,7,25–29</sup> Structural domains are often related to particular functions. Hence, the domain organization of a protein may contain information crucial for understanding structure and function. Structural biologists also care about guessing the domain organization before experimentally solving the structure of a protein, because crystallization is more likely to succeed when expressing fragments that constitute domains,<sup>17</sup> and because NMR spectroscopy—limited by protein length—is more likely to unravel

*Abbreviations:* 3D structure, three-dimensional coordinates of protein structure; CHOP, dissection into structural domain-like fragments introduced here (Fig. 1); CLUP, simple clustering algorithm introduced here; ORF, open reading frame (for simplicity, we usually refer to ORFs from genome-sequencing projects as “proteins”); PDB, Protein Data Bank of experimentally determined 3D structures of proteins<sup>1</sup>; Pfam-A, expert curated database of protein families<sup>2</sup>; PrISM, automatic method assigning sequence-consecutive structural domains from PDB coordinates<sup>3</sup>; ProDom, automatic assignment of domain-like fragments from alignment information<sup>4–6</sup>; SCOP, structural classification of proteins (i.e., expert-based classification and domain dissection of protein structures)<sup>7</sup>; SWISS-PROT, a database of protein sequences.<sup>8</sup>

The Supplementary Materials Referred to in this article can be found at <http://www.interscience.wiley.com/jpages/0887-3585/suppmat/index.html>

Grant sponsor: Protein Structure Initiative of National Institutes of Health; Grant number: P50 GM62413.

\*Correspondence to: Burkhard Rost, CUBIC, Department of Biochemistry and Molecular Biophysics, Columbia University, 650 W. 168th St., BB217, New York, NY. E-mail: [rost@columbia.edu](mailto:rost@columbia.edu)

Received 5 June 2003; Accepted 5 December 2003

Published online 1 April 2004 in Wiley InterScience ([www.interscience.wiley.com](http://www.interscience.wiley.com)). DOI: 10.1002/prot.20095

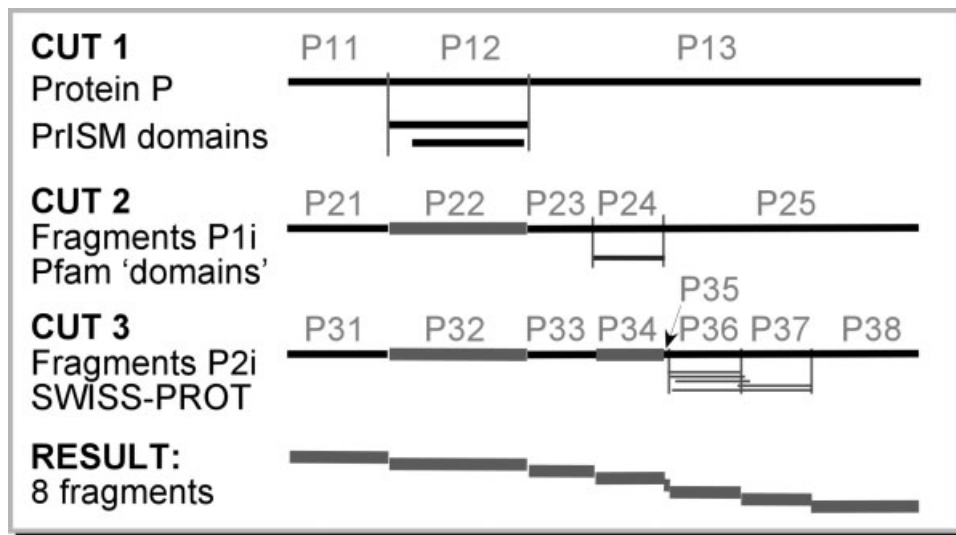


Fig. 1. Concept of CHOP. We dissect proteins in a hierarchical manner beginning from PrISM/PDB domains<sup>3,96,97</sup> [i.e., very reliable data about boundaries of structural domains (CUT1)]. If we find a similarity to many PrISM domains, we use the longest to chop (P12). For the example shown, this leaves us with two fragments (P11 and P13) to be processed. The next step (CUT2) is the identification of matches in Pfam-A<sup>2</sup>. In the example, fragments P21, P23, and P25 remain; all three are aligned to all SWISS-PROT proteins and are cut further if we find full-length proteins in SWISS-PROT the N- or C-termini of which align to part of the fragments (longest alignment that covers >80% of the SWISS-PROT protein). In the example, this dissects P25 into P36, P37 and into two undigested fragments P35 and P38.

the structure for long proteins when dissecting these into structural domains.<sup>30</sup>

### Progress in Identifying Structural Domains Without Structures

An increasing number of methods and databases address the problem of identifying structural domains from sequence.<sup>2,5,6,31–33</sup> The first automatic method, ProDom, identified likely domains through “boundaries” in multiple alignments.<sup>4–6</sup> The major problem of basing domain boundaries on alignments is that proteins are dissected into too small fragments<sup>6,31</sup>; in fact, entirely conserved segments as captured in the BLOCKS database,<sup>34,35</sup> usually span over short fragments of structural domains. This observation is by no means self-evident; rather, it points to the complexity of evolutionary constraints. DOMAINATION,<sup>36</sup> which delineates domains through analyzing iterative PSI-BLAST alignments, explicitly attempts to elongate domain-like regions. Other automatic methods apply concepts from protein structure prediction (SnapDRAGON),<sup>37</sup> statistics about domain size distributions,<sup>38</sup> a statistical approach toward combining various sources of information,<sup>39</sup> artificial neural networks,<sup>40,41</sup> or other ways of exploring alignment information.<sup>42,43</sup> Most recently, DomSSEA uses information from alignments of predicted secondary structure segments to identify structural domains.<sup>44</sup> Another unique idea is to first predict interresidue contacts by exploring correlated mutations and to then choose the domain so that the quotient intra/interdomain contact becomes minimal.<sup>45</sup> None of these more recent methods has yet been experimentally verified on a large scale.

### Clustering the Protein Universe Without Domains

Given the explosion of protein sequences, the necessity to cluster these data becomes increasingly urgent.<sup>31,46–51</sup> One of the most practical reasons is to speed up databases comparisons; another is to reduce the bias and, hence, sharpen such searches.<sup>46,52–56</sup> Maps of the universe of proteins, such as ProtoNet,<sup>57</sup> ProtoMap,<sup>49,58</sup> or BioSphere,<sup>47</sup> tend to group proteins with similar function.<sup>59</sup> One problem is that such clustering methods typically begin with full-length proteins. Two groups attempted to first dissect proteins into domain-like fragments and to then cluster these fragments through GeneRage<sup>42</sup> and PICASSO<sup>51</sup>; the GeneRage algorithm fails to handle long, complex eukaryotic proteins,<sup>42,60</sup> and PICASSO is not available for public testing.

Here we introduce a hierarchical approach for chopping proteins into fragments that resemble structural domains (CHOP, Fig. 1). The hierarchy imposed begins from the most reliable information (proteins of known structure), continues to families that are well characterized by experts (Pfam-A), and finally explores the reliable information about N- and C-terminal ends of full-length proteins that have been characterized experimentally (SWISS-PROT). The objective is not to obtain all domain boundaries, but rather to identify only those boundaries for which we are confident; 70% of the proteins from 62 entirely sequenced organisms can be dissected by CHOP; the length distribution of CHOP fragments resembles that of known structural domains. Next, we clustered all these fragments and unchopped full-length proteins (CLUP). The CHOP fragments and the CLUP clusters have been successfully applied to the target selection process in

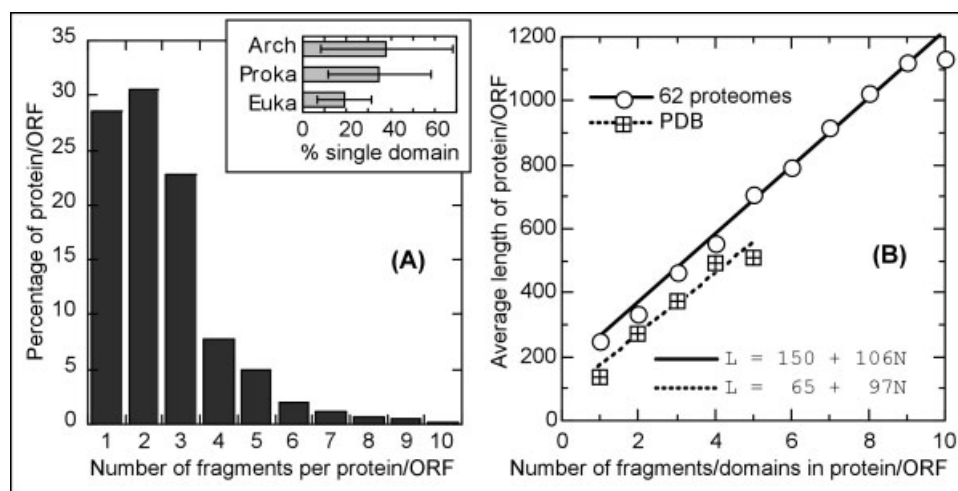


Fig. 2. Fraction of proteins chopped and number of fragments per protein. **A:** Distribution of number of CHOP fragments per protein/ORF; for example, although CHOP is incomplete, we found that 70% of all chopped proteins appeared to have more than one structural domain-like fragment. The inset shows the percentages of single-domain proteins for eukaryotes, prokaryotes, and archae; the bars mark the minimal and maximal numbers within each kingdom (more details in Fig. S3, Supplement). **B:** Relation between number of fragments and protein length: on average, the number of CHOP fragments appeared to increase linearly with protein length. The line constituted a linear regression fit of  $L = 150 + 106N$  ( $R > 0.99$ , with  $L$  being the length of the protein and  $N$  the number of fragments). For comparison, the average length of PDB proteins also increases with the number of domains they have, with linear fit of  $L = 65 + 97N$  ( $R = 0.979$ ).

North East Structural Genomics Consortium (NESG, <http://www.nesg.org/>) and are publicly available through an SRS<sup>61</sup> interface at <http://cubic.bioc.columbia.edu/srs/><sup>60</sup>; data in flat files format will also be available on request).

## RESULTS

### Most Proteins Had More Than One Fragment

We applied hierarchically the three steps of CHOP (Fig. 1) to all proteins/ORFs from 62 entirely sequenced organisms (Table in Supplement): 164,433 (69%) of the 238,492 proteins were dissected by CHOP. More than 70% of these chopped proteins had more than one fragment [Fig. 2(A)]. As expected, eukaryotes have fewer single domain proteins than do prokaryotes and archae [Fig. 2(A), inset; detailed graph in Fig. S3, Supplement]. In contrast to these data, most of the 30,309 PDB chains available in fall 2003 constituted single-domain proteins (18,292 = 60%). However, structural biologists have strong incentives to choose domain-like proteins and to determine the structure for fragments of longer proteins. In fact, for one fourth (4,854) of the seemingly single-domain proteins, we found the corresponding proteins in SWISS-PROT to be at least 50 residues longer than the PDB version. Assuming that these do then also constitute multidomain proteins, <45% (13,438) of the PDB proteins have single domains. One extreme case of a mega multidomain—*Drosophila melanogaster* protein bt (flybase identifier: FBan0001479) with 7,107 residues—was chopped into 79 fragments. This is not surprising, because bt is a member of the titin family, well known for its titanic usage of immunoglobulin-like and fibronectin type 3 (Fn3) domains.<sup>62</sup> The longest human protein in our data set was the ovarian cancer-related tumor marker CA125 (TrEMBL id q8wxi7<sup>8</sup>) with 11,721 residues; it was cut into seven fragments in its middle,

whereas the C-terminus remained uncut for almost 9,000 residues; >2,000 of the residues were in low-complexity regions, distributed more or less equally over the entire protein.

### Most Domains Span on Average Over ~100 Residues

When averaging over the lengths of all proteins with  $N$  CHOP fragments, we observed that the number of CHOP fragments was directly proportional to the protein length [Fig. 2(B)]. On average, most CHOP fragments for multifragment proteins stretched over ~106 residues [slope of linear fit in Fig. 2(B)]. When we compared structurally known PrISM domains on the same plot, we noted a very similar slope [Fig. 2(B)]. Although the fit for PDB proteins was supported by less data, in particular for proteins with many domains, the two fits were strikingly parallel, suggesting that our result was not caused by the lack of precision and/or completeness of the CHOP procedure. The linear fits also unraveled another surprising observation: single CHOP fragments from the 62 proteomes extended over ~256 residues, and those from PrISM domains extended over 163 residues. Both numbers were significantly larger than the averages for subsequent domains in multidomain proteins. Thus,  $N-1$  domains in proteins with  $N$  domains extend over 97–106 residues, whereas one extends over 163–256 residues. (Note that the average length of the first fragment exceeded 300 residues for eukaryotic proteins.<sup>63</sup>) To establish that this unexpected finding was not caused by the particular way of presenting the data (average length vs. number of domains), we pooled all domain-like fragments and randomly “assembled proteins” according to the observed distributions for the number of fragments per protein (data not shown).

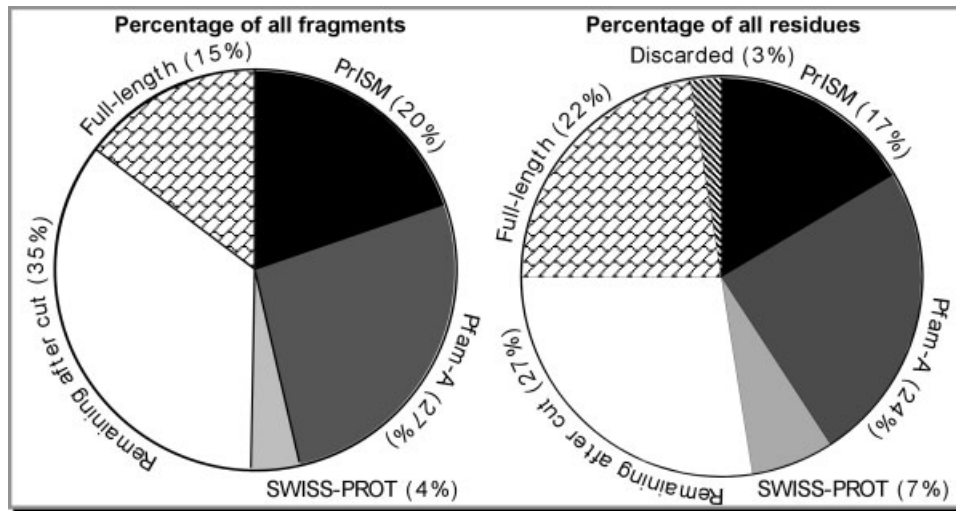


Fig. 3. Sources of chopping. **Left:** One fifth of all fragments were obtained from homology to PrISM/PDB domains, and half did not find any similar region to cut; 15% of the fragments were full-length proteins that were not cut at all. **Right:** Residue-wise, the PrISM fragments covered 17% of all the 87,732,248 residues in the 62 proteomes; Pfam-A fragments 24%, and 3% of the residues were discarded during the chopping because they were from fragments that were smaller than 30 residues. In other words, ~47% of all CHOP fragments appeared supported by very conservative data, and they covered 41% of all residues.

As expected, this control experiment yielded a line passing through 0. Thus, our finding is not explained by the particular presentation of the data. Thus, the detailed fit is likely to constitute a more precise estimate of the average length of a structural domain than the one that is obtained from compiling a simple average over all domains currently annotated in PDB.

#### 47% of the Fragments Were Similar to PrISM and Pfam-A

The number of domains for proteins of known structure fully agrees for only 40% of all proteins contained in PDB/PrISM and Pfam-A; for ~10% of the common proteins, PrISM/PDB and Pfam-A disagree by more than two domains (Fig. S2, Supplement). Because our goal is to generate structural domain-like fragments, we favored the structure-derived (PrISM/PDB) over the sequence-derived (Pfam-A) domain assignment. PrISM domains were the origin of 20% of all CHOP fragments [Fig. 3(A)]. Only about 15% of all 499,465 CHOP fragments constituted full-length proteins (i.e., were never touched by our hierarchical procedure) [Fig. 3(A)]. These untouched, full-length proteins covered ~22% of all residues [Fig. 3(B)]. Conversely, ~20% of the fragments originated from known structures (PrISM domains). Overall, ~47% of all CHOP fragments appeared supported by very conservative data; these fragments covered 41% of all residues.

#### Lengths of CHOP Fragments Resembled Structural Domains

Are the remaining fragments and the uncut full-length proteins similar to structural domains, or did we simply not find the necessary data to dissect these? Although we could not answer this question conclusively, the length distribution of the remaining fragments suggested that at

least, on average, these differed significantly from the whole set of full-length proteins (Fig. 4). In fact, the major problem in terms of length distribution appeared the overrepresentation of fragments shorter than 50 residues in the set of fragments that remained after chopping. In contrast, the length distribution of those full-length proteins that had not been chopped at all appeared more similar to Pfam-A regions than to the entire set of all full-length proteins. In other words, these untouched proteins constituted a subset of short proteins, many of which might in fact have single domains. The length distribution of all CHOP fragments (including the short remaining fragments and the untouched proteins) were most similar to the Pfam-A distribution in which short (50% shorter than 106 residues) and long fragments (>350) are overrepresented in comparison to “real” structural domain as taken from PrISM. The obvious outliers were the fragments obtained through homology to full-length SWISS-PROT proteins: these tended to be much longer than fragments chopped according to PrISM (Fig. 4). However, because they accounted for only 4% of all CHOP fragments [Fig. 3(A)], they did not affect the length distribution of all CHOP fragments significantly. The length distribution for all CHOP fragments was similar for all three kingdoms (eukaryotes, prokaryotes, and archae; data not shown). However, we observed significant differences for those fragments that originated from SWISS-PROT termini: although almost 30% of the eukaryote SWISS-PROT fragments were longer than 500 residues, <5% of the archae and <10% of the prokaryote SWISS-PROT fragments were as long. Prokaryotic fragments from Pfam-A were slightly longer than eukaryotic ones, and long eukaryotic fragments were overrepresented in both the sets of untouched full-length proteins and that of remaining fragments (data not shown).

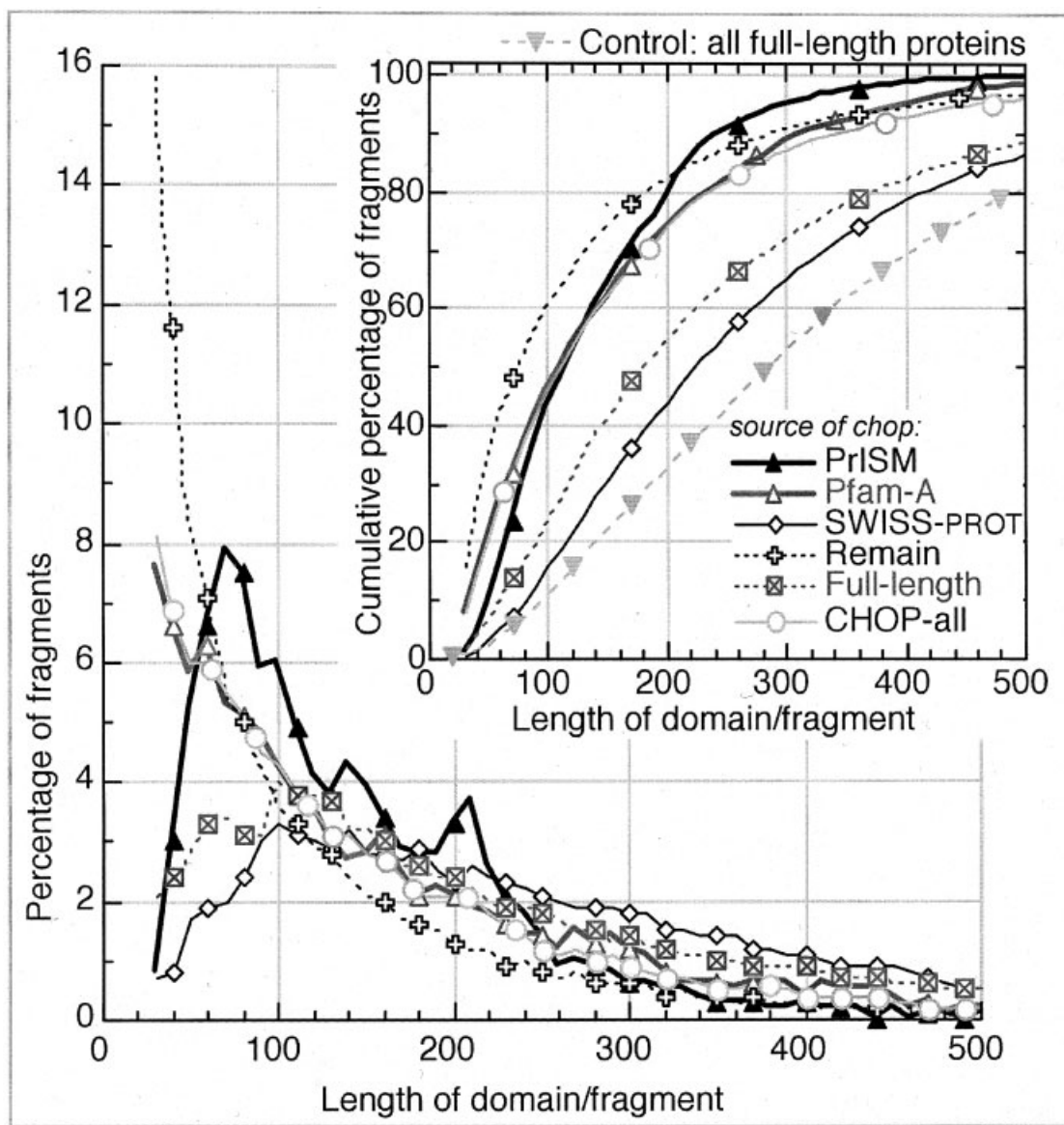


Fig. 4. Length distribution of CHOP fragments. Note that except for the controls shown in the cumulative percentages in the top-right inlet (full-length proteins), all curves described the CHOP fragments (e.g., the thick black line with filled triangles showed the distributions for fragments that were chopped through similarity to PrISM domains). "Remain" marks those fragments that remained N- and/or C-terminal from a region cut out according to similarity to either PrISM domains, Pfam-A regions, or SWISS-PROT termini; "Full-length" mark proteins that were not touched at all by the CHOP algorithm. Both the cumulative and noncumulative curves for all CHOP fragments (gray with open circles) were almost indistinguishable from those for Pfam-A fragments. Fragments cut according to SWISS-PROT termini (4%; Fig. 3) were closer to the distribution of all full-length proteins (control, light gray with downward pointing triangles) than the subset of proteins that remained untouched.

#### Partial Agreement With SUPERFAMILY Assignments for Yeast

One difficulty in evaluating the CHOP procedure was that we used all available information. The method of cross-validation was not easily applicable because the different sources of domain dissection overlapped only partially, and even where they did overlap, we could at best verify that our procedure was self-consistent (see above). Therefore, the overall length distribution provided an independent perspective on the global performance of

CHOP. Lacking unused standards of truth, we compared CHOP with another method that has implicitly a partially similar goal, namely, SUPERFAMILY.<sup>64</sup> SUPERFAMILY is a library of hidden Markov models (HMMs) for SCOP structural superfamilies. SUPERFAMILY contains data for >100 proteomes; because of CPU restrictions, we had to limit the comparison to one particular proteome, namely, yeast, the smallest entirely sequenced eukaryote with 6349 proteins. SUPERFAMILY assigned 4794 domains in a subset of 3359 proteins. In contrast, CHOP found 6000

**TABLE I. Failure of Single-Linkage Clustering for Full-Length Proteins<sup>†</sup>**

Lg(E value)	Singleton	Non-singleton clusters	Largest cluster
0	9546	1879	86601
-1	11489	3462	74144
-2	12485	4318	66207
-3	13217	4746	62192
-6	15385	5974	50668
-10	18094	7153	38936

<sup>†</sup>Data set: all 102,932 proteins from yeast, fly, worm, weed, and human.

domain-like fragments from PrISM and Pfam-A in a subset of 3915 proteins. About 72% of the SUPERFAMILY domains are continuous in sequence (3435); thus, their domain boundaries were directly comparable to CHOP fragments: 40% (1380) of these SUPERFAMILY domains agreed with CHOP fragments (defined as 80% overlap between both assignments), 30% were significantly longer than CHOP fragments, and 28% shorter. SUPERFAMILY associations imply predictions for structure, because SUPERFAMILY models are based on proteins of known structure. The same is true, on average, for only 20% of the CHOP fragments [Fig. 3(A)]. Therefore, it may appear surprising that both methods cover such a similar number of proteins in yeast (3359 by SUPERFAMILY vs. 3915 by CHOP). The reason is that we apply much more stringent thresholds in sequence similarity when we fragment a protein. When we use CHOP fragments to pinpoint putative targets for structural genomics,<sup>63</sup> we add another step that is conceptually similar to SUPERFAMILY and thereby increase the coverage to levels more similar to those achieved by SUPERFAMILY.

### Clustering Sequence Space Must Begin From Fragments

Multidomain proteins constitute the most troublesome challenge to clustering sequence space. When we ignored this challenge and single-linkage clustered full-length proteins from five entirely sequenced eukaryotes (yeast, fly, worm, weed, and human), we observed that no matter how we chose the thresholds for the clustering, we ended up with one big “cluster” that pulled in almost half of all proteins<sup>65</sup> (Table I). Thus, we cannot cluster entire proteomes in a way that ascertains all proteins in one cluster to have a structurally similar domain-like region without dissecting proteins into fragments. Could we succeed with single-linkage clustering if we knew all structural domains? We addressed this question by clustering all PrISM domains of known structure (47,582): We found 835 clusters with one member (singleton) and 3039 clusters with more than one. Not surprisingly, the largest cluster with 2156 members was mostly immunoglobulin related, and the second largest had 1224 serine proteases/hydrolases. Although the results from clustering structural domains looked much more reasonable than those from clustering full-length eukaryotic proteins, the largest cluster still contained many unrelated domain pairs (e.g., two

immunoglobulin domains unrelated in sequence). To avoid this problem, we developed a clustering scheme (CLUP) that started with proteins that have few homologues and avoided merging families based on transitive homology. For the same set of PrISM/PDB domains, CLUP yielded 835 singletons, 4061 multimember groups, the largest of which had 558 structural domains from serine protease family. Visual inspection of the largest clusters did not reveal any grouping of unrelated proteins. Furthermore, no cluster was degenerate in the sense that each domain belonged to only one cluster.

### Over 63,000 Multimember Clusters From the CHOP Fragments

We clustered all 499,465 CHOP fragments from 62 entirely sequenced genomes, including those uncut full-length proteins that were treated as single fragments (CLUP, Methods). CLUP grouped these fragments into 118,108 single- and 63,300 multimember clusters. About 43% of the multimember clusters had more than three members, and ~13% had >10 members [Fig. 5(A)]<sup>1</sup>. The largest cluster contained 3343 members; the seed for this cluster was an ATP-binding cassette from the ABC transporters. Our premise was that structural domains constitute something like the atom or basic unit of sequence space. If true, two such units should not occur in different clusters. We could explicitly build such a constraint into our clustering scheme. However, the benefit of the simplicity of CLUP was that no such constraint was applied. Hence, for our clustering scheme, the percentage of ambivalently clustered fragments constituted an indirect measure for the reliability of the CHOP fragmentation. Most of the CHOP fragments (76%) were single-cluster fragments, whereas only 3% of the fragments are associated with more than three clusters [Fig. 5(B)].

## DISCUSSION AND CONCLUSIONS

### CHOP Is the First Step Toward Complete Domain Dissection

On the one hand, CHOP failed to identify all domain boundaries: 31% of the proteins remained unchopped; these accounted for 15% of all final CHOP fragments [Fig. 3(A)]. On the other hand, even the remaining CHOP fragments and those proteins that were not chopped, on average, were more similar to structural domains than to full-length proteins (Fig. 4). The obvious difference between the length distribution of CHOP fragments and PrISM domains or Pfam-A regions was an abundance of short fragments. Because this fragmentation was largely due to the fragments remaining after chopping (Fig. 4), it is not clear whether these fragments indicated the limitation of the CHOP procedure or simply constituted a large pool of domain-linking fragments. For example, about one fifth of the remaining fragments were probably signal

<sup>1</sup>Incidentally, the distribution of the number of fragments could be forced to fit a popular power-law Percentage (N)  $\sim N^{-2.3}$  as the one initially proposed by the Harvard linguist George Kingsley Zipf<sup>66</sup> to-amongst many other purposes-describe the frequency of English words and the population density of cities.

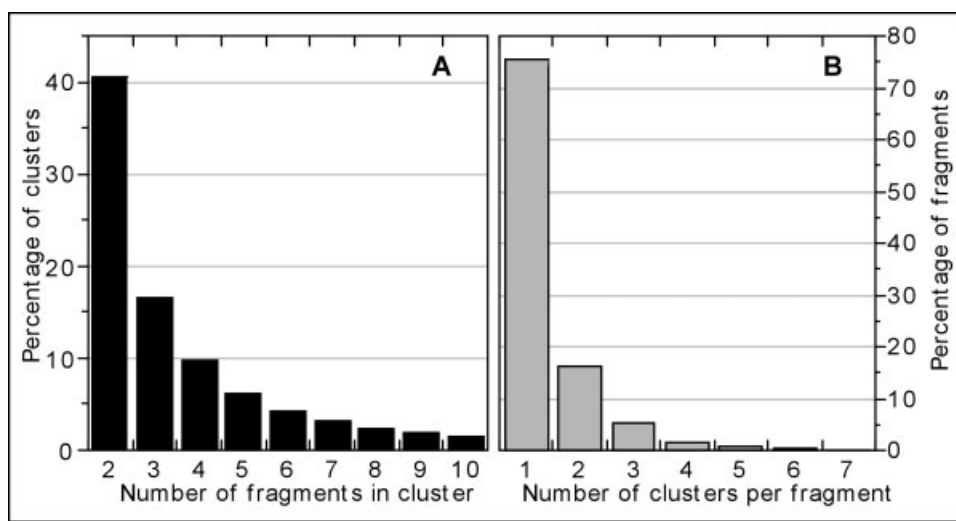


Fig. 5. Cluster sizes and degeneracy. **A:** For the subset of multimember clusters, 41% have two, about two thirds more than four, and ~13% >10 members. **B:** Most CHOP fragments (76%) were associated with a single cluster (not degenerate), whereas only 3% of the fragments were associated with more than three clusters (highly degenerate).

peptides, and many originated from membrane spanning helices.<sup>9,65,67</sup> We did find some examples in these remaining fragments that appeared to be valid functional regions rather than just leftover fragments. For example, the yeast Tup1 protein was dissected by CHOP into a C-terminal domain (residues 315–713) homologous to the PDB entry 1ERJ<sup>68</sup> and into an N-terminal fragment (1–315); it has been reported that this N-terminal portion of Tup1 protein is responsible for the oligomerization of Tup1 and for mediating the repression of transcription.<sup>69</sup> Overall, CHOP fragments appeared not as atomized as those obtained through prediction from, for example, ProDom<sup>6</sup> or through expert annotation from SMART<sup>31,70</sup> (Fig. S4, Supplement). Obviously, the length distribution of CHOP fragments constituted only a very crude means of evaluating the reliability of CHOP. However, given the nature of our procedure (we used all reliable information to chop), we had no data left that enabled a more comprehensive evaluation. Cross-validating our procedure by chopping according to Pfam-A and SWISS-PROT only for proteins of known structure confirmed that CHOP was somehow consistent. However, even this cross-validation was somehow circular. The CHOP procedure never dissected proteins without strong evidence; this reduced the data set considerably without introducing too many mistakes. Currently, we work on the next step, namely, the development of a de novo prediction of domain boundaries that may help to distinguish “single domain” from “uncut because of lacking homology” for the remaining fragments and for the untouched full-length proteins.

#### Most Proteins Had More Than One Structural Domain-Like Fragment

On average, we found a CHOP fragment for every 106 residues [Fig. 2(B)]. Thus, most of the proteins from the 62 entirely sequenced organisms contained more than one domain-like fragment; in fact, >70% of the proteins that

could be chopped had multiple domain-like fragments [Fig. 2(A)]. This number is significantly higher than the multidomain proteins previously described in *E. coli* (6% with 2–4 domains),<sup>71</sup> and it also exceeds the percentage of multidomain proteins analyzed in a detailed comparison of structural known enzymatic domain “mosaics” (32%).<sup>21,72</sup> Supposedly, the latter differs so much from what we observed, because it is biased by the bias of known structures in PDB: most PDB proteins appear to be single-domain proteins.<sup>7</sup> However, we observed that the bias toward single-domain proteins in PDB partially originated from the fact that proteins are often cut to determine structure: <44% of our PrISM domain set appeared to correspond to single-domain proteins. Supposedly, our estimate of that ratio of single/multidomain proteins provides a lower limit to the real number, because the CHOP algorithm misses domains that have not been characterized previously.

#### Do Domains From Single and Multidomain Proteins Differ?

Given the CHOP algorithm, fragments left and right of a structural domain will often be too short (e.g., a signal peptide or a few helices before the globular domain in G-coupled receptors), we expected that the relation between length and number of fragments differed between single- and multifragment proteins. More specifically, we expected that this leftover effect would be more extreme for proteins with two than for those with five fragments (less leftover). Although the data confirmed the expected tendency, we observed that the number of CHOP fragments was directly proportional to the protein length [Fig. 2(B)]. In fact, on average, CHOP fragments for multifragment proteins stretched over ~106 residues [slope of linear fit in Fig. 2(B)], whereas single-fragment proteins were ~256 residues long (value of fit for  $N = 1$ ). We might argue that this difference between the average lengths of single- and multi-fragment proteins originated from the

leftover problem. However, this interpretation was not consistent with the data, because the leftover effect between proteins with two and five fragments appeared negligible. Did this suggest that there really is a genuine difference between the domains used in single- and those used in multiple-domain proteins? Is it maybe less expansive to shuffle short domains than long ones? Do proteins need a minimal length for entropic reasons to fold? Are shorter domains relics from ancient, longer domains? Or do catalytic functions require a minimal length and do many of the shorter domains in multidomain proteins increase the complexity of regulation?<sup>73</sup> Our data could not shed light onto these speculations.

### CHOP Procedure Rather Robust With Respect to Local Parameter Changes

The two major parameters for the CHOP procedure are the minimal coverage of a known domain (i.e., the minimal fraction of a PrISM or Pfam-A region that we require to be similar in sequence to chop) and the minimal level of sequence similarity. By default, we required 80% coverage of and BLAST E-values  $<10^{-3}$ . However, the overall number of fragments did not alter significantly when modifying these parameters. For example, the number of CHOP fragments varied only 0.3% when the coverage threshold was changed from 90% to 70%, and 0.6% when changing the BLAST E-value from  $10^{-3}$  to  $10^{-1}$  (Fig. S1, Supplement). We also checked the consistency of the sources for chopping by chopping according to only one source and verifying the fragments with another. The domain boundaries from similarity to SWISS-PROT proteins were rarely in conflict with those from similarity to PrISM domains (0.2%) and Pfam domains (1%). For proteins that could be chopped according to both PrISM and Pfam-A, the number of domain-like fragments resulting from applying the two methods independently was largely consistent: 40% of the proteins showed no difference, and 34% differed by one domain (Fig. S2, Supplement). When the two methods differed substantially, often Pfam-A detected multiple copies of a repeat, whereas PrISM treated them altogether as one structural domain. Examples for this are typically very short fragments such as helix-turn-helix motifs, or the Pfam-A dissection of beta-helices into single helices annotated independently. We removed some of such cases by introducing another parameter that removed all Pfam-A entries shorter than 30 residues. Thus, overall CHOP was rather consistent and fairly robust to local parameter changes.

### CLUP Succeeded Somehow in Clustering Protein Space

Previously, we showed that single-linkage clustering failed to generate reasonable clusters from full-length eukaryotic proteins.<sup>60</sup> In this study, we demonstrated the failure of such a simple clustering scheme more systematically (Table I). Even when clustering PDB domains with single linkage, we found many unrelated protein pairs in the resulting clusters. We presented a novel clustering scheme that did not merge clusters based on transitive homology. Overall, CLUP yielded clusters that were reason-

able by the following criteria: 1) the largest cluster did not appear to join many completely unrelated proteins; 2) only ~24% of the clusters had cross-relations (i.e., two different clusters shared one or more fragments); and 3) the overall length distribution of the clusters appeared reasonable (Fig. S4, Supplement). As much as CHOP is a first step toward domain dissection, CLUP is a first step toward clustering these domain-like fragments. Although systems such as ProtoNet,<sup>57</sup> ProtoMap,<sup>49,58</sup> or BioSphere<sup>47</sup> explicitly map out the protein universe, CLUP has no notion of distance between two clusters other than that we cannot merge the two. Instead, CLUP only groups all fragments that are likely to have some common structural core. We imagine that systems comprehensively mapping sequence space could benefit directly from using our CHOP fragments as input to their clustering. On the one hand, the success of CLUP relies on the fact that CHOP succeeded in dissecting a reasonable fraction of the proteomes into domain-like fragments. Much more advanced and intelligent clustering schemes have been proposed.<sup>42,47,49–51,53,57,58,74–76</sup> Some of these seemingly even cope with the complexity of eukaryotic proteomes. On the other hand, the advantage of CLUP over more complex systems may be that the clustering is based exclusively on pair relations: for example, although ProtoNet and BioSphere enable the discovery of nontrivial connections, both advanced systems cannot always “name” the relation between any pair of proteins in two connected clusters. In contrast, all pairs in CLUP clusters most likely share a common structural foldlike region. This feature was crucial to applying CLUP to the task of selecting targets for structural genomics.<sup>63</sup> Our structural domain-like fragments may also constitute good starting points for reducing the noise in two-hybrid, TAP, and mass-spectrometry experiments that probe protein–protein interactions between fragments rather than between full-length proteins.

### How Many Clusters of Close Homologues Are There?

A back-of-the-envelope calculation made Cyrus Chothia challenge that there are only 1000 different folds in nature.<sup>77</sup> This short note set off an avalanche of alternative estimates, most of which corrected the number upward.<sup>9,78–84</sup> Structural genomics initiatives thrive at experimentally determining most existing folds.<sup>17,83,85–92</sup> How many structures will these initiatives have to determine to cover fold space? Even if Chothia's estimate were right within an order of magnitude (i.e., if there were  $<10,000$  folds in nature), it has been estimated that because of technical reasons, structural genomics would have to determine  $>2–10$  times more structures to get a representative for each fold.<sup>65,83,91,93</sup> However, all these estimates were based on a number of assumptions generalizing from statistics about proteins of known structure; none of these estimates actually clustered sequences and explicitly selected targets for structural genomics. The methods that we presented here can be used for target selection. The only step that remains to actually select targets for structural genomics is to exclude all clusters for which we either already have structural information or

which may not constitute high-priority targets for structural genomics.<sup>63</sup> About 30,000 of the multimember clusters have no obvious sequence similarity to known structures, hence, could constitute a minimal set of targets for structural genomics (data not shown). However, CLUP also created 118K singletons; some of these might find a connection to the multimember clusters when we know more protein sequences. Yet, many may eventually support another daring challenge put forward by Coulson and Moulton,<sup>79</sup> namely, that many folds have been realized only once in evolution. In any case, our clusters of structural domain-like fragments are likely to constitute a good starting point for both structural and functional genomics.

## MATERIALS AND METHODS

### Obtaining Structural Domains From PrISM

Structural domains for PDB<sup>1</sup> proteins were extracted by PrISM.<sup>3</sup> We chose PrISM rather than other databases such as SCOP<sup>7</sup> and CATH<sup>25</sup> for the following reasons: 1) All domains defined by PrISM are continuous in sequence. This was important to us to maximize the probability that a domain identified by sequence similarity is correct. (2) Because we have the program locally available, we can obtain the domain classification in real time (i.e., while structures are being added to PDB).

### Chopping Full-Length Proteins Into Domain-Like Fragments (CHOP)

CHOP implements three hierarchical steps that were applied by decreasing confidence in the accuracy of the information (Fig. 1). In particular, we discarded all fragments from step S that overlapped with fragments identified in the previous step S-1 (more reliable identification of domain boundaries). At any step, we discarded fragments with <30 residues. For the set of all proteins in 62 organisms {P62}, we applied the following three steps.

#### 1. High-reliability: PrISM domains

We applied a pairwise BLAST<sup>94</sup> search with each protein against all PrISM domains (49,300 domains); we marked all hits at expectation values  $< 10^{-2}$  that aligned at least 80% of the PrISM domain. If this criterion applied to more than one PrISM domains, we chose the longest. Next, we cut these domain fragments from the list of all proteins and repeated the search with the remaining fragments until we no longer found any similarity to PrISM. Note that the new BLAST is no longer strictly local. Therefore, the repeated search with fragments rather than full-length proteins may uncover similarities that were previously overlooked.

#### 2. Acceptable confidence: Pfam families

When comparing the length distribution of public curated databases of protein families, we found Pfam-A to come closest to the notion of structural domains.<sup>31</sup> Therefore, we assumed that Pfam-A constitutes the next best resource to increase coverage in our domain dissection. For each of the remaining fragments, we searched for similarities in Pfam-A with HMMER<sup>95</sup> (global mode, Pfam release 7.0 with 5049 families, threshold  $10^{-2}$ ). If a similarity to

any Pfam family was detected for the fragment, we again dissected it in the same way as before for the PrISM domains. The procedure was repeated until no homology of Pfam was detected in any of the remaining fragments.

### 3. Evolutionary relation unraveled by SWISS-PROT termini

If the N-terminal N1 residues of protein A are similar to  $\geq 80\%$  of an experimentally characterized, full-length protein B, whereas the N2 C-terminal residues of A find no similarity in known databases, we have a good reason to suspect that A has at least two domains. SWISS-PROT<sup>8</sup> contains full-length proteins that have typically been studied independently by experimentalists; in particular, SWISS-PROT is not prone to "pollution" by short EST fragments. SWISS-PROT entries marked as "fragment" or shorter than 30 residues were excluded from our search. We identified all homologues of full-length proteins contained SWISS-PROT<sup>8</sup> through pairwise BLAST searches with all the remaining fragments (BLAST E-value  $< 10^{-2}$ , covering at least 80% of the SWISS-PROT protein). As before, we iterated until no similarity remained.

To retain the integrity of domains, we imposed the restriction that cutting points were only introduced when the homology covered most ( $> 80\%$ ) of the structural domain (PrISM), domain-like region (Pfam), or full-length protein. The final set of fragments was the combination of all fragments identified in the three steps and all remaining fragments that were longer than 30 residues. The dissection was not sensitive to our choice of parameters (BLAST E-value of  $10^{-2}$  and coverage of 80%).

### Clustering CHOP Protein Fragments (CLUP)

Our goal for the final clustering is that all members of one cluster share a region of common structure that basically spans an entire structural domain. We approached this goal by clustering all the fragments (and uncut full-length) proteins obtained from CHOP. Toward this end, we simply ran an all-against-all PSI-BLAST<sup>94</sup> search with a threshold E-value  $< 10^{-3}$ . The final clustering step involved the following iteration.

#### 1. Initialize

Put all CHOP fragments onto a stack C, sort (a) according to number of homologues found by PSI-BLAST (ascending), and (b) by ascending sequence length. In other words, we began with the smallest group and the shortest sequence.

#### 2. Iterate

Select the first fragment from the stack (small group, short sequence), consider it as seed, and create a cluster that contains this seed and all related fragments. Finally, remove the seed and all the related fragments from the stack. Note that although seeds and related fragments that are removed from the stack at this point may still become members of other clusters, they will not seed any other cluster.

#### 3. Repeat until completion:

We repeat step 2 until no fragment is left in our stack.

#### 4. Merge clusters

Two clusters are merged if homologous regions of any common member with regards to the seeds significantly overlap (80%).

#### ACKNOWLEDGMENTS

We thank our experimental colleagues at the Northeast Structural Genomics Consortium (NESG) for their advice and strong support of our project. In particular, we thank Guy Montelione (Rutgers) for his invaluable optimism in leading the NESG team. We thank the other experimental teams around Tom Acton (Rutgers), Cheryl Arrowsmith and Aled Edwards (Toronto), Wayne Hendrickson, John Hunt, and Liang Tong (Columbia), Mike Kennedy (Pacific Northwest Natl Laboratory, Richland), and George DeTitta (Buffalo). Thanks to our colleagues from target selection for crucially helpful discussions: Barry Honig and Sharon Goldsmith (Columbia) and Diana Murray (Cornell). To Mark Gerstein and his group (Yale) for pushing us to develop PEP. Particular thanks to Phil Carter (Columbia, New York and Imperial College, London) for building the databases PEP, CHOP, and CLUP, and to An-Suei Yang (Columbia) for providing and helping with PrISM. Thanks also to Michal Linial (Jerusalem) for insightful discussions that kept the fun part of science alive. Last not least, thanks to all those who deposit their experimental data in public databases, in particular in the context of structural genomics, and to the teams around PDB (Helen Berman, Rutgers, and Phil Bourne, UCSD), Pfam (Alex Bateman, Sanger, and Erik Sonnhammer, Stockholm), and SWISS-PROT (Amos Bairoch, SIB, Geneva) who maintain these databases that were central to this work.

#### REFERENCES

- Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. *Nucleic Acids Res* 2000;28:235–242.
- Bateman A, Birney E, Cerruti L, Durbin R, Eddy SR, Griffiths-Jones S, Howe KL, Marshall M, Sonnhammer EL. The Pfam protein families database. *Nucleic Acids Res* 2002;30:276–280.
- Yang AS, Honig B. An integrated approach to the analysis and modeling of protein sequences and structures. III. A comparative study of sequence conservation in protein structural families using multiple structural alignments. *J Mol Biol* 2000;301:691–711.
- Sonnhammer EL, Eddy SR, Durbin R. Pfam: a comprehensive database of protein domain families based on seed alignments. *Proteins* 1997;28:405–420.
- Corpet F, Servant F, Gouzy J, Kahn D. ProDom and ProDom-CG: tools for protein domain analysis and whole genome comparisons. *Nucleic Acids Res* 2000;28:267–269.
- Servant F, Bru C, Carrere S, Courcelle E, Gouzy J, Peyruc D, Kahn D. ProDom: automated clustering of homologous domains. *Brief Bioinform* 2002;3:246–251.
- Lo Conte L, Brenner SE, Hubbard TJ, Chothia C, Murzin AG. SCOP database in 2002: refinements accommodate structural genomics. *Nucleic Acids Res* 2002;30:264–267.
- Bairoch A, Apweiler R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res* 2000;28:45–48.
- Liu J, Rost B. Comparing function and structure between entire proteomes. *Protein Sci* 2001;10:1970–1979.
- Rost B. Did evolution leap to create the protein universe? *Curr Opin Struct Biol* 2002;12:409–416.
- Rose GD. Hierarchic organization of domains in globular proteins. *J Mol Biol* 1979;134:447–470.
- Jaennicke R. Folding and association of proteins. *Prog Biophys Mol Biol* 1987;49:117–237.
- Hao MH, Scheraga HA. Molecular mechanisms for cooperative folding of proteins. *J Mol Biol* 1998;277:973–983.
- Sham YY, Ma B, Tsai CJ, Nussinov R. Thermal unfolding molecular dynamics simulation of Escherichia coli dihydrofolate reductase: thermal stability of protein domains and unfolding pathway. *Proteins* 2002;46:308–320.
- Thornton JM, Orengo CA, Todd AE, Pearl FM. Protein folds, functions and evolution. *J Mol Biol* 1999;293:333–342.
- Ponting CP, Russell RR. The natural history of protein domains. *Annu Rev Biophys Biomol Struct* 2002;31:45–71.
- Hurley JH, Anderson DE, Beach B, Canagarajah B, Ho YS, Jones E, Miller G, Misra S, Pearson M, Saidi L, Suer S, Trievel R, Tsujishita Y. Structural genomics and signaling domains. *Trends Biochem Sci* 2002;27:48–53.
- Holm L, Sander C. Parser for protein folding units. *Proteins* 1994;19:256–268.
- Baron M, Norman DG, Campbell ID. Protein modules. *Trends Biochem Sci* 1991;16:13–17.
- Bennett MJ, Schlunegger MP, Eisenberg D. 3D domain swapping: a mechanism for oligomer assembly. *Protein Sci* 1996;5:2455–2468.
- Teichmann SA, Rison SC, Thornton JM, Riley M, Gough J, Chothia C. The evolution and structural anatomy of the small molecule metabolic pathways in Escherichia coli. *J Mol Biol* 2001;311:693–708.
- Apic G, Gough J, Teichmann SA. Domain combinations in archaeal, eubacterial and eukaryotic proteomes. *J Mol Biol* 2001;310:311–325.
- Lupas AN, Ponting CP, Russell RB. On the evolution of protein folds: are similar motifs in different protein folds the result of convergence, insertion, or relics of an ancient peptide world? *J Struct Biol* 2001;134:191–203.
- Zuckerklund E, Pauling L. Evolutionary divergence and convergence in proteins. In: Bryson V, Vogel HJ, editors. *Evolving genes and proteins*. New York and London: Academic Press; 1965. p 97–166.
- Orengo CA, Bray JE, Buchan DW, Harrison A, Lee D, Pearl FM, Sillitoe I, Todd AE, Thornton JM. The CATH protein family database: a resource for structural and functional annotation of genomes. *Proteomics* 2002;2:11–21.
- Dietmann S, Park J, Notredame C, Heger A, Lappe M, Holm L. A fully automatic evolutionary classification of protein folds: DALI domain dictionary version 3. *Nucleic Acids Res* 2001;29:55–57.
- Dengler U, Siddiqui AS, Barton GJ. Protein structural domains: analysis of the 3Dee domains database. *Proteins* 2001;42:332–344.
- Marchler-Bauer A, Panchenko AR, Ariel N, Bryant SH. Comparison of sequence and structure alignments for protein domains. *Proteins* 2002;48:439–446.
- Holm L, Sander C. Dictionary of recurrent domains in protein structures. *Proteins* 1998;33:88–96.
- Montelione GT, Zheng D, Huang YJ, Gunsalus KC, Szyperski T. Protein NMR spectroscopy in structural genomics. *Nat Struct Biol* 2000;7(Suppl):982–985.
- Liu J, Rost B. Domains, motifs and clusters in the protein universe. *Curr Opin Chem Biol* 2003;7:5–11.
- Ponting CP, Schultz J, Milpetz F, Bork P. SMART: identification and annotation of domains from signalling and extracellular protein sequences. *Nucleic Acids Res* 1999;27:229–232.
- Haft DH, Selengut JD, White O. The TIGRFAMs database of protein families. *Nucleic Acids Res* 2003;31:371–373.
- Henikoff JG, Henikoff S. Blocks database and its applications. *Methods Enzymol* 1996;266:88–104.
- Henikoff JG, Greene EA, Pietrokovski S, Henikoff S. Increased coverage of protein families with the blocks database servers. *Nucleic Acids Res* 2000;28:228–230.
- George RA, Heringa J. Protein domain identification and improved sequence similarity searching using PSI-BLAST. *Proteins* 2002;48:672–681.
- George RA, Heringa J. SnapDRAGON: a method to delineate protein structural domains from sequence data. *J Mol Biol* 2002;316:839–851.
- Wheeler SJ, Marchler-Bauer A, Bryant SH. Domain size distributions can predict domain boundaries. *Bioinformatics* 2000;16:613–618.
- Kulikowski CA, Muchnik I, Yun HJ, Dayanik AA, Zhang D, Song Y, Montelione GT. Protein structural domain parsing by consensus reasoning over multiple knowledge sources and methods. *Medinfo* 2001;10:965–969.
- Murvai J, Vlahovicek K, Szepesvari C, Pongor S. Prediction of protein functional domains from sequences using artificial neural networks. *Genome Res* 2001;11:1410–1417.

41. Miyazaki S, Kuroda Y, Yokoyama S. Characterization and prediction of linker sequences of multi-domain proteins by a neural network. *J Struct Funct Gen* 2002;2:37–51.
42. Enright AJ, Ouzounis CA. GeneRAGE: a robust algorithm for sequence clustering and domain detection. *Bioinformatics* 2000;16:451–457.
43. Marcotte EM, Pellegrini M, Thompson MJ, Yeates TO, Eisenberg D. A combined algorithm for genome-wide prediction of protein function. *Nature* 1999;402:83–86.
44. Marsden RL, McGuffin LJ, Jones DT. Rapid protein domain assignment from amino acid sequence using predicted secondary structure. *Protein Sci* 2002;11:2814–2824.
45. Rigden DJ. Use of covariance analysis for the prediction of structural domain boundaries from multiple protein sequence alignments. *Protein Eng* 2002;15:65–77.
46. Kriventseva EV, Biswas M, Apweiler R. Clustering and analysis of protein families. *Curr Opin Struct Biol* 2001;11:334–339.
47. Yona G, Levitt M. Within the twilight zone: a sensitive profile-profile comparison tool based on information theory. *J Mol Biol* 2002;315:1257–1275.
48. Sasson O, Linial N, Linial M. The metric space of proteins-comparative study of clustering algorithms. *Bioinformatics* 2002;18(Suppl 1):S14–S21.
49. Yona G, Linial N, Tishby N, Linial M. A map of the protein space—an automatic hierarchical classification of all protein sequences. In: Glasgow J, Littlejohn T, Major F, Lathrop R, Sankoff D, Sensen C, editors. Montreal, Canada: AAAI Press; 1998. p 212–221.
50. Krause A, Haas SA, Coward E, Vingron M. SYSTERS, GeneNest, SpliceNest: exploring sequence space from genome to protein. *Nucleic Acids Res* 2002;30:299–300.
51. Heger A, Holm L. Picasso: generating a covering set of protein family profiles. *Bioinformatics* 2001;17:272–279.
52. Wise MJ. The POPPs: clustering and searching using peptide probability profiles. *Bioinformatics* 2002;18(Suppl 1):S38–S45.
53. Kriventseva EV, Servant F, Apweiler R. Improvements to CluSTr: the database of SWISS-PROT+TrEMBL protein clusters. *Nucleic Acids Res* 2003;31:388–389.
54. Li W, Jaroszewski L, Godzik A. Clustering of highly homologous sequences to reduce the size of large protein databases. *Bioinformatics* 2001;17:282–283.
55. Przybylski D, Rost B. Alignments grow, secondary structure prediction improves. *Proteins* 2002;46:195–205.
56. Rost B. Enzyme function less conserved than anticipated. *J Mol Biol* 2002;318:595–608.
57. Sasson O, Vaaknin A, Fleischer H, Portugaly E, Bilu Y, Linial N, Linial M. ProtoNet: hierarchical classification of the protein space. *Nucleic Acids Res* 2003;31:348–352.
58. Yona G, Linial N, Linial M. ProtoMap: automatic classification of protein sequences and hierarchy of protein families. *Nucleic Acids Res* 2000;28:49–55.
59. Linial M, Linial N, Tishby N, Yona G. Global self-organization of all known protein sequences reveals inherent biological signatures. *J Mol Biol* 1997;268:539–556.
60. Carter P, Liu J, Rost B. PEP: Predictions for entire proteomes. *Nucleic Acids Res* 2003;31:410–413.
61. Etzold T, Argos P. SRS—an indexing and retrieval tool for flat file data libraries. *CABIOS* 1993;9:49–57.
62. Amodeo P, Fraternali F, Lesk AM, Pastore A. Modularity and homology: modelling of the titin type I modules and their interfaces. *J Mol Biol* 2001;311:283–296.
63. Liu J, Hegyi H, Acton TB, Montelione GT, Rost B. Automatic target selection for structural genomics on eukaryotes. *Proteins* 2004; In press.
64. Gough J, Karplus K, Hughey R, Chothia C. Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. *J Mol Biol* 2001;313:903–919.
65. Liu J, Rost B. Target space for structural genomics revisited. *Bioinformatics* 2002;18:922–933.
66. Zipf GK. Human behavior and the principle of least effort. Reading: Addison-Wesley Press; 1949.
67. Nielsen H, Brunak S, von Heijne G. Machine learning approaches for the prediction of signal peptides and other protein sorting signals. *Protein Eng* 1999;12:3–9.
68. Sprague ER, Redd MJ, Johnson AD, Wolberger C. Structure of the C-terminal domain of Tup1, a corepressor of transcription in yeast. *EMBO J* 2000;19:3016–3027.
69. Tzamarias D, Struhl K. Functional dissection of the yeast Cyc8-Tup1 transcriptional co-repressor complex. *Nature* 1994;369:758–761.
70. Letunic I, Goodstadt L, Dickens NJ, Doerks T, Schultz J, Mott R, Ciccarelli F, Copley RR, Ponting CP, Bork P. Recent improvements to the SMART domain-based sequence annotation resource. *Nucleic Acids Res* 2002;30:242–244.
71. Serres MH, Gopal S, Nahum LA, Liang P, Gaasterland T, Riley M. A functional update of the Escherichia coli K-12 genome. *Genome Biol* 2001;2:RESEARCH0035.
72. Teichmann SA, Rison SC, Thornton JM, Riley M, Gough J, Chothia C. Small-molecule metabolism: an enzyme mosaic. *TIBTECH* 2001;19:482–486.
73. Dueber JE, Yeh BJ, Chak K, Lim WA. Reprogramming control of an allosteric signaling switch through modular recombination. *Science* 2003;301:1904–1908.
74. Abascal F, Valencia A. Clustering of proximal sequence space for the identification of protein families. *Bioinformatics* 2002;18:908–921.
75. Remm M, Storm CE, Sonnhammer EL. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J Mol Biol* 2001;314:1041–1052.
76. Vlahovicek K, Murvai J, Barta E, Pongor S. The SBASE protein domain library, release 9.0: an online resource for protein domain identification. *Nucleic Acids Res* 2002;30:273–275.
77. Chothia C. One thousand protein families for the molecular biologist. *Nature* 1992;357:543–544.
78. Wolf YI, Grishin NV, Koonin EV. Estimating the number of protein folds and families from complete genome data. *J Mol Biol* 2000;299:897–905.
79. Coulson AF, Moutl J. A unfold, mesofold, and superfold model of protein fold use. *Proteins* 2002;46:61–71.
80. Crippen GM, Maiorov VN. How many protein folding motifs are there? *J Mol Biol* 1995;252:144–151.
81. Finkelstein AV, Ptitsyn OB. Why do globular proteins fit the limited set of folding patterns? *Prog Biophys Mol Biol* 1987;50:171–190.
82. Orengo CA, Jones DT, Thornton JM. Protein superfamilies and domain superfolds. *Nature* 1994;372:631–634.
83. Vitkup D, Melamud E, Moutl J, Sander C. Completeness in structural genomics. *Nat Struct Biol* 2001;8:559–566.
84. Blundell TL, Johnson MS. Catching a common fold. *Protein Sci* 1993;2:877–883.
85. Montelione GT, Anderson S. Structural genomics: keystone for a human proteome project. *Nat Struct Biol* 1999;6:11–12.
86. Burley SK, Almo SC, Bonanno JB, Capel M, Chance MR, Gaasterland T, Lin D, Sali A, Studier FW, Swaminathan S. Structural genomics: beyond the human genome project. *Nat Gen* 1999;23:151–157.
87. Christendat D, Yee A, Dharamsi A, Kluger Y, Savchenko A, Cort JR, Booth V, Mackereth CD, Saridakis V, Ekiel I, Kozlov G, Maxwell KL, Wu N, McIntosh LP, Gehring K, Kennedy MA, Davidson AR, Pai EF, Gerstein M, Edwards AM, Arrowsmith CH. Structural proteomics of an archaeon. *Nat Struct Biol* 2000;7:903–909.
88. Hendrickson WA. Synchrotron crystallography. *Trends Biochem Sci* 2000;25:637–643.
89. Rost B. Marrying structure and genomics. *Structure* 1998;6:259–263.
90. Shapiro L, Lima CD. The Argonne Structural Genomics Workshop: Lamaze class for the birth of a new science. *Structure* 1998;6:265–267.
91. Sali A. 100,000 protein structures for the biologist. *Nat Struct Biol* 1998;5:1029–1032.
92. Thornton J. Structural genomics takes off. *Trends Biochem Sci* 2001;26:88–89.
93. Linial M, Yona G. Methodologies for target selection in structural genomics. *Prog Biophys Mol Biol* 2000;73:297–320.
94. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 1997;25:3389–3402.
95. Eddy SR. Profile hidden Markov models. *Bioinformatics* 1998;14:755–763.
96. Yang AS, Honig B. An integrated approach to the analysis and modeling of protein sequences and structures. II. On the relationship between sequence and structural similarity for proteins that are not obviously related in sequence. *J Mol Biol* 2000;301:679–689.
97. Yang AS, Honig B. An integrated approach to the analysis and modeling of protein sequences and structures. I. Protein structural alignment and a quantitative measure for protein structural distance. *J Mol Biol* 2000;301:665–678.