Automatic Target Selection for Structural Genomics on Eukaryotes

Jinfeng Liu,[1,2,3,4] **Hedi Hegyi,**[1,2] **Thomas B. Acton,**[5,6] **Gaetano T. Montelione,**[5,6] **and Burkhard Rost**[1,2,3*]

[1]*CUBIC, Department of Biochemistry and Molecular Biophysics, Columbia University, New York, New York*

[2]*NorthEast Structural Genomics Consortium (NESG), Department of Biochemistry and Molecular Biophysics, Columbia University, New York, New York*

[3]*Columbia University Center for Computational Biology and Bioinformatics (C2B2), New York, New York*

[4]*Department of Pharmacology, Columbia University, New York, New York*

[5]*Center for Advanced Biotechnology and Medicine (CABM), Rutgers University, and Department of Biochemistry, Robert Wood Johnson Medical School, Piscataway, New Jersey*

[6]*Northeast Structural Genomics Consortium (NESG), Department of Molecular Biology and Biochemistry, Rutgers University, Piscataway, New Jersey*

ABSTRACT A central goal of structural genomics is to experimentally determine representative structures for all protein families. At least 14 structural genomics pilot projects are currently investigating the feasibility of high-throughput structure determination; the National Institutes of Health funded nine of these in the United States. Initiatives differ in the particular subset of "all families" on which they focus. At the North*E*ast Structural *G*enomics consortium (NESG), we target eukaryotic protein domain families. The automatic target selection procedure has three aims: 1) identify all protein domain families from currently five entirely sequenced eukaryotic target organisms based on their sequence homology, 2) discard those families that can be modeled on the basis of structural information already present in the PDB, and 3) target representatives of the remaining families for structure determination. To guarantee that all members of one family share a common foldlike region, we had to begin by dissecting proteins into structural domain-like regions before clustering. Our hierarchical approach, CHOP, utilizing homology to PrISM, Pfam-A, and SWISS-PROT chopped the 103,796 eukaryotic proteins/ORFs into 247,222 fragments. Of these fragments, 122,999 appeared suitable targets that were grouped into >27,000 singletons and >18,000 multifragment clusters. Thus, our results suggested that it might be necessary to determine >40,000 structures to minimally cover the subset of five eukaryotic proteomes. Proteins 2004;56:188–200. © 2004 Wiley-Liss, Inc.

Key words: structural genomics; target selection; protein structure family; cluster; domains; proteome analysis

INTRODUCTION

Structural genomics: determine a structure for each sequence-structure family. In 2000, the National Institute of Health (NIH) in the United States began to finance pilot projects for large-scale protein structure determination (structural genomics).[14] One goal of structural genomics[15–28] is to determine at least one structure for each representative protein family for which a structure cannot be inferred by comparative modeling. An important benefit is the basic understanding of biology and biological processes that will result from the determina-

TABLE I. Structural Genomics Initiatives

Acronym	Name	Country	URL[a]
BSGC	Berkeley Structural Genomics Center	USA	www.strgen.org/
CESG	Center for Eukaryotic Structural Genomics	USA	www.uwstructuralgenomics.org/
JCSG	The Joint Center for Structural Genomics	USA	www.jcsg.org/
MCSG	The Midwest Center for Structural Genomics	USA	www.mcsg.anl.gov/
NYSGRC	New York Structural Genomics Research Consortium	USA	www.nysgrc.org/
NESG	Northeast Structural Genomics Consortium	USA	www.nesg.org/
SECSG	The Southeast Collaboratory for Structural Genomics	USA	www.secsg.org/
SGPP	Structural Genomics of Pathogenic Protozoa Consortium	USA	depts.washington.edu/sgpp/
TB	TB Structural Genomics Consortium	USA	www.doe-mbi.ucla.edu/TB/
S2F	Sequence To Function	USA	s2f.umbi.umd.edu/
BSGI	Montreal-Kingston Bacterial Structural Genomics Initiative	Canada	euler.bri.nrc.ca/brimsg/bsgi.html
SGC	Structural Genomics Consortium	Canada	www.uhnres.utoronto.ca/proteomics/
SPINE	Structural Proteomics in Europe	Europe	www.spineurope.org/
ASG	After Sequencing Genomes	France	afmb.cnrs-mrs.fr/stgen/tglist.html
BIGS	Bacterial targets Genomics and Structural Information	France	igs-server.cnrs-mrs.fr/Str_gen/
NWSGC	North West Structural Genomics Centre	England	www.nwsgc.ac.uk/
OPPF	Oxford Protein Production Facility	England	www.oppf.ox.ac.uk
PSB	Partnership for Structural Biology	France	psb.esrf.fr/
PSF	Protein Structure Factory	Germany	www.rzpd.de/psf/
SGM	Structural Genomics of Micobacteria	France	feu.sis.pasteur.fr/cgi-bin/WebObjects/ MINISGP
WSPC	Weizmann Structural Proteomics Center	Israel	www.weizmann.ac.il/~wspc/
YSG	Yeast Structural genomics	France	genomics.eu.org/
BIRC	Biological Information Research Center	Japan	www.aist.go.jp/aist_e/research_units/ research_center/birc/birc_main.html
RSGI	RIKEN Structural Genomics Initiative	Japan	www.rsgi.riken.go.jp/

[a]URL without http://, e.g., http://www.nesg.org

tion of structural scaffolds for most basic functional elements. Nevertheless, with the advances of many pilot projects, it becomes increasingly apparent that "structures for all families" is only one contribution of structural genomics. Structural genomics also pioneers high-throughput projects targeting proteins rather than genes. This challenge requires the development of techniques and protocols for large-scale expression, purification, crystallization, and structure determination. Such techniques semi or fully automating the work with proteins are likely to simplify many aspects of, for example, biochemistry and cell biology, and to add many techniques from biophysics to the standard battery of tools. Thus, these tools may ultimately become the most profound impact of structural genomics on everyday wet laboratory biology.

One structure per family: simple concept, tough task! The goal to choose one structure per sequence-structure family appears conceptually trivial: 1) establish thresholds for levels of sequence similarity that accurately imply structural similarity[7,29–36], 2) begin from any protein and pull in all those proteins from the sequence universe that have the same structure. Unfortunately, this simple concept fails in practice.[1,37] The detailed reasoning for our conclusion is beyond the scope of this manuscript. However, the two major points are that 1) we have to begin the clustering from fragments that resemble structural domains and that 2) the task at hand is to cluster entire proteomes, rather than to build sequence-structure fami-

lies for selected proteins as in the spirit of the HSSP database[29,38], or the PIR,[39,40] CATH,[41] and SCOP[42] superfamilies. Therefore, we have to first chop proteins as reliably as possible into structural domain-like fragments and to then cluster these fragments before we can systematically choose one fold per sequence-structure family.

NESG focus on eukaryotic domain families. The NorthEast Structural Genomics consortium (NESG http://www.nesg.org/), one of the NIH-funded structural genomics pilot projects, has focused on proteins from the fully sequenced eukaryotic model organisms *Saccharomyces cerevisiae*, *Arabidopsis thaliana*, *Caenorhabditis elegans*, *Drosophila melanogaster,* and *Homo sapiens*. We initially decided to focus on protein targets shorter than 340 residues to reduce the problem of multidomain proteins: >90% of the structural domains in SCOP[42] and PrISM[7] are shorter than 340 residues.[1] The primary goal is to experimentally determine one structure per sequence-structure family not represented in the PDB from these eukaryotic organisms. In this aim, we target, clone, and express proteins from these eukaryotic model proteomes (target proteomes) and in some cases, those from a number of bacterial and archael reagent proteomes that belong to the eukaryotic sequence-structure families.

First-stage automatic target selection at NESG. The first stage of automated target selection has to solve the following four tasks: 1) cluster all proteins from the eukaryotic model organisms so that each cluster repre-

sents one particular fold; 2) for each cluster, pull in those proteins from the non-eukaryotic reagent proteomes that share the fold represented by this cluster; 3) exclude all clusters for which the fold is known; and 4) mark regions that are likely to hamper experimental progress, namely, helical membrane proteins (PHDhtm[43,44]; note that we most likely do not make any major mistake by ignoring β-membrane proteins in eukaryotes[45]), signal peptides (SignalP[11]), proteins dominated by coiled-coil regions (COILS[2]), low-complexity regions (SEG[10]), or long regions without regular secondary structure (NORS[3,4]). All proteins that have <50 residues left after steps 3 (PDB) and 4 (difficult/unwanted) were excluded from further analysis; consequently, entire clusters may be excluded. Note that our procedure does not exclude secreted or membrane proteins; rather, it only excludes proteins that are structurally characterized except for "unwanted" regions such as signal peptides or membrane helices. The resulting clusters comprise the NESG target list. One advantage gained from working with these clusters is realized in the protein production aspect of our project: The differences in sequence between members of a cluster can result in proteins that differ crucially in their characteristics of expression, solubility, crystallizability, and amenability to structure determination by NMR. This multiplex scheme in which many members of each cluster are cloned and expressed in parallel increases the chances of eventually obtaining a sample suitable for structure determination. For example, although a target protein from yeast might not be soluble when overexpressed in *Escherichia coli,* a homologue from *Aquifex aeolicus* might prove soluble in the same system. If purification succeeds for more than one member of a cluster, a second stage of target selection is invoked, the details of which will be described elsewhere (Diana Murray et al., in preparation). For each selected target, the amino acid sequence, nucleic acid sequence, and other key information required for the cloning process is organized for our molecular biology and protein production efforts in the Web-based ZebaView database.[46] Details of progress in cloning, expression, purification, and structure determination for each NESG protein target are then tracked by the SPiNE laboratory information management system.[47] All target clusters are linked to public databases and information about protein structure and function through our PEP database.[48]

Here, we describe the results from the first-stage automatic target selection. To group proteins into sequence-structure families, we first chopped all proteins from the eukaryotic target proteomes and from the reagent proteomes into structural domain-like fragments by the procedure CHOP.[1] CHOP imposes a hierarchy beginning from the most reliable information about structure domains (PrISM[8,9] domains for proteins of known structure), continues to families that are well characterized by experts (Pfam-A[6]), and finally explores the reliable information about N- and C-terminal ends of proteins that are characterized by experimentalists (SWISS-PROT[13,49]). The objective is not to obtain all domain boundaries, but rather, to identify only those boundaries for which we are confident.

Our clustering strategy identified 18,000–21,000 nonsingleton clusters representing the five entirely sequenced eukaryotes. These 18,000–21,000 domain-family clusters can be viewed as the minimal set of protein domains that structural genomics has to determine experimentally for eukaryotes.

RESULTS

Most proteins had more than one fragment. For the 103,796 proteins in the five eukaryotic target proteomes, the CHOP algorithm[1] generated 247,222 fragments; 47% of these resulted from sequence similarity to PrISM, Pfam-A, or to SWISS-PROT termini [Fig. 1(A)]. Only 14% of the final fragments were full-length proteins that remained untouched by our algorithm. To illustrate these percentages by numbers: 34,914 proteins (14% of all fragments and 34% of all proteins) were not chopped, whereas 115,601 fragments (47% of fragments) originated directly from similarities to PrISM, Pfam-A, or SWISS-PROT; another 96,707 fragments (39% of fragments) were left over after chopping. The subset of these untouched "leftover" fragments differed significantly in its length distribution from all full-length proteins [Fig. 1(B)]. The distribution of the number of fragments per protein differed slightly from that for 62 entirely sequenced proteomes. In particular, for the subset of all proteins that were chopped by our algorithm, 28% had only one fragment in the set of all 62 proteomes[1] and about 19% in the five eukaryotic proteomes; about 1% of the chopped eukaryotic proteins—corresponding to 1026 proteins—had >10 fragments [Fig. 1(C)]. On average, the number of CHOP fragments was directly proportional to the protein length [fit: average length of protein = 202 + 103 * number of fragments, Fig. 1(D) open circles]. This linear fit for the average length of a structural domain-like fragment unravelled two rather remarkable features. The first was that most fragments extend over ~100 residues. Although the current PDB is biased toward certain types of proteins, in particular, proteins that are shorter than the averages observed from entire genome sequencing, the corresponding fit for structurally known protein domains—defined by PrISM—was parallel [fit: average length of protein = 65 + 97 * number of fragments, Fig. 1(D) crosses]. Thus, most domains in multidomain proteins in PDB were also ~100 residues long. The second remarkable feature was that the linear fit for both the eukaryotic proteomes and the PDB did not begin at 0 but at 65 for PrISM/PDB domains and at 202 for our eukaryotic fragments. Thus, N-1 domains in proteins with N domains extend >100 residues, whereas one extends >160–300 residues. 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). 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
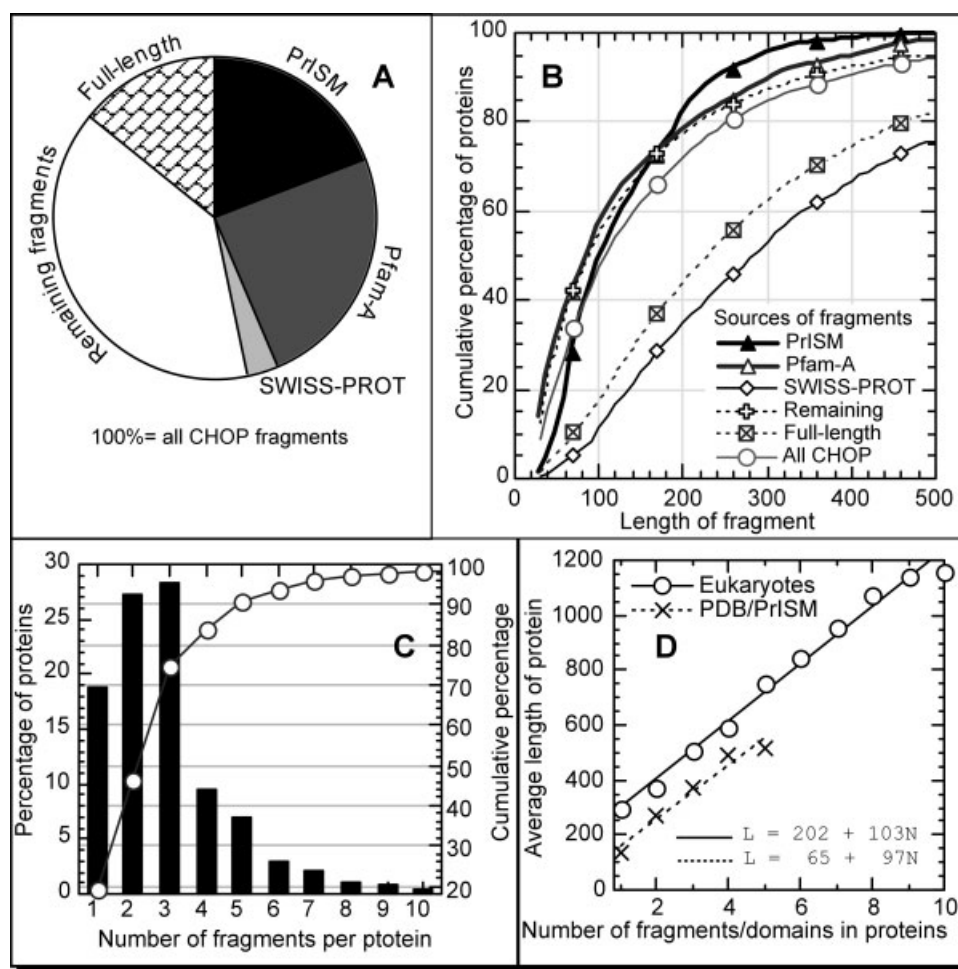
Fig. 1. Statistics on eukaryotic CHOP fragments. **A:** Percentage of all CHOP fragments (including remaining regions): 47% of all the final CHOP fragments originated from cuts according to PrISM, Pfam-A, or SWISS-PROT; 14% of the "fragments" were full-length proteins untouched by CHOP. **B:** Note that 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. All CHOP fragments (gray with open circles) were similar to the subset of Pfam-A fragments. Fragments cut according to SWISS-PROT termini (5%, Fig. 2) were more similar to the distribution of all full-length proteins from the eukaryotic proteomes (not shown) than the subset of proteins that remained untouched. **C:** Distribution of number of CHOP fragments per protein (as percentage of all proteins that were chopped). For example, we found that <20% of the eukaryotic proteins had a single structural domain-like fragment. **D:** Relation between number of fragments and protein length: on average, the number of CHOP fragments appeared to increase linearly with protein length (open circles). The basic functional form for this plot was similar for domains from proteins of known structures (crosses, taken from PrISM). The lines fit the data with L = a + b N, with L being the length of the protein and N the number of fragments.

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.

Half of all fragments were not suitable for structural genomics. For each of the 247,222 eukaryotic structural domain-like fragments generated by CHOP, we searched for similarities to PDB structures (Methods) and applied a variety of prediction methods. The objective of this step was to exclude fragments from further analysis that either had known structure or constituted low-priority targets for structural genomics; 167,717 fragments did not match to any known structure. To filter out

potentially difficult cases for structure determination (low-priority targets), we discarded all fragments that were dominated by predicted membrane helices,[44] coiled-coil helices,[2] low-complexity regions,[10] long regions of low secondary structure contents (NORS regions[3,4]), and those of insufficient length (<50 residues). The precise criterion to accept a fragment was that we found at least 50 consecutive residues without known structure and without any of the "problematic" regions listed; this step left 122,999 globular eukaryotic fragments. Albeit conceptually easy, our criterion for exclusion of fragments made it impossible to directly investigate the relative contribution
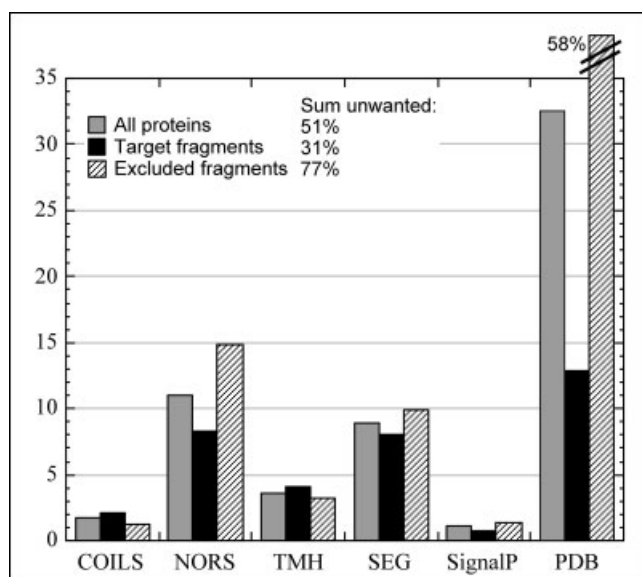
Fig. 2.   Statistics on exclusion of CHOP fragments for NESG. Although we cannot predict from sequence which proteins are the "best" targets for structural genomics, we can predict which ones are likely not suitable. Most obviously, these are proteins with known structure and proteins with many regions that hamper high-throughput structure determination because they have long coiled-coil (COILS) or membrane helices (TMH), long regions without regular secondary structure (NORS), or with low-complexity regions (SEG), or fragments that basically contain only a signal peptide (SignalP). Before clustering CHOP fragments, we excluded all segments that would clearly not be suitable. Our criterion for exclusion was simply that we could not find at least 50 consecutive residues without any unwanted region. Note that this definition implies that we did choose targets with "problematic regions." In particular, many targets originated from secreted proteins; only the signal peptides were excluded. Here, we compared the percentage of residues in such unwanted regions for all eukaryotic target proteins (gray bars) with that of the excluded fragments (stippled bars) and the target fragments clustered in this work (black bars).

of each of the possible reasons for exclusion. For example, a fragment might have residues 1–40 in a signal peptide, 60–65, 70-75, and 80–85 in SEG, and a membrane helix close to the C-terminus. However, we could measure the per residue contribution of each reason for exclusion: overall, our filtering procedure considered ~51% of the residues in all eukaryotic proteins as either "known structure" or as "unwanted"; this percentage was 46% higher for fragments excluded before clustering (77%) than for the 122,999 fragments considered further (31%, Fig. 2). Most "unwanted" residues originated from similarity to very short segments of known structure (58% of the residues in excluded fragments). The second most common reason for exclusion was the presence of a long region lacking regular secondary structure (NORS). The contribution from signal peptides (SignalP), membrane helices (TMH) and coiled-coil helices (COILS) was rather small in comparison: NORS+PDB accounted for 95% of the "unwanted" residues in excluded fragments. Because our condition for inclusion was that we could find at least one region of ≥50 consecutive residues without unwanted residues, the 122,999 fragments used for clustering still contained considerable fractions of such residues (31%).

Simple clustering on CHOP fragments yielded reasonable groups. We grouped all 122,999 eukaryotic fragments that constituted potential targets for structural genomics by a simple clustering procedure (CLUP[1]). Most resulting clusters (27,669) were singletons (i.e., contained only one fragment); 21,309 of the clusters contained multiple eukaryotic fragments. Of these 21,309 nonsingleton clusters, about one half contained three or more members, and ~13% had >10 eukaryotic fragments [Fig. 3(A), dark gray]. Although NESG aims at experimentally determining structures for eukaryotic proteins (target proteomes), we also target homologues of these protein domain families from non-eukaryotic reagent proteomes (Table II). Prokaryotic members of these domain families are often easier to produce in *E. coli* expression systems and often correspond to full-length versions of domains that occur within large multidomain eukaryotic proteins. On average, the non-eukaryotic reagent proteomes contributed fewer members to each cluster than the target proteomes [Fig. 3(A), light gray gives reagent + target proteomes]; 143 clusters contained >100 fragments; the largest cluster contained 643 fragments; the seed of this cluster was the worm protein caeel_fr26007 annotated by Pfam-A (PF00153) as "Mitochondrial carrier protein." By definition, all members of one CLUP cluster share one region that might constitute a common fold. This implies that the same fragment can belong to more than one cluster [Fig. 3(B)]. One reason could be that the fragment actually consists of two structural domains that were not recognized as such by CHOP. Most fragments mapped to a single cluster (74%), and only 1% were associated with more than five clusters [Fig. 3(B)]. The nonsingleton clusters united a total of 94,678 fragments; 21,290 of these constituted full-length proteins that had been left untouched by CHOP, and only 2,817 of these were fragments from proteins for which CHOP identified a single domain-like region.

DISCUSSION AND CONCLUSIONS

Do structural domains constitute the "atom of evolution"? Although we failed to cluster full-length proteins,[1] even our simple clustering strategy yielded a reasonable grouping for domain-like fragments. All members of one cluster share a region that is likely to constitute a common fold. The number of fragments that belong to more than one cluster (i.e., the degeneracy of our clusters) reflects the success of combining CHOP and CLUP: 74% of all eukaryotic fragments amendable to structural genomics mapped exclusively to one eukaryotic cluster [Fig. 3(B)]. When we applied CLUP to known structural domains (PrISM), the degeneracy was negligible[1]; these data may suggest that the remaining degeneracy is largely caused by CHOP being incomplete: 14% of all fragments did not match to known domain-like regions [Fig. 1(A)]. However, the lack of degeneracy for PrISM domains might also just be a size effect (i.e., we might observe a higher degeneracy if PDB were 10 times larger). We also evaluated the degeneracy when we merged all clusters that had at least two in three fragments in common (data not
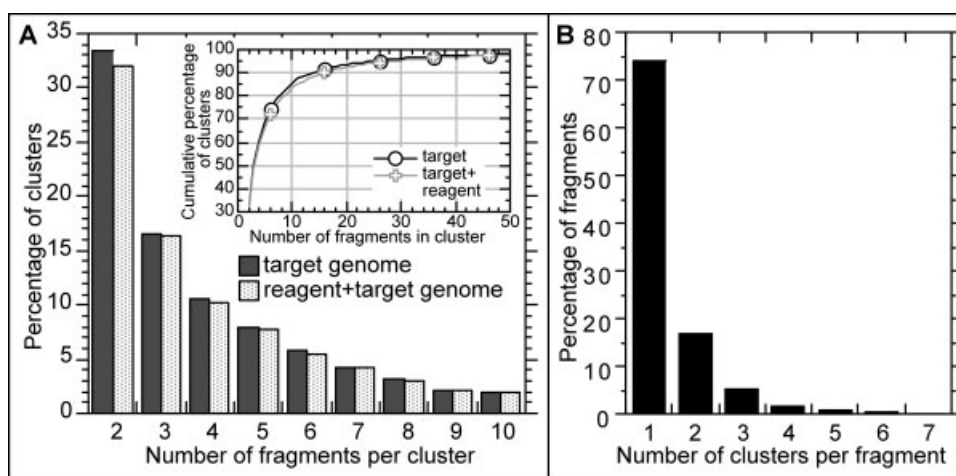
Fig. 3. Statistics on target clusters. **A:** Number of fragments per target cluster for the eukaryotic target proteomes (dark) and the target + reagent proteomes (light). More than one third of the nonsingleton clusters contains two fragments; about half of all clusters have more than three members (inlet: cumulative percentages). **B:** Degeneracy of clusters. Most CHOP fragments (74%) are associated with a single cluster, whereas ~1% of the proteins are associated with more than five clusters. Note that only those clusters that constitute valid targets for structural genomics were considered.

TABLE II. Target and Reagent Proteomes at NESG[†]

Organism	No. of ORFs	No. of CHOP fragments	No. of CLUP clusters
Eukaryotic target proteomes			
Arabidopsis thaliana	17,401	26,867	8,487
Caenorhabditis elegans	12,390	18,094	8,529
Drosophila melanogaster	7,655	10,994	8,152
Homo sapiens	22,547	34,413	12,605
Saccharomyces cerevisiae	3,120	4,401	3,225
Archaeal reagent proteomes			
Aeropyrum pernix K1	296	334	430
Achaeoglobus fulgidus	317	364	440
Methanobacterium thermoautotrophicum	292	330	401
Pyrococcus furiosus	290	340	369
Pyrococcus horikoshii	261	296	364
Thermoplasma acidophilum	235	265	330
Prokaryotic reagent proteomes			
Aquifex aeolicus	304	380	511
Bacillus subtilis	470	535	527
Brucella melitensis	260	309	340
Campylobacter jejuni	193	220	274
Caulobacter crescentus	450	517	540
Deinococcus radiodurans	390	451	476
Escherichia coli	694	812	747
Fusobacterium nucleatum	224	267	316
Haemophilus influenzae	226	278	341
Helicobacter pylori	200	233	289
Lactococcus lactis	248	285	375
Neisseria meningitides	251	296	385
Staphylococcus aureus	312	357	475
Streptomyces coelicolor	827	941	752
Streptococcus pyogenes	192	227	348
Thermotoga maritime	275	325	432
Vibrio cholerae	332	405	488

[†]All numbers refer to the subset of proteins in our final 22,037 nonsingleton clusters. Note that the number of clusters to which each proteome contributes does not sum to the number of all target clusters, because all the clusters considered have more than one member.

shown). Although such a permissive merging decreased the degeneracy considerably (91% only in 1 cluster, and only 1% in >3), after such merging, we can no longer ascertain that all members in one cluster share a common fold. Our combined chopping and clustering strategy implicitly assumed that structural domains constitute some-

thing like the "atoms of evolution." This becomes evident when considering the alternative. If evolution proceeded by a cut-and-paste mechanism of units shorter than structural domains—as recently suggested[50,51]—a simple clustering would not separate the groups in the sense of 0 degeneracy. Presumably, the degeneracy that we observe to some extent originated from the partial error in our initial assumption and to some extent from the fact that often does move by cut-and-past of subdomains.[50,51] Structural domains may not be "the atom of evolution"; however, our data suggested that even our incomplete chopping procedure constituted a rather successful starting point on the quest for evolutionary units. Because most protein–protein interactions are between single domains,[52–56] our structural domain-like fragments may also help reduce noise in two-hybrid experiments by probing protein–protein interactions between fragments rather than between full-length proteins.

Thresholds optimized for high rate of sequence-unique structures. Although our automatic target selection strategy is conceptually rather simple, it required choosing many more or less arbitrary thresholds about when to consider a similarity sufficient to chop or to cluster and when to exclude fragments. Overall, our results were rather stable with respect to minor changes in these thresholds with one prominent exception: the sequence similarity to known PDB structures that exclude fragments. Toward this end, we applied a threshold (PSI-BLAST expectation value of 1 or an HSSP value of 0^{34}) that is fairly conservative in the sense that we are likely to exclude "trivial" similarities. The success of this choice is apparent in the high percentage of structures determined by NESG for proteins that could not have been modeled by homology: 59% of the solved structures constituted sequence-unique proteins (Fig. 4); this was >7 times higher than the corresponding percentage for the entire PDB from the same period (and it was the second highest for all structural genomics consortia, surpassed only by the S2F structure-to-function consortium headed by John Moult[57]). Although successful to avoid overlap with known structures, these thresholds are far too permissive to ascertain that all protein pairs with this level of similarity are within homology-modeling distance of each other. Therefore, we had to choose a different threshold (PSI-BLAST E value $< 10^{-3}$) to assign proteins to a particular cluster. At this level, comparative modeling correctly predicts about 30–50% of the proteins at a main-chain root-mean-square deviation (RMSD) of ~2–5 Å (Marc Marti-Renom, data unpublished[58,59]). Assume that a structural biologist determines the structure for one of the fragments in our cluster X. Should we automatically remove that cluster from our list? Obviously, our threshold for grouping is still too permissive to rule out that we would benefit from determining additional members of X. Therefore, NESG carries out an additional step, namely, an expert-driven manual detailed examination of cluster X that aims at providing a more reliable answer to the question of whether multiple structures are required to characterize the entire cluster (Diana Murray, Cornell, unpublished).

Structural genomics already contributed many sequence-unique structures. About 8,000 structures containing a total of about 14,000 chains were added to the PDB while structural genomics consortia have existed; 1,100 of these chains (~8%) were sequence-unique,[5,60] based on the criteria described above. One way of evaluating the success of structural genomics pilot projects in the United States over the first 3 years of funding is by measuring the percentage of new sequence-unique structures determined by all pilot projects (49%; Fig. 4). (Note that only the U.S. projects have consistently deposited their data into TargetDB, the special database for structural genomics provided by the PDB.) To illustrate the impact of this high number, although structural genomics projects in the United States contributed fewer than 3% of all structures, they solved almost 20% of the sequence-unique structures deposited into PDB in that period. Nevertheless, given that, for instance, the NESG threshold for excluding proteins with similarities to known structures from the target list is very permissive (E value of 1), it seems surprising that the rate of sequence-unique structures from NESG was—albeit the highest for all consortia— "only" 67%. Part of the reason is a legacy of initial "technology development" targets that were selected and brought into the structure analysis pipeline before the target selection process outlined here had been developed. Indeed, almost 80% of the targets selected by the initial realization of the concept described here were sequence-unique—still not 100% because we considered targets as sequence-unique at the time of deposition in the PDB rather than at the time of selection (we did not have this information for all consortia). Thus, by this definition, the structural genomics consortia compete with each other and all other structural biologists. Because many consortia make an effort to choose "high-leverage targets" (as demonstrated by the above average values for the leverage of each structure deposited, Fig. 4 upper panels), the high rate of sequence-unique structure illustrates another aspect of success: speed. Although the consortia differ in their focus, they obviously overlap in the attempt to prioritize targets that appear promising and interesting. This may explain the high degree of overlap between the consortia (on average 50%; Fig. 5). In fact, the smallest U.S. consortium (S2F) reached the second highest ratio of sequence-unique structures because S2F selected proteins not considered by others and/or determined structures faster than the others.

Alarm system detecting recently solved structures: work in progress. Each week we compare all proteins added to PDB against our target PSI-BLAST profiles. If we find a new structure that has sequence similarity to any of our targets, we notify the group of Diana Murray at Weil Medical College of Cornell University, members of the NESG, who investigate the case in more detail. We currently apply three different thresholds, depending on the experimental stage of the target. 1) If the target is already expressed, soluble, and purified and we have a good or promising first HSQC spectrum and/or a promising crystal, we require a similarity to a known structure at either a
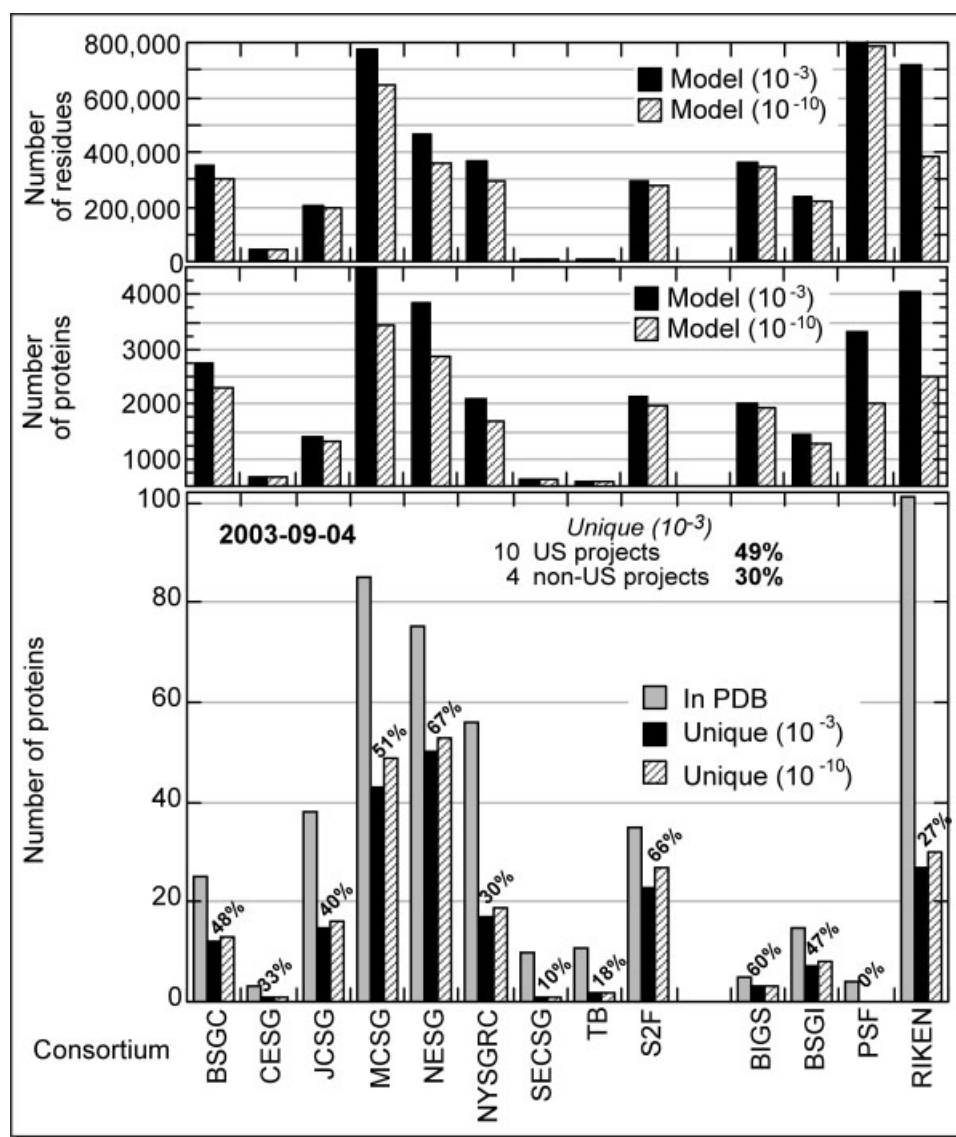
Fig. 4.   Early success of structural genomics. One goal of structural genomics is to determine structures for proteins for which we cannot predict structure through comparative modeling. The lower panel shows the number of structures deposited into the PDB (light gray) by structural genomics consortia (list Table I); the dark and stippled bars show the subsets of these for which comparative modeling was not applicable at the time of deposition assuming that we can build all models when a homologue in PDB is similar at a BLAST expectation value of 10^{-3} (black) or 10^{-10} (stippled); the numbers on top of the black bars give the percentage of sequence-unique proteins at 10^{-3}. About 49% of all the structures solved by all 10 U.S. consortia could not have been modeled (for comparison: the corresponding number for the entire PDB for the same time period was ~8%; percentage values on top of black bars in lower graph). Thus, in their second or third year of progress, the structural genomics consortia are already on track. The middle panel gives a coarse-grained estimate for an upper limit on the number of proteins that could possibly be modeled (dark bars assume we can model all proteins with a PSI-BLAST E value $< 10^{-3}$; stippled bars mark E values $< 10^{-10}$); the upper panel shows the same information for the number of residues modeled in all these proteins. These leverage values show the impact of a single well-chosen experimental structure: all consortia determined 189 unique structures that might yield new models—at least at low resolution—for ~23,000 proteins and >5 million residues (i.e., on average each unique structure gave rise to 120 new models). However, the number of sequence-structure families with >600 members is rather limited; <40% of all families have >100 members.[64,71] Consequently, the leverage values will decrease with increasing success of structural genomics in structurally covering all sequence-structure families. Thus, leverage values may not constitute the most reasonable measures of success for structural genomics.

PSI-BLAST E value $< 10^{-10}$ or at an HSSP proximity > 10 to "stop work" on the corresponding target. 2) If the target is cloned, expressed, soluble, and purified, but has not yet yielded success in preliminary efforts of structure analysis, the thresholds are E value $< 10^{-3}$ or HSSP proximity > 2 for "stop work." 3) If the target has not been touched
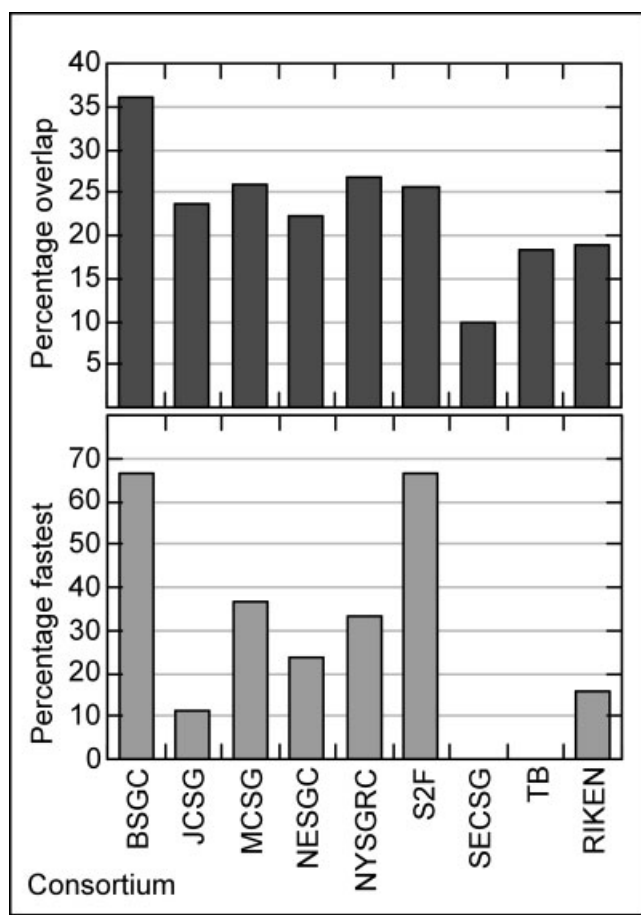
Fig. 5. Overlap between structural genomics consortia. The fraction of sequence-unique proteins in Figure 4 implicitly reflects the overlap between the structural genomics consortia and all structure deposited into the PDB (overlap means PSI-BLAST E values below 10^{-3}). Here, we addressed the question to which extent the structural genomics consortia overlapped (10^{-3}) with each other (top panel: number of proteins that overlap as a percentage of the number of proteins deposited in PDB by 2003-03-26). Although all consortia have similar implicit constraints (as many sequence-unique structures as possible to substantiate funding), the overall overlap remained below 30% (84 of 315 proteins deposited into the PDB by 2003-03-26). The consortia differed substantially in the percentage of targets for which they overlap and deposited the structure faster than any other consortium (lower panel: number of proteins determined before all other consortia as percentage of number of overlapping proteins).

experimentally or has been cloned but not yet expressed in soluble form, we simply remove it from the target list for E values < 1 and for HSSP proximity > 0. These values reflect our experiences with the reliability of inferring fold similarity from sequence. However, there are no large-scale data available yet, how these thresholds may be revised in the context of our efforts to cover the variety of significant local changes of a given fold within one sequence-structure family. The Murray group is currently testing our preliminary "stop-work" thresholds. In addition, given the large number of targets processed by the NESG, the task of identifying targets with significant homology to recently determined structures would be overwhelming if done manually. Therefore, we have created two automated

systems to detect such events, one of which is embedded into our official target Web site ZebaView (http://www-nmr.cabm.rutgers.edu/bioinformatics/ZebaView/). Every week, ZebaView blasts targets that have progressed into the data collection phase against the PDB, PDB-on-hold, and the TargetDB; any significant hits (E values $< 10^{-3}$) are alerted.

Over 40,000 targets for five eukaryotes, alone! The largest funding for structural genomics worldwide originates from the Science & Technology Agency in Japan and is concentrated at the RIKEN Structural Genomics Initiative (RSGI) at the Institute of Physical and Chemical Research.[61] The second largest funding originates from the National Institute of General Medical Sciences (NIGMS) at the National Institutes of Health (NIH) in the United States. The NIGMS protein structure initiative (PSI) formulates as one of the goals of structural genomics "to determine representative structures from all protein families."[14] The goal of our automatic target selection is exactly along this line: 1) find all representative families from entirely sequenced eukaryotes, 2) exclude those for which we already have knowledge about structure and which are likely to hinder progress in a high-throughput enterprise, and 3) develop the means that allow rapid structure determination for each of these families. We were surprised by the amount of technical difficulties that we had to surmount to approximate a solution for the tasks imposed on target selection by the first two steps; the results that we presented here refine in many ways the simple concepts laid out earlier.[62–66] Possibly the most surprising result of our work is summarized by the following two numbers: 27,669 singleton and 18,079 (after permissive merging of clusters) to 21,309 nonsingleton clusters, summing up to >40,000 proteins for which we have to determine structure if we want to "minimally cover" the proteomes of only five eukaryotes. In fact, these numbers still underestimate the actual workload because they ignore membrane regions, domains that appear depleted of regular secondary structure,[3,4,67,68] as well as proteins that clearly have similar folds but differ in function. About 6700 clusters contain at least one full-length protein (untouched by CHOP) that falls into the NESG length range and belonged to only one cluster; most of these target clusters are already in the experimental pipeline of NESG. Currently, we work on ways to connect clusters through threading. Although such connections will be very unreliable, they may enable us to further raise the probability of spanning as many structural families as possible with as few structures as we can determine experimentally. Nevertheless, at the current rate of structure determination, structural genomics projects will not run out of targets for many years to come.

METHODS
Sequences From Entirely Sequenced Organisms

We obtained all ORFs for the archael and prokaryotic reagent proteomes for *Saccharomyces cerevisiae* and for *Arabidopsis thaliana* from the NCBI Web site: ftp://ftp.ncbi.nih.gov/genbank/genomes/. For the remaining eu-

karyotic proteomes, we used the following sources: *Homo sapiens*, from SWISS-PROT (release 39) and TrEMBL (release 18), *Drosophila melanogaster* from http://www.fruitfly.org/ (release 2), and *Caenorhabditis elegans* from http://www.sanger.ac.uk/Projects/C_elegans/wormpep/ (wormpep 65).

Database Searches and Prediction Methods

Search for similar proteins. We detected similar sequences in the following way: Run PSI-BLAST[69] searches against all known sequences contained in SWISS-PROT,[13,49] TrEMBL,[13,49] and PDB.[5] For simplicity, we refer to the combination of these three databases as the set BIG. We first searched against a "filtered" version of BIG (regions of low complexity were marked by SEG[10]; proteins with >98% pairwise sequence identity were removed) and then used the final profile to search against the unfiltered BIG.[70,71] When building a sequence-structure family, we included all hits below a PSI-BLAST E-value of 10^{-3} or above an HSSP value of 0.[34] The HSSP curve relates alignment length to pairwise sequence identity or similarity[29,34]; for alignments of 100 residues, HSSP = 0 corresponds to 33% pairwise sequence identity; for alignments longer than 250 residues, to about 20%; we refer to values > 0 as HSSP proximity (the larger the more similar) and to values < 0 as HSSP distance (the larger the more distant).

Membrane proteins. We used only the filtered MaxHom alignments[29,34] for predicting membrane regions by the program PHDhtm[43,44] using the default threshold of 0.8. (Note our notion of "membrane proteins" is restricted to integral helical membrane proteins.) In particular, we ignored proteins anchoring helices in the membrane because these classes of proteins cannot be identified generally from sequence information alone. We also ignored proteins inserting β-strands (porins) into the membrane because 1) these proteins are not assumed to exist in any major fraction in any eukaryote[45] and 2) no method that is currently publicly available detects all these proteins with sufficient reliability.

Signal peptides. We predicted signal peptides by using the program SignalP.[11,12] We considered a protein to contain a signal peptide if the "mean S" value in the prediction was above the default threshold. The accuracy of SignalP was estimated to be around 90%.[11,12,72] We excluded archae-bacterial reagent proteomes from the analysis because SignalP was developed for prokaryotes and eukaryotes. As for all other "problematic regions," the presence of a signal peptide did not exclude the respective protein from our target list; rather, we only excluded the signal peptide itself and proteins for which, except for the signal peptide, no region of interest was longer than 50 residues.

Coiled-coil helices. We used the program COILS[2] to predict coiled-coil region, with the window size set to 28 residues and the threshold for probability set to 0.9.

Low-complexity (SEG) and nonregular secondary structure (NORS). We labeled regions of low complexity by using the program SEG[10] using the default parameters.

Using the filtered MaxHom alignments, we used PROFsec[43,73–75] to predict secondary structure and PROFacc[43,75,76] to predict solvent accessibility. We considered stretches of >70 consecutive residues with <12% predicted helix or strand as "NORS."[3,4]

Chopping into domain-like fragments (CHOP) and clustering (CLUP).

CHOP and CLUP (i.e., our methods for dissecting proteins into structural domain-like fragments and for clustering these fragments) are described elsewhere.[1] Here, we only sketched the idea of both procedures (Fig. 6 gives a simplified flowchart for the entire automatic selection stage at the NESG).

Domain-like fragments. CHOP implements three hierarchical steps that were applied by decreasing confidence in the accuracy of the information: 1) high reliability: PrISM domains, 2) acceptable confidence: Pfam-A families, and 3) protein termini from SWISS-PROT. 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. We applied CHOP to all proteins in the five eukaryotic target proteomes and the 23 reagent proteomes.

Clustering CHOP fragments. We clustered all the fragments (and uncut full-length) proteins obtained from CHOP (except those that were removed; see below). Toward this end, we simply ran an all-against-all PSI-BLAST[69] search with a threshold E value of $<10^{-3}$ Then we applied a simple clustering algorithm starting from the shortest fragments in the groups with fewest members identified by the previous PSI-BLAST. Finally, we merged all clusters that have >10 members and share 90% of their members.

Conservative and permissive merging of CLUP clusters. In our final step, we merged all clusters that had >10 members and shared 90% of their members. We introduced this step after analyzing visually the relations between the largest degenerate clusters. For all examples that we looked at, this particularly conservative threshold did not challenge our goal that all members of a given cluster have a structurally similar region that is likely to constitute a structural domain. We also analyzed a more permissive merging strategy by joining all clusters that shared at least 50% of their members (2 in 3 for clusters with only 3 members).

Thresholds and Exclusion of Fragments

Removing fragments. Many proteins of known structure contain regions of low complexity.[3,68] However, proteins that contain almost no high-complexity regions constitute—at best—low-priority targets for structural genomics. Before clustering, we removed all fragments that had fewer than 50 residues in nonmembrane, noncoiled, nonsignal peptide, non-SEG, or non-NORS regions.

Different thresholds for different objectives. Our procedure required introducing different thresholds for sequence similarity depending on whether we wanted to chop proteins into fragments, to ascertain that two pro-
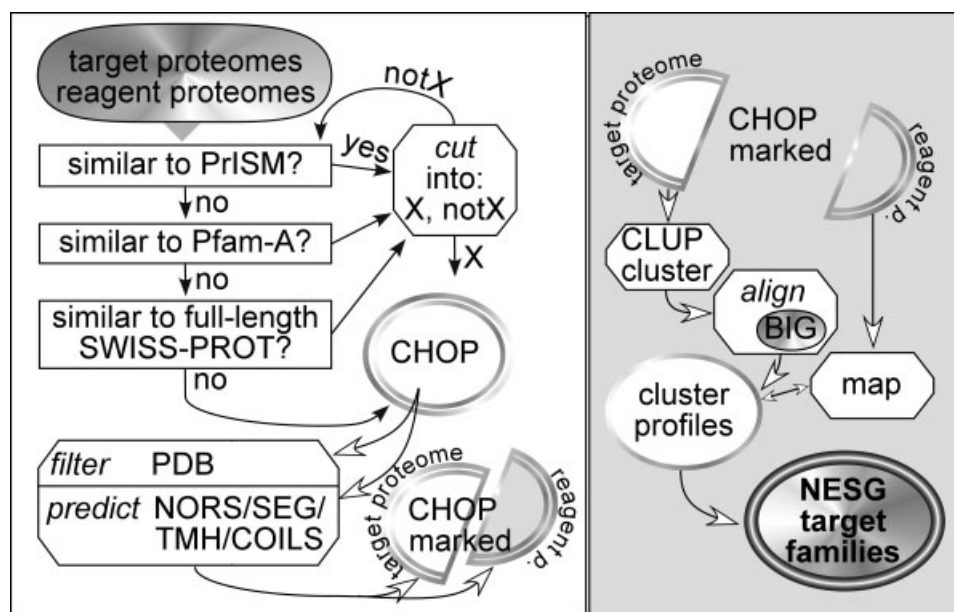
Fig. 6.   Simplified flowchart for automatic target selection stage at the NESG. Our first objective is to dissect as many of the proteins from the target and reagent proteomes (Table II) into structural domain-like fragments (CHOP procedure[1]) and to label these fragments (left panel). CHOP imposes a hierarchy in the sense that it explores first the most reliable information (PrISM domains of known structure), then the next reliable (Pfam-A regions annotated by experts), and finally pulls in the least reliable information (experimentally verified full-length proteins from SWISS-PROT). At each step, proteins are cut into the region that is homologous to any of the three sources (PrISM, Pfam-A, SWISS-PROT, labeled X in the flowchart) and the fragments left and right of this cut (labeled notX). Fragments shorter than 30 residues are removed after this cut. Next, we filter out all CHOP fragments that match to known structures and mark the regions in CHOP fragments from the eukaryotic target proteomes that may pose problems to rapid structure determination (NORS, SEG, TMH, COILS). Fragments that have <50 consecutive residues without any such "problematic" regions are also removed at this step. (Note one of the many additional complications left out here is that we based these predictions on multiple alignments rather than on single sequences.) The second major objective is to group the CHOP fragments into domain-like sequence-structure families and to map proteins from the non-eukaryotic reagent proteomes into these families (right panel). We first cluster all CHOP fragments from the eukaryotic target proteomes (CLUP[1]). Then we generate PSI-BLAST profiles for each cluster and finally map the non-eukaryotic reagent proteins into these cluster families by aligning them into the CLUP PSI-BLAST profiles.

teins have a common fold, or to guarantee that a target cluster is really unlikely to be modeled by comparative modeling. CHOP was restricted to pairwise BLAST E values $< 10^{-2}$, >80% of the PrISM domain, and to E $< 10^{-2}$ in HMMER[77] for Pfam-A regions. We included proteins into a sequence-structure family when they overlapped at least 50 residues, with the representative chosen by CLUP at a similar threshold for sequence similarity as used to chop.

REFERENCES

1. Liu J, Rost B. CHOP proteins into structural domain-like fragments. Proteins 2003. In press.
2. Lupas A. Prediction and analyis of coiled-coil structures. Methods Enzymol 1996;266:513–525

3. Liu J, Tan H, Rost B. Loopy proteins appear conserved in evolution. J Mol Biol 2002;322:53–64.
4. Liu J, Rost B. NORSp: predictions of long regions without regular secondary structure. Nucleic Acids Res 2003;31:3833–3835.
5. Berman HM, Westbrook J, Feng Z, Gillliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res 2000;28:235–242.
6. Bateman A, Birney E, Cerruti L, Durbin R, Etwiller L, Eddy SR, Griffiths-Jones S, Howe KL, Marshall M, Sonnhammer EL. The Pfam protein families database. Nucleic Acids Res 2003;30:276–80.
7. 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;679–689.
8. 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.
9. 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.
10. Wootton JC, Federhen S. Analysis of compositionally biased regions in sequence databases. Methods Enzymol 1996;266:554–571.
11. Nielsen H, Engelbrecht J, Brunak S, von Heijne G. Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. Protein Eng 1997;10:1–6.
12. 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.
13. Bairoch A, Apweiler R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic Acids Res 2000;28:45–48.
14. NIGMS Protein structure initiative. National Institute of General Medical Sciences (NIGMS), 2003, Available online at http://www.nigms.nih.gov/psi.
15. Gaasterland T. Structural genomics taking shape. Trends in Genetics 1998;14:135.
16. Rost B. Marrying structure and genomics. Structure 1998;6:259–263.
17. Sali A. 100,000 protein structures for the biologist. Nat Struct Biol 1998;5:1029–1032.
18. Shapiro L, Lima CD. The Argonne Structural Genomics Workshop: Lamaze class for the birth of a new science. Structure 1998;6:265–267.
19. Brenner SE, Barken D, Levitt M. The PRESAGE database for structural genomics. Nucleic Acids Res 1999;27:251–253.
20. 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.
21. Montelione GT, Anderson S. Structural genomics: keystone for a Human Proteome Project. Nat Struct Biol 1999;6:11–12.
22. Teichmann SA, Chothia C, Gerstein M. Advances in structural genomics. Curr Opin Struct Biol 1999;9:390–399.
23. 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, Gestein M, Edwards AM, Arrowsmith CH. Structural proteomics of an archaeon. Nat Struct Biol 2000;7:903–909.
24. Hendrickson WA. Synchrotron crystallography. Trends Biochem Sci 2000;25:637–643.
25. Montelione GT, Zheng D, Huang YJ, Gunsalus KC, Szyperski T. Protein NMR spectroscopy in structural genomics. Nat Struct Biol 200;7:982–985.
26. Moult J, Melamud E. From fold to function. Curr Opin Struct Biol 2000;10:384–389.
27. Skolnick J, Fetrow JS, Kolinski A. Structural genomics and its importance for gene function analysis. Nat Biotechnol 2000;18:283–287.
28. Thornton J. Structural genomics takes off. Trends Biochem Sci 2001;26:88–89.
29. Sander C, Schneider R. Database of homology-derived structures and the structural meaning of sequence alignment. Proteins 1991;9:56–68.
30. Abagyan RA, Batalov S. Do aligned sequences share the same fold? J Mol Biol 1997;273:355–368.
31. Park J, Teichmann SA, Hubbard T, Chothia C. Intermediate sequences increase the detection of distant sequence homologies. J Mol Biol 1997;273:349–354.
32. Brenner SE, Chothia C, Hubbard TJP. Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. Proc Natl Acad Sci USA 1998;95:6073–6078.
33. Park J, Karplus K, Barrett C, Hughey R, Haussler D, Hubbard T, Chothia C. Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. J Mol Biol 1998;284:1201–1210.
34. Rost B. Twilight zone of protein sequence alignments. Protein Eng 1999;12:85–94
35. Li W, Pio F, Pawlowski K, Godzik A. Saturated BLAST: an automated multiple intermediate sequence search used to detect distant homology. Bioinformatics 2000;16:1105–1110.
36. Blake JD, Cohen FE. Pairwise sequence alignment below the twilight zone. J Mol Biol 2001;307:721–735.
37. Liu J, Rost B. Domains, motifs and clusters in the protein universe. Curr Opin Chem Biol 2003;7:5–11.
38. Schneider R, de Daruvar A, Sander C. The HSSP database of protein structure-sequence alignments. Nucleic Acids Res 1997;25:226–230.
39. Barker WC, Garavelli JS, McGarvey PB, Marzec CR, Orcutt BC, Srinivasarao GY, Yeh LS, Ledley RS, Mewes HW, Pfeiffer F, Tsugita A, Wu C. The PIR-International Protein Sequence Database. Nucleic Acids Res 1999;27:39–43.
40. Wu CH, Yeh LS, Huang H, Arminski L, Castro-Alvear J, Chen Y, Hu Z, Kourtesis P, Ledley RS, Suzek BE, Vinagaka CR, Zhang J, Barker WC. The Protein Information Resource. Nucleic Acids Res 2003;31:345–347.
41. 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.
42. 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.
43. Rost B. PHD: predicting one-dimensional protein structure by profile based neural networks. Methods Enzymol 1996;266:525–539.
44. Rost B, Casadio R, Fariselli P. Topology prediction for helical transmembrane proteins at 86% accuracy. Protein Sci 1996;5:1704–1718.
45. Schulz GE. beta-Barrel membrane proteins. Curr Opin Struct Biol 2000;10:443–447.
46. Wunderlich Z, Liu J, Kornhaber G, Acton TB, Rost B, Montelione GT. ZebaView: the official protein target list of the northeast structural genomics consortium. Proteins 2003. Forthcoming.
47. Bertone P, Kluger Y, Lan N, Zheng D, Christendat D, Yee A, Edwards AM, Arrowsmith CH., Montelione GT, Gerstein M. SPINE: an integrated tracking database and data mining approach for identifying feasible targets in high-throughput structural proteomics. Nucleic Acids Res 2001;29:2884–2898.
48. Carter P, Liu J, Rost B. PEP: predictions for entire proteomes. Nucleic Acids Res 2003;31:410–413.
49. Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I, Pilbout S, Schneider M. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res 2003;31:365–370.
50. de Souza SJ, Long M, Schoenbach L, Roy SW, Gilbert W. Intron positions correlate with module boundaries in ancient proteins. Proc Natl Acad Sci USA 1996;93:14632–14636.
51. 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.
52. Eisenberg D, Marcotte EM, Xenarios I, Yeates TO. Protein function in the post-genomic era. Nature 2000;405:823–826.
53. Jones S, Marin A, Thornton JM. Protein domain interfaces: characterization and comparison with oligomeric protein interfaces. Protein Eng 2000;13:77–82.

54. 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.
55. Liu Y, Eisenberg D. 3D domain swapping: as domains continue to swap. Protein Sci 2002;11:1285–1299.
56. Ofran Y, Rost B. Analysing six types of protein-protein interfaces. J Mol Biol 2003;325:377–387.
57. Moult J. Structure to function consortium (S2F). Rockville, MD: Center for Advanced Research in Biotechnology, University of Maryland; 2003.
58. Eyrich V, Marti-Renom MA, Przybylski D, Fiser A, Pazos F, Valencia A, Sali A, Rost B. EVA: continuous automatic evaluation of protein structure prediction servers. Bioinformatics 2001;17:1242–1243.
59. Marti-Renom MA. Accuracy of comparative modelling. San Francisco, CA: University of California, San Francisco; 2003, Available online at http://eva.compbio.ucsf.edu/~eva.
60. Koh IYY, Eyrich VA, Marti-Renom MA, Przybylski D, Madhusudhan MS, Narayanan E, Graña O, Valencia A, Sali A, Rost B. EVA: evaluation of protein structure prediction servers. Nucleic Acids Res 2003;31:3311–3315.
61. Yokoyama S, Kuramitsu S. RIKEN Structural Genomics Initiative (RSGI). Yokohama, Japan: Institute of Physical and Chemical Research; 2003.
62. Linial M, Yona G. Methodologies for target selection in structural genomics. Prog Biophys Mol Biol 2000;73:297–320.
63. Brenner SE. A tour of structural genomics. Nature 2001;2:801–809.
64. Liu J, Rost B. Comparing function and structure between entire proteomes. Protein Sci 2001;10:1970–1979.
65. Vitkup D, Melamud E, Moult J, Sander C. Completeness in structural genomics. Nat Struct Biol 2001;8:559–566.
66. Liu J, Rost B. Target space for structural genomics revisited. Bioinformatics 2002;18:922–933.
67. Wright PE, Dyson HJ. Intrinsically unstructured proteins: reassessing the protein structure-function paradigm. J Mol Biol 1999;293:321–331.
68. Dunker AK, Obradovic Z. The protein trinity-linking function and disorder. Nat Biotechnol 2001;19:805–806.
69. 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.
70. Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 1999;292:195–202.
71. Przybylski D, Rost B. Alignments grow, secondary structure prediction improves. Proteins 2002;46:195–205.
72. Emanuelsson O, von Heijne G. Prediction of organellar targeting signals. Biochim Biophys Acta 2001;1541:114–119.
73. Rost B, Sander C. Prediction of protein secondary structure at better than 70% accuracy. J Mol Biol 1993;232:584–599.
74. Rost B, Sander C. Combining evolutionary information and neural networks to predict protein secondary structure. Proteins 1994;19:55–72.
75. Rost B, Liu J. The PredictProtein server. Nucleic Acids Res 2003;31:3300–3304.
76. Rost B, Sander C. Conservation and prediction of solvent accessibility in protein families. Proteins 1994;20:216–226.
77. Eddy SR. Profile hidden Markov models. Bioinformatics 1998;14:755–763.
78. Coulson AF, Moult J. A unifold, mesofold, and superfold model of protein fold use. Proteins 2002;46:61–71.