Completeness in structural genomics

Dennis Vitkup[1,2], Eugene Melamud[3], John Moult[3] and Chris Sander[1,4]

Structural genomics has the goal of obtaining useful, three-dimensional models of all proteins by a combination of experimental structure determination and comparative model building. We evaluate different strategies for optimizing information return on effort. The strategy that maximizes structural coverage requires about seven times fewer structure determinations compared with the strategy in which targets are selected at random. With a choice of reasonable model quality and the goal of 90% coverage, we extrapolate the estimate of the total effort of structural genomics. It would take ~16,000 carefully selected structure determinations to construct useful atomic models for the vast majority of all proteins. In practice, unless there is global coordination of target selection, the total effort will likely increase by a factor of three. The task can be accomplished within a decade provided that selection of targets is highly coordinated and significant funding is available.

The success of genome sequencing projects and advances in protein structure determination have led the structural biology community to propose a comprehensive effort, often referred to as structural genomics[1–6], to map protein structure space. Structural genomics may develop into a major international collaboration similar to the human genome sequencing project. Several pilot projects are currently under way in the United States (*http://www.x12c.nsls.bnl.gov/StrGen.htm*), Europe (*http://userpage.chemie.fu-berlin.de/~psf/index.html*) and Asia.

What are the specific goals of structural genomics? A variety of objectives have been proposed[7] ranging from obtaining representative structures for all protein folds (estimated at 2,000–4,000 structures[8]) to solving the structures of all human proteins (~35,000 protein genes)[9,10]. Here we explore the goal of obtaining a set such that accurate atomic models can be built for almost all functional domains, including those resulting from alternative splicing and post-translational modifications. The goals of structural genomics are qualitatively different from those of genome sequencing projects. Sequencing genomics has a well-defined scope — the experimental determination of the complete (consensus) nucleotide sequence of a particular organism — for example, the approximately three billion base pairs for the human genome. For structural genomics, the overall scope is less well defined because it depends on the ratio of structures determined experimentally and structural models built computationally. Additionally, although the sequences of each additional organism have to be determined experimentally, most structures of related organisms can be determined with little additional experimental effort using homology modeling methods.

This paper explores a number of alternative approaches to yield completeness in structural coverage of protein sequence space and to estimate the total effort required under different scenarios. First, we consider the accuracy of current modeling methods in order to quantify the ratio of experimental and computational effort. Second, we investigate the extent to which the known fraction of protein space is covered by experimental or computational structures. Third, we quantify the relationship between the scope of structural genomics and the quality of structural models obtained. We then calculate the number of experimental structure determinations required to cover a well-defined set of currently known protein families. Several possible scenarios of protein space coverage are explored. Finally, we estimate the number of experimental structure determinations required to provide structural models for the vast majority of proteins from all organisms.

Accuracy of structural modeling

The number of experimental structure determinations required to cover protein space depends critically on the reliability of homology modeling methods. We aim at a careful choice of requirements for minimal sequence similarity using data on modeling accuracy from the Critical Assessment of Techniques for Protein Structure Predictions (CASP)[11]. CASP collects and analyzes *bona fide* structure predictions from a large number of participating research groups, spanning a wide range of relationships between the model and template structures. To quantify modeling quality we use the reported spatial deviations and alignment errors between model and structural template as a function of the sequence identity (Fig. 1*a*,*b*). We drew the conclusion, as had others[12,13], that models based on <30% sequence identity have significant alignment errors, resulting in large errors in main chain positions. Structural models based on >30–35% sequence identity tend to have reasonably low alignment and structural errors. For this and higher sequence identity, it is often possible to correlate differences in function within protein families with structural variations. In this paper, we call models based on at least 30% sequence identity 'reasonably accurate' or, for brevity, 'accurate'.

Structural coverage of currently known proteins

Before estimating how many experimental structures will be needed to build models of all proteins, computing the structural coverage of the currently known ones is instructive. Structural coverage for ~300,000 sequences in the databases SWISS-PROT (SP, release 37) plus TrEMBL (release 11)[14] was calculated using the profile method PSI-BLAST[15] and reported as a function of sequence identity and the fraction of the sequence aligned between the sequence of the modeling target and that of a protein of known structure (Fig. 2). Often, these alignments correspond to distinct structural domains[16–18] and do not cover the full length of the modeling target sequence.

[1]MIT Center for Genome Research, One Kendall Square, Building 300, Cambridge, Massachusetts 02139, USA. [2]Department of Chemistry and Chemical Biology, Harvard University, Cambridge, Massachusetts 02138, USA. [3]Center for Advanced Research in Biotechnology, University of Maryland Biotechnology Institute, 9600 Gudelsky Drive, Rockville, Maryland 20850, USA. [4]Millennium Pharmaceuticals, 640 Memorial Drive, Cambridge, Massachusetts 02139, USA.

Correspondence should be addressed C.S. *email: sander@genome.wi.mit.edu*

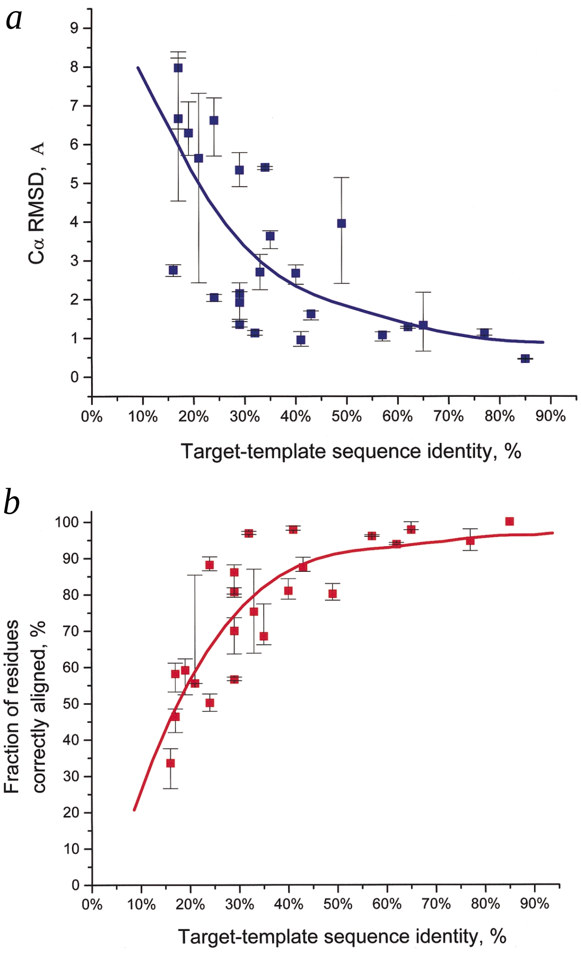


Fig. 1 Accuracy of CASP protein structure models as a function of target-template sequence identity. Data are from all models from CASP2 and CASP3 (Critical Assessment of Techniques for Protein Structure Predictions[11]) for which >80% of the protein residues are modeled. Each data point represents an average over the six best predictions for a single target. The range bars delineate the most and the least accurate models out of each set of six best predictions. ***a***, As sequence identity falls below 30%, errors in Cα coordinates rapidly increase. RMSD is root mean square (r.m.s.) positional deviation. ***b***, The primary cause of this effect is alignment errors between target and template sequences. The alignment error is quantified as the percentage of misaligned residues in 3D. Model data were obtained from the CASP Web site (*http://prediction center.llnl.gov*).

Using these alignments as structural templates, accurate structural models covering the full length of the protein (upper right quadrant, Fig. 2) can be constructed for 19% of the proteins in SP + TrEMBL. For an additional 10% of the proteins, such modeling is possible for part of the sequence (lower right quadrant). In all, some structural information, whether full length or not, is available for 43% (19%+10%+10%+4%) of the proteins in SP + TrEMBL. How many reasonably accurate models can be built for each experimental structure? As the PDB (Protein Data Bank, January 2000) has 3,100 nonredundant structures filtered at 95% sequence identity[19], the modeling ratio of full length structural models to the number of experimental protein structures is currently (0.19 × 300,000) / 3,100, or ~20. That is, for every unique protein in the PDB, on average 20 reasonably accurate full length models may be built from SP + TrEMBL sequences. We expect this ratio to increase as more and more sequences become available. At the same time, the growing database of experimental structures will in general put more and more sequences within reach of model building based on several alternative structural templates, with an attendant potential gain of modeling accuracy[20,21].

Structural coverage of fully sequenced genomes

Reliable extrapolation from the different current sequence data sets to all natural protein sequences is difficult because of uneven representation of types of proteins and species. Therefore, we computed the structural coverage of a representative set of completely sequenced genomes (not including the recently completed human genome) as a basis for more reliable extrapolation.

The coverage differs significantly depending on whether one focuses on the fraction of protein sequences with a link to a known structure (using a more permissive threshold in results from PSI-BLAST similarity searches — that is, optimistic view; Table 1) or the fraction of the total number of amino acid residues that can be accurately modeled (using the less permissive threshold of 30% minimal sequence identity over aligned regions in results from FASTA searches — that is, realistic view; Table 1). In the optimistic view, some structural information is available for 30–35% of sequences of genomes. In contrast, in the realistic view, only ~5–10% of residues from complete genomes can be placed in accurate structural models. Note that by using more sensitive methods and data from more recent structures, it is possible to increase the fraction of genomic sequences with a link to a fold to ~50% and the fraction of all residues with such a link to ~40% (J. Gough and C. Chothia, pers. comm., see also *http://stash.mrc-lmb.cam.ac.uk/superfamily*).

Compared with structural coverage of complete genomes, the SP + TrEMBL sequence database (Fig. 2) is clearly biased in that it contains a larger fraction than found in complete genomes of accurately modelable residues (23% *versus* 5–10%). To estimate the overall effort in structural genomics, we use genome-based numbers. These numbers are similar in spirit, but different in detail, from those of other studies of structural assignment across completely sequenced genomes[22–25].

Scope as a function of desired model quality

We now take a detailed look at the way in which the number of experimental structure determinations required to cover protein space depends on the reliability of homology modeling methods. Anticipating complete organization of protein sequences into domain families across all species, we use the current Pfam collection of protein alignments[26] and perform data simulations with models built using template structures at differing levels of model quality. More precisely, in each simulation run, we set the minimal model quality in terms of a maximal modeling distance or minimal model-template sequence identity.

As an illustration of coverage at different levels of minimal model quality, consider the example of the ras-like protein family (G-domain) in yeast. This is best visualized in a two-dimensional projection (Fig. 3) of a higher-dimensional protein sequence space in which distances between points (each point is a family member) represent the modeling distance between proteins, quantified as percent residue differences between amino acid sequences. At a given minimal modeling quality, a certain number of structural templates (centers of circles) are needed to provide modeling coverage to sequence neighbors (contained in a circle). A decrease in maximal modeling distance (smaller circles) leads to higher accuracy structural models; however, more experimental structure determinations (more circles) must be made to cover family members.

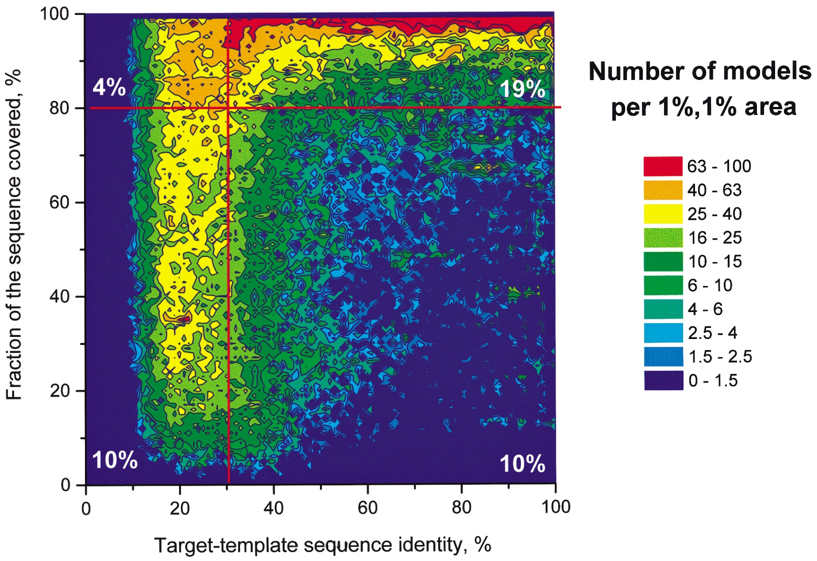


Fig. 2 Current structural coverage of proteins in SP + TrEMBL. Color contours show the density of models that can be currently built as a function of the fraction of the sequence included in a model (vertical axis), and the sequence identity between the modeled protein and the closest known experimental structure (horizontal axis). The distribution was constructed for SP + TrEMBL (release 7 and 11, respectively; SP = SWISS-PROT) using sequence profile searches with PSI-BLAST (see Methods) against proteins with known structures in the PDB (Protein Data Bank). Models to the right of the vertical red line are based on >30% sequence identity over the length of the aligned subsequences and are of relatively high quality; these are called 'accurate' or 'reasonably accurate' models in the text and form the basis of the estimates of modeling density. Models above the horizontal red line cover 80% or more of the sequence. The most useful models (upper right, 19% of the total) are at high levels of sequence identity and cover most of the length of the protein.

To simulate structural coverage as a function of minimal model quality, we use release 4.4 of PfamA, which contains 2,000 domain families (including, by our definition, 1,626 nonmembrane families) constructed from ~260,000 domain sequences. Most Pfam families represent structural domains[27] and are assembled using sequence profiles in the form of hidden Markov models (HMMs)[28]. Manual curators aim at ensuring high quality alignments and accurate definition of domain boundaries. Of the proteins in SP + TrEMBL, ~63% have at least one domain in Pfam[26].

The number of structure determinations required was estimated using a greedy coverage algorithm[19]. The greedy algorithm first selects the structural target (template structure) that would generate the maximum number of models for sequences within the maximal modeling distance, then the target that returns the maximum number of models for the remaining sequences in the family is selected, and so on, until there are no sequences left that cannot be modeled. The algorithm is run repeatedly on a given collection of family alignments for different values of maximal modeling distance. The results are as follows.

Assuming that 30% or better sequence identity is required for accurate modeling, ~13,000 experimental structures are required to cover models for all nonmembrane domains (in 1,626 families) in Pfam. Many of these have already been done (we estimate ~35% of all residues using the criteria of Table 1, second column), but the point here is to derive numbers for modeling density in a known family collection and use these for extrapolation to less well-known regions of protein sequence space. Inclusion of membrane associated families in Pfam increases the number of structure determinations required for accurate modeling of all 260,000 sequences in Pfam to 17,000.

How does the number of structure determinations required to cover the Pfam collection depend on desired model quality? First of all, there is a clear trade-off between model quality and experimental effort (Fig. 4): as minimal modeling quality (horizontal axis) increases, more template structures (vertical axis) are required. Above 30% sequence identity, the number of experimental structure determinations increases approximately linearly with the minimal sequence identity between model and template. The slope change at ~20% sequence identity represents the current limit of sensitivity for reliably grouping protein domains into families. For modeling distances in the twilight zone of sequence identity (~10–20%), modeling density is highest, so that a single structure determination of any protein from the family is often sufficient to cover all family members but with a high penalty in model quality. The shape of the curve (Fig. 4) is such that a minimal modeling distance at 30% sequence identity captures most of the savings in effort (decrease in the number of structure determinations), confirming our choice of minimal model quality for the purposes of estimating the scope of structural genomics.

Practical considerations in covering protein space

A number of factors may modify our simple estimates. Here, we consider (i) substantial savings from a slight relaxation of completeness requirements; (ii) realistic success rates of structure determination; (iii) special types of protein sequences; and (iv) variation in target selection strategy.

(i) Quasi-completeness: in computing the minimal number of structure determinations for complete model coverage of all pro-

Table 1 Two views of current structural coverage of complete genomes[1]

Organism	Optimistic view[2]	Realistic view[3]
M. genitalium	36%	10.5%
M. pnemonie	32%	7.7%
H. pylori	29%	7.8%
M. jannashii	32%	6.0%
H. influenza	39%	10.3%
B. subtilis	31%	10.1%
E. coli	31%	9.9%
S. cerevisiae	36%	6.8%
C. elegans	35%	6.5%

[1] Full genus names: *Mycoplasma genitalium, Mycoplasma pnemonie, Helicobacter pylori, Methanoccoccus jannashii, Haemophilus influenza, Bacillus subtilis, Escherichia coli, Saccharomyces cerevisiae, Caenhorabditis elegans*

[2] Percentage of genomic sequences with a link to a known structure.

[3] Percentage of genome residues in accurate homology models (based on 30% or higher sequence identity). The fraction of genome sequences for which a partial structural model, including those at low accuracy, can be constructed. The coverage of genome sequences and residues was obtained using sequence searches with PSI-BLAST and FASTA (see Methods).

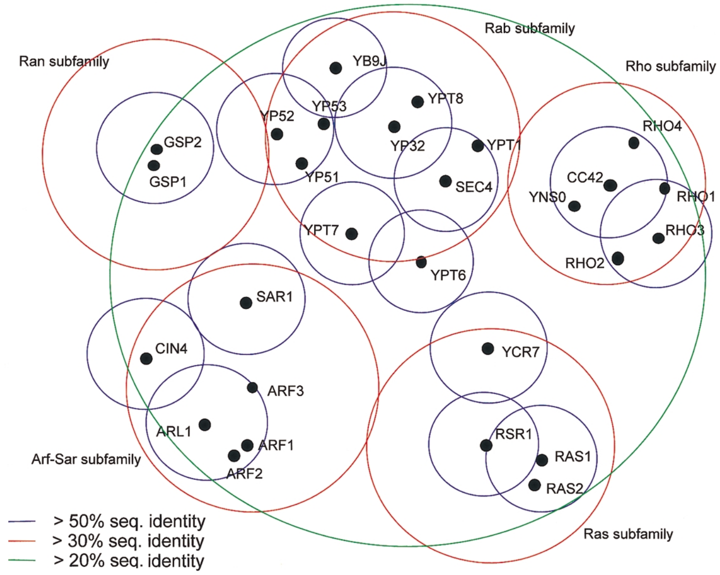


Fig. 3 Structural coverage of a protein family, illustrated using the Ras family in yeast as an example. Members of the family (labeled dots) are projected onto a plane (see Methods). The distance between points is approximately proportional to the modeling distance (modeling distance = 100% – sequence identity). The circles represents structural coverage based on different levels of model-template sequence identity: 20% (green), 30% (red) or 50% (blue). Increasing the number of structure determinations results in more accurate models; for example, structure determination of a single protein (YPT6 in the center of the green circle) allows modeling of all family members based on >20% sequence identity to the structural template. Solving the structure of five proteins (red circles) allows modeling based on >30% sequence identity; solving the structure of 15 proteins (blue circles), modeling based on >50% identity. For clarity, experimental structural information already available for the Ras family is not taken into account.

tein domains in Pfam, we relaxed the completeness requirement from 100% to 90% and found a substantial decrease in effort from 13,000 to ~3,300 structure determinations (4,000 structures if membrane proteins are included). This large drop is a consequence of the uneven distribution of proteins in sequence space. Coverage of relatively dense regions of protein space — for example, Ras, Rab, Rho, Arf-Sar (Fig. 3) — requires few structure determinations to obtain many models. In contrast, coverage of sparse regions of protein space — for example, Ran subfamily (Fig. 3) — returns much fewer models per structure determination. In light of these data, it is impractical to aim for achieving 100% structural coverage of protein space. Instead, a reasonable objective for structural genomics is to focus on the denser parts of protein space (except as dictated by particular biological interest), aiming in general to obtain accurate models for 90% of all sequences.

(ii) Realistic success rates: in practice, structure determination for certain proteins, whether by crystallography or by nuclear magnetic resonance spectroscopy, can be difficult or impossible because of a variety of problems — for example, in cloning, expression, purification, concentration, labeling and/or crystal growth. In an attempt to estimate how experimental difficulties affect structural coverage, we simulated coverage of Pfam families by assuming different success rates of structure determination. A less than perfect success rate was represented in a modified version of the greedy algorithm as follows: whenever a possible target is selected for structure determination, a random number generator is used to simulate if a structure determination would be successful. For a success, the algorithm proceeded as usual. Otherwise, structure determination was assumed to be impossible and was never attempted again for that protein.

Simulation of the impact of different success rates on the structural coverage of Pfam families (Fig. 5*a*) shows that coverage is not seriously affected by a decrease of success rate down to values of ~0.2 (one success in five attempts). Even a success rate as low as 0.1 does not decrease coverage by >10 percentage points. The reason is that large families usually provide several alternative targets, often from different organisms, that may be equally suitable as templates for structural modeling. As genome sequencing continues, families will grow larger, and structural genomics will be able to accommodate even smaller success rates without compromising overall coverage. Early returns from pilot structural genomics projects suggest success rates considerably better than 0.1. For example, in one project[6] a first set of 62 target proteins from a single organism has already yielded 15 structures (success rate of 0.2).

(iii) Nonstandard sequence regions: in addition to globular and transmembrane domains, a currently undetermined frac-

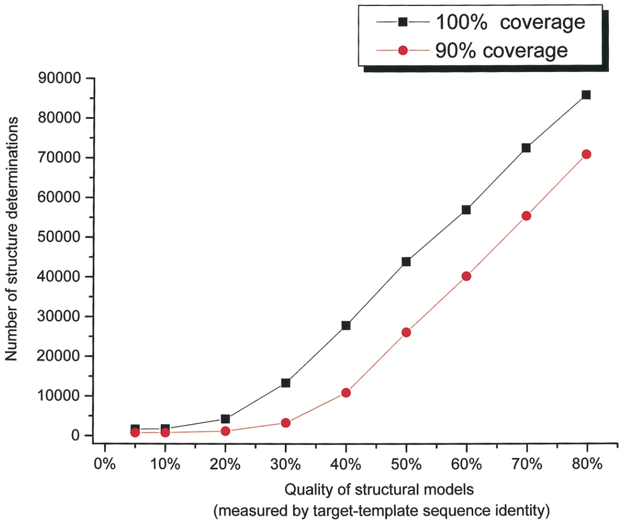
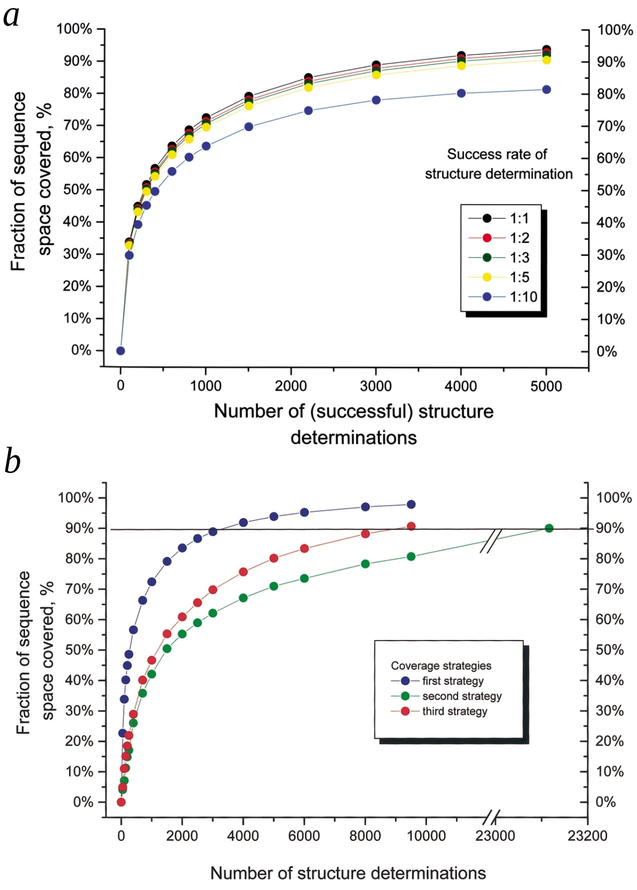


Fig. 4 Scope of structural coverage as a function of model quality. The number of experimental structure determinations required to model all sequences in nonmembrane associated Pfam families as a function of sequence identity between the modeling target and the experimentally determined template structure. A greedy coverage algorithm was used to approximate the minimal number of structure determinations required to cover 100% (red) or 90% (black) of nonmembrane protein sequences in Pfam. Above 30% sequence identity, the number of structure determinations is approximately proportional to model quality (measured by sequence identity). Note that for sequence identity in the 25–35% range there is a significant reduction (by a factor of 3–4) in the number of structures required to cover 90% rather than 100% of all sequences. Pfam release 4.4 containing 2,000 families was used for all calculations.

Fig. 5 Two factors affecting the scale of structural genomics. ***a***, Protein space coverage at different success rates of structure determination. The coverage algorithm assumes that any desired structure can be obtained experimentally. In practice, many choices will restrict that choice. We estimate the reduction in the coverage of protein sequence space by models based on >30% sequence identity as a function of the success rate in getting the desired experimental structures. All nonmembrane proteins from Pfam4.4 were used in the calculation. Success rates between 1:1 (100% successful) and 1:10 are considered. As many families provide a number of alternative structural targets, it is possible to achieve nearly the same final coverage of protein space, within 10 percentage points, down to a success rate of about 1:10. ***b***, Different scenarios of protein space coverage. The coverage algorithm assumes that the experimental community will focus on proteins that maximize the number of models that can be built. In practice, such a universal coordination in target selection is unlikely. Here we compare three scenarios and how they affect the number of models that can be built for a given number of structure determinations. In the first (blue), the greedy strategy that maximizes the number of built models is used. In the second (green), proteins are selected at random for structure determination; this strategy requires seven times more structure determinations to achieve 90% coverage of protein space compared to the first. In the third (red), intermediate strategy, it is assumed that proteins are chosen for structural study randomly, but only if they are <30% in sequence identity to a known structure The third strategy requires 2.5× more structure determinations for 90% coverage compared with the first. All nonmembrane associated proteins in Pfam4.4 were used in the calculation.



tion of proteins are not suitable for mainstream structural genomics. These are filamentous proteins, such as simple coiled coils, which can easily be modeled computationally, as well as regions having unusual amino acid composition, sometimes called low-complexity regions[29], for which structure solution presents an unsolved problem. Finally, an unknown fraction of domains are likely to attain a well-defined structure only as part of a large complex with attendant experimental difficulties. We have not included these effects here.

(iv)Alternative strategies: A major practical consideration is the choice of targets for structure determination, with different criteria applied by different laboratories — for example, medical or functional significance, or species representation. We simulated how different strategies will affect coverage of protein space by varying the target selection process (Fig. 5*b*). A maximal modeling distance of 30% was used in all cases. The first strategy focused exclusively on the maximal return of accurate structural models (simulated using the simple greedy algorithm, see above). The payoff is measured by the fraction of the protein space covered (the blue curve in Fig. 5*b*). As noted earlier, this strategy results in ~3,300 structure determinations in order to cover 90% of protein space. The second possible strategy is completely random selection of the targets (red curve in Fig. 5*b*). Assuming that, to a first approximation, functionally important proteins are randomly distributed in protein sequence space, this strategy approximates space coverage with selection of structural targets based exclusively on functional importance. To achieve 90% coverage, about seven times as many structure determinations are needed compared with the first strategy. In practice it is likely that an intermediate strategy will be adopted by the structural genomics community. For example, in the third strategy considered, targets are selected at random but only if they have <30% sequence identity to an already determined structure (green curve in Fig. 5*b*). The third strategy requires about two to three times as many structure determination to achieve 90% coverage compared with the optimal first strategy.

Table 2 Coverage of complete genomes/chromosomes by current Pfam domain families[1]

Organism name[2]	Fraction of sequences with at least one Pfam domain	Fraction of residues covered by Pfam domains	Number of unique Pfam domains found
E. coli (4,257)	52%	38%	753
M. jannashii (1,715)	50%	36%	499
S. cerevisiae (6,406)	46%	23%	752
C. elegans (16,332)	48%	22%	819
A. thaliana, chromosome 2 (4,038)	49%	23%	510
D. melanogaster (13,710)	53%	22%	939
Human, chromosome 22 (482)	60%	30%	174
SWISS-PROT + TrEMBL (~300,000)	63%	45%	2,000

[1] Coverage was calculated using HMM profiles searches with HMMer[37]. Typically, about half of the proteins in a genome contain at least one Pfam domain, representing approximately one quarter of the residues.
[2] Number of proteins in a genome or chromosome is indicated in parenthesis.

Towards comprehensive clustering of protein space

Can the structural genomics approach be efficiently applied to cover all protein space? The answer depends on the feasibility of clustering the vast majority of protein space into a set of domain families of nontrivial size, such that models can be built of many family members based on one or a few structural representatives. This depends on the diversity and number of genomes sequenced. With many genome sequencing projects underway, time favors the aggregation of proteins into families in our databases. For example, domain families of size one, sometimes called singletons, have been declining as a fraction of the total number of families. Even if singletons as a fraction of families within genomes remained at 10%, aggregation of homologs between genomes will lead to their rapid disappearance[30].

Eukaryotic genomes in particular appear to have a significant degree of sequence similarity within and between organisms, indicative of paralogy or homology. For example, a study of a continuous stretch of genomic sequence from the Adh region in *Drosophila melanogaster*[31] reported that ~72% of genes have homology to sequences in other eukaryotic organisms. Such a high level of sequence homology is impressive if one considers that only a small fraction of eukaryotic genomes is currently available. In addition, recent analysis of the complete genome sequence of *D. melanogaster* showed that 60% of human 'disease genes' have full-length homologs in the fly genome[32]. A high level of sequence similarity was also found in the genome analysis of *Caenhorabditis elegans*[33], *Arabidopsis thaliana*[34,35] and human chromosome 22 (ref. 36) genomes.

Once representative organisms in major branches of the tree of life have been sequenced, we expect that additional sequencing will turn up increasingly less sequence diversity, but the likelihood that new coding sequences will join existing families will steadily increase. Therefore, most of protein space can be clustered into large families of homologous/paralogous protein domains. Possible exceptions are proteins or segments with unusual composition, sometimes called low complexity regions, as well as some filamentous regions. We could, therefore, reasonably and cautiously extrapolate the results of data simulations using the Pfam domain family database to estimate the number of structure determinations required to structurally cover most sequences from all organisms.

Structure determinations to cover most protein space

We estimate the total number of experimental structure determinations required to cover all of protein space in two steps: (i) we estimate the total number of domain families by assessing which fraction of all coding regions in key genomes can be assigned to known Pfam domain families (Table 2), and (ii) we then assume, to a first approximation, that the modeling density within the Pfam domain database applies to all protein space, including currently unknown families. This estimation based on Pfam is a good starting point, but may be biased by the tendency of known Pfam families to be large and well characterized and may likely underrepresent both filamentous proteins and proteins with amino acid composition atypical of globular proteins. The HMMer package[37] was used to calculate the fraction of genome residues that can be assigned to known Pfam families (Table 2). The results of our calculations are consistent with earlier studies by Bateman *et al.*[26] and do not depend on the particular value of maximal modeling distance chosen. For a wide variety of genomes available to us (and for all sequences in SP + TrEMBL), about one-half of the sequences and one-quarter of the residues can be assigned to known Pfam families. Given 2,000 families in Pfam, this puts the estimate for the total number of protein domain families at 2,000 / 0.25, or 8,000. This number is compatible with other recent estimates[38].

Given this estimate of the total number of families, how many structures will it take to cover these? To cover 90% of all protein domain sequences, ~4,000 structure determinations are needed in 2,000 Pfam families using an optimal strategy. Extrapolating 2,000–8,000 families and assuming the Pfam modeling density can be applied to all of protein space, the total number of structure determinations required to produce models for 90% of protein space is 4,000 / 0.25 = 16,000, using the optimal strategy for target selection (see above). Nonredundant domains structures already solved (~10%, Table 1) are included in this total.

In practice, departures from a strategy that maximizes the number of models per experimental structure are likely. For example, simulation of (uncoordinated) target selection, followed by deselection of potential targets already covered by an accurate model, leads to about three times that number — ~50,000 structure determinations (see (iv) in the section on practical considerations and Fig. 5*b*).

Conclusions

The principal goal of structural genomics is to construct a complete and accurate map of protein structure space. The map can be constructed by experimental structures of representatives from protein families in combination with computational homology modeling. Here, we have explored comprehensive structural coverage based on a curated collection of protein families[26]. Qualitatively similar results were obtained using other databases of domain families[39–44]. We draw several conclusions based on simulations in a limited data set of sequence families and structures, with cautious extrapolation to a much larger fraction of all natural proteins:

(i) The fraction of protein space for which reasonably accurate structural modeling is currently possible is relatively small. Although some structural information is available for domains in about one-third of sequences from complete genomes, the fraction of residues in a given genome that can be included in an accurate model is generally below 10%.

(ii) The number of experimental structure determinations required to cover protein space increases monotonically with the desired quality of homology models, where model quality is quantified by the minimal sequence identity used in modeling. The increase is approximately linear above 40% sequence identity.

(iii) The number of structure determinations required to cover 90% of protein domains is about four times smaller than the number required to cover 100%. Coverage of the remaining 10% would come at a disproportionately high cost. Assuming that at least 30% sequence identity to a structural template is required for structural modeling, on average eight structure determinations would be needed to cover 100% of a protein family *versus* two structure determinations to cover 90% of a protein family.

(iv) Large protein families provide a number of alternative structural targets useful as structural templates. Consequently, broad structural coverage can be achieved economically with a relatively small success rate of experimental structure determination (as low as one in five).

(v) The number of structure determinations required for complete coverage of protein space crucially depends on the strategy used for the selection of structural targets. For a goal of 90% coverage, a strategy of completely random selection of targets requires almost seven times more structure determinations than a strategy that optimizes the average number of computational models per experimental structure.

(vi) Improvements in computational modeling methods would lead to a significant reduction in experimental effort. For example, a 10% decrease in the threshold needed for accurate modeling, from 30% to 20% sequence identity, would reduce the number of experimental structures required by more than a factor of two.

How soon can the goals of structural genomics be achieved? We estimate at least 16,000 experimental structure determinations would be required to accurately model almost all proteins when using an optimal strategy for target selection. However, unless there is tight coordination of target selection, as many as 50,000 structure determinations may be necessary. A more accurate estimate will be possible once the genome sequences of more eukaryotes are complete and reliable protein sequences have been deduced.

The current rate of experimental structure determination is ~50 structures per week[45] (or 25,000 per decade). However, only about one in five solved structures is nonredundant in that it represents a new protein family, as defined by a family radius of 25–30% sequence identity[46,47]. Consequently, only ~10 nonredundant structures are solved per week (or 5,000 per decade). Over the next few years, the rate of traditional (low-throughput) structure determination is likely to further increase (since 1990 the rate has increased 10-fold). In addition, emerging structural genomics projects (both in industry and academia) are aiming for a total high-throughput production level of thousands of nonredundant structures per year. Based on these projections, it is possible that a combination of low-throughput and high-throughput approaches will yield a near-complete map of protein structure space in about a decade.

Methods

Family plane projection. The projection of the yeast Ras family onto a plane (Fig. 3) was obtained using the package SOM_PAK[48] with slight manual adjustments. The sequence YPT6 was chosen as the center of projection. Note that the projection of a multidimensional space (such as protein sequence space) onto a plane cannot accurately represent all distances.

Structural coverage of SP+TrEMBL. Structural coverage distribution of proteins in SP + TrEMBL (Fig. 2) was calculated by PSI-BLAST sequence profile searches against the PDB[49]. For each nonredundant protein in PDB95 (PDB proteins filtered at 95% sequence identity), three rounds of iterative PSI-BLAST searches were conducted against SP + TrEMBL. Sequences with an expectation score (E-value) <0.001 were collected in each iteration. These sequences were used as potential structural templates for proteins in SP + TrEMBL.

The rate of false positives was estimated using the SCOP databases of structural domains release 1.37 (ref. 16). For all nonredundant sequences representing SCOP structural domains (SCOP filtered at 95% sequence identity), three iterations of PSI-BLAST were run against the SP + TrEMBL. All SP + TrEMBL sequences hits with E-scores above 0.001 were collected for each SCOP domain. Two or more hits to the same region (defined by a 90% overlap) of a SP + TrEMBL sequence from SCOP domains with different folds are clearly spurious (false positives)[24]. The percentage of such hits relative to the total number of hits to SP + TrEMBL sequences was ~3%.

Structural coverage of complete genomes/chromosomes. Structural coverage of complete genomes was estimated using PSI-BLAST and FASTA searches. The fraction of genome sequences with a link to a known fold was calculated by running three rounds of PSI-BLAST using the PDB95 dataset[19]. Only sequences with an E-value <0.001 and >50 amino acids in length were considered as a link. The fraction of residues in genomes that could be included in homology models was estimated by FASTA searches using PDB95. Only FASTA hits with an E-value <0.001 and sequence identity >30% to a PDB protein were considered. The BLAST and FASTA searches were performed using a parallel Beowulf cluster (*www.beowulf.org*).

Detection of membrane families in Pfam. The program TopPred[50] was used to detect transmembrane families in Pfam. Families in which >30% of the sequences had a confident prediction of at least one transmembrane domain were considered as transmembrane families. This criterion is similar to the one used recently by Elofsson *et al.*[51].

Coverage of genomes by current Pfam families. Coverage of different genomes by the Pfam families was calculated using profile searches with the HMMer package[37]. The PVM (parallel virtual machine) version of HMMer was run on a Beowulf cluster. The family specific gathering cutoffs (GA) used in the compilation of Pfam families were applied.

Greedy coverage algorithm. The following greedy algorithm was used to estimate the number of experimental structure determinations required to cover each family: (i) For a given maximal modeling distance, the number of models that can be built based on each (remaining) protein in the family (model yield) is calculated. (ii) The protein with the highest model yield is selected. This protein and the proteins structurally covered by it are removed from further calculations. (iii) The number of structure determinations required to cover this family is increased by one. Step (i) is repeated until there are no proteins left in the family.

Acknowledgments

We thank L. Holm, D. Marks and J. Norvell for discussions and C. Venclosas for providing CASP template/target sequence identity data. This work was supported in part by research grants from the US National Institute of Health (NIGMS) and the Department of Energy to J.M. and C.S.

Received 6 March, 2001; accepted 23 March, 2001

1. Kim, S.H. Shining a light on structural genomics. *Nature Struct. Biol.* **5**, 643–645 (1998).
2. Terwilliger, T.C. *et al.* Class-directed structure determination: foundation for a protein structure initiative. *Protein Sci.* **7**, 1851–1856 (1998).
3. Sali, A. 100,000 protein structures for the biologist. *Nature Struct. Biol.* **5**, 1929–1932 (1998).
4. Montelione G.T. & Anderson, S. Structural genomics: keystone for a human proteome. *Nature Struct. Biol.* **6**, 11–12 (1999).
5. Burley, S.K. *et al*. Structural genomics: beyond the human genome project. *Nature Genet.* **23**, 151–157 (1999).
6. Eisenstein, E. *et al.* Biological function made crystal clear – annotation of hypothetical proteins *via* structural genomics. *Curr. Opin. Biol.* **11**, 25–30 (2000).
7. NIGMS Structural Genomics workshop. *http://www.nigms.nih.gov/news/meetings/structural_genomics_targets.html* (NIH campus, USA; 1999).
8. Govindarajan, S., Recabarren, R. & Goldstein, R.A. Estimating the total number of protein folds. *Proteins* **35**, 408–414 (1999).
9. Venter, J.C. *et al.* The sequence of the human genome. *Science* **291**, 1304–1351 (2001).
10. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
11. Moult, J., Hubbard, T., Fidelis, K. & Pedersen, J.T. Critical assessment of methods of protein structure prediction (CASP): Round3. *Proteins* **S3**, 2–6 (1999).
12. Martin, A.C., MacArthur, M.W. & Thornton, J.M. Assessment of comparative modeling in CASP2. *Proteins Suppl.* **1**, 14–28 (1997).
13. Sanchez, R. & Sali, A. Advances in comparative modeling. *Curr. Opin. Struct. Biol.* **7**, 206–214 (1997).
14. Bairoch, A. & Apweiler, R. The SWISS-PROT protein sequence data bank and its new supplement. TrEMBL. *Nucleic Acids Res.* **24**, 17–21 (1996).
15. Altschul, S.F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acid Res.* **25**, 3389–3402 (1997).
16. Murzin, A.G., Brenner, S.E., Hubbard, T. & Chothia, C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247**, 536–540 (1995).
17. Holm, L. & Sander, C. Dali/FSSP classification of three-dimensional protein folds. *Nucleic Acids Res.* **25**, 231–234 (1997).
18. Orengo, C.A. *et al.* CATH – a hierarchic classification of protein domain structures. *Structure* **5**, 1093–1108 (1997).
19. Hobohm, U., Sander, C., Scharf, M. & Schneider, R. Selection of representative protein datasets. *Protein Sci.* **1**, 409–417 (1992).
20. Sanchez, R. & Sali, A. Large-scale protein structure modeling of *Saccharomyces cerevisiae* genome. *Proc. Natl. Acad. Sci. USA* **95**, 13597–13602 (1998).
21. Guex, N., Diemand, A. & Peitsch, M.C. Protein modeling for all. *Trends Biochem. Sci.* **24**, 364–367 (1999).
22. Teichmann, S.A., Chothia, C. & Gerstein, M. Advances in structural genomics. *Curr. Opin. Struct. Biol.* **9**, 390–399 (1999).
23. Sanchez, R. & Sali, A. ModBase: A database of comparative protein structural models. *Bioinformatics* **15**, 1060–1061 (1999).
24. Wolf, Y.I., Brenner, S.E., Bash, P.A. & Koonin, E.V. Distribution of protein folds in the three superkingdoms of life. *Genome Res.* **9**, 17–26 (1999).
25. Gerstein, M. Patterns of protein-fold usage in eight microbial genomes: a comprehensive structural census. *Proteins* **33**, 518–534 (1998).
26. Bateman, A. *et al.* The Pfam protein families database. *Nucleic Acids Res.* **27**, 263–266 (2000).
27. Holm, L. & Sander, C. Dictionary of recurrent domains in protein structures. *Proteins* **33**, 88–96 (1998).
28. Eddy, S.R. Hidden Markov models. *Curr. Opin. Struct. Biol.* **6**, 361–365 (1996).
29. Wootton, J.C. Non-globular domains in protein sequences: automated segmentation using complexity measures. *Comput. Chem.* **18**, 269–274 (1994).
30. Fischer, D. & Eisenberg, D. Finding families for genomics ORFans. *Bioinformatics* **15**, 759–762 (1999).
31. Ashburner, M. *et al*. An exploration of the sequence of a 2.9-megabase region of the genome of *Drosophila melanogaster* - The Adh region. *Genetics* **15**, 179–219 (1999).
32. Rubin, M.G *.et al*. Comparative genomics of the eukaryotes. *Science* **287**, 2204–2215 (2000).
33. Sonnhammer, E.L.L. & Durbin, R. Analysis of protein domain families in *Caenorhabditis elegans*. *Genomics* **46**, 200–216 (1997).
34. Lin, X. *et al.* Sequence and analysis of chromosome 2 of the plant *Arabidopsis thaliana*. *Nature* **402**, 761–768 (1999).
35. Mayer, K. *et al*. Sequence and analysis of chromosome 4 of the plant *Arabidopsis thaliana*. *Nature* **402**, 769–777 (1999).
36. Dunham, I. *et al*. The DNA sequence of human chromosome 22. *Nature* **402**, 489–495 (1999).
37. Eddy, S., Mitchison, G. & Durbin, R. Maximum discrimination hidden Markov models of sequence consensus. *J. Comput. Biol.* **2**, 9–23 (1995).
38. Wolf, Y.I., Grishin, N.V. & Koonin, E.V. Estimating the number of protein folds and families from complete genome data. *J. Mol. Biol.* **299**, 897–905 (2000).
39. Krause, A., Nicodeme, P., Bornber-Bauer, E., Rehmsmeier, M. & Vingron, M. WWW access to the SYSTERS protein sequence cluster set. *Bioinformatics* **15**, 262–263 (1999).
40. Heger, A. & Holm, L. Towards a covering set of protein family profiles. *Prog. Biophys. Mol. Biol.* **73**, 321–337 (2000).
41. Yona, G., Linial, N., Tishby, N. & Linial, M. A map of the protein space – an automatic hierarchical classification of all protein sequences. *ISMB* **6**, 212–221 (1998).
42. Corpet, F., Gouzy, J. & Kahn, D. Recent improvements of the ProDom database of protein domain families. *Nucleic Acids Res.* **27**, 263–267 (1999).
43. Tatusov, R.L., Koonin, E.V. & Lipman, D.J. A genomic perspective on protein families. *Science* **278**, 631–637 (1997).
44. Wu, C.H., Shivakumar, S. & Huang, H. ProClass protein family database. *Nucleic Acids Res.* **27**, 272–274 (1999).
45. Bourne, P.E. Editorial in bioinformatics. *Bioinformatics* **15**, 715–716 (1999).
46. Holm, L. & Sander, C. Protein folds and families: sequence and structure alignments. *Nucleic Acids Res.* **27**, 244–247 (1999).
47. Brenner, S.E. & Levitt, M. Expectations from structural genomics. *Protein Sci.* **9**, 197–200 (2000).
48. Kohonen, T., Hynninen, J., Kangas, J. & Laaksonen, J. SOM_PAK: The self-organizing map program package. (Helsinki University of Technology, Helsinki; 1996).
49. Bernstein, F.C. *et al.* The Protein Data Bank: a computer based archival file for macromolecular structures. *J. Mol. Biol.* **122**, 535–542 (1977).
50. Czero, M., Wallin, E., Simon, I., von Heijne, G. & Elofsson, A. Prediction of transmembrane α-helices in prokaryotic membrane proteins: the dense alignment surface method. *Protein Eng.* **17**, 673–676 (1997).
51. Elofsson, A. & Sonnhammer, E.L.L. A comparison of sequence and structure protein domain families as a basis for structural genomics. *Bioinformatics* **15**, 480–500 (1999).