Title

Implications of structural genomics target selection strategies: Pfam5000, whole genome, and random approaches

Permalink

https://escholarship.org/uc/item/7520r807

Authors

Chandonia, John-Marc
Brenner, Steven E.

Publication Date

2004-07-14

Peer reviewed

Manuscript Prot-00199-2004-R1

Implications of structural genomics target selection strategies: Pfam5000, whole genome, and random approaches.

John-Marc Chandonia[1] and Steven E. Brenner[1,2]

Address for correspondence:
Steven E. Brenner
Department of Plant and Microbial Biology
461A Koshland Hall
University of California
Berkeley, CA 94720-3102
email: brenner@compbio.berkeley.edu

Affiliations:

1 - Berkeley Structural Genomics Center, Physical Biosciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA

2 - Department of Plant and Microbial Biology, University of California, Berkeley, CA 94720, USA

ABSTRACT:

The structural genomics project is an international effort to determine the three-dimensional shapes of all important biological macromolecules, with a primary focus on proteins. Target proteins should be selected according to a strategy which is medically and biologically relevant, of good value, and tractable. As an option to consider, we present the "Pfam5000" strategy, which involves selecting the 5000 most important families from the Pfam database as sources for targets. We compare the Pfam5000 strategy to several other proposed strategies that would require similar numbers of targets. These include including complete solution of several small to moderately sized bacterial proteomes, partial coverage of the human proteome, and random selection of approximately 5000 targets from sequenced genomes. We measure the impact that successful implementation of these strategies would have upon structural interpretation of the proteins in Swiss-Prot, TrEMBL, and 131 complete proteomes (including 10 of eukaryotes) from the Proteome Analysis database at EBI. Solving the structures of proteins from the 5000 largest Pfam families would allow accurate fold assignment for approximately 68% of all prokaryotic proteins (covering 59% of residues) and 61% of eukaryotic proteins (40% of residues). More fine-grained coverage which would allow accurate modeling of these proteins would require an order of magnitude more targets. The Pfam5000 strategy may be modified in several ways, for example to focus on larger families, bacterial sequences, or eukaryotic sequences; as long as secondary consideration is given to large families within Pfam, coverage results vary only slightly. In contrast, focusing structural genomics on a single tractable genome would have only a limited impact in structural knowledge of other proteomes: a significant fraction (about 30-40% of the proteins, and 40-60% of the residues) of each proteome is classified in small families, which may have little overlap with other species of interest. Random selection of targets from one or more genomes is similar to the Pfam5000 strategy in that proteins from larger families are more likely to be chosen, but substantial effort would be spent on small families.

Background

Structural genomics aims at the discovery, analysis, and dissemination of three-dimensional structures of protein, RNA, and other biological macromolecules representing the entire range of structural diversity found in nature (http://www.nigms.nih.gov/news/meetings/airlie.html#agree)[1-5]. Once a single structure in a protein family is solved, the basic fold of the other members of the family may be predicted, even if the similarity of the other sequences is too low to allow accurate modeling[6-8]. Often, the protein structure allows elucidation of molecular function, for example through inference of homology that was too distant to detect from sequence[9-11].

In the United States, the National Institutes of Health are supporting structural genomics projects at 9 pilot centers through the Protein Structure Initiative (PSI). In the first (pilot) phase of PSI, each center independently developed a list of targets to study; in the second (production) phase, beginning in 2005, the majority of targets for all centers are expected to be chosen using a more centralized strategy (http://grants2.nih.gov/grants/guide/rfa-files/RFA-GM-05-001.html)[12].

The target selection strategy for the second phase of PSI must meet several competing goals. First and foremost, it must represent sound biological research that will ultimately have benefits for human health. At the same time, it must present a sufficiently clear succinct motivation to be compelling to participants, to other scientists, and to the public. The work in structural genomics must complement and enrich biological studies beyond structural genomics, while not inhibiting other research in structural biology[13]. Finally, at the risk of stating the obvious, the project must be tractable and provide good value for the considerable resources expended.

Several approaches have been suggested to try to address these goals. One approach is to pursue structures of all proteins encoded in a complete pathogenic genome[14,15]. Completion of a complete structural repertoire will have intrinsic biological value; at the most fundamental level, we will learn for the first time the complete structural repertoire of an organism's proteome. In addition to the breakthrough this offers for basic science, the better understanding of the pathogen—and of how to inhibit its proteins—will have clear medical importance. The project will be clearly understandable to a large audience. It will likely be tractable in scope, though it will be challenging if structures must come from the specific pathogen's genome, to provide high enough resolution structures for drug design. Additional drawbacks of this approach are that it directs the entire thrust of structural genomics in a narrow direction, and that it may lead to pursuit of specialized proteins in the pathogen—many with little medical importance—at the expense of others with much greater broad biological significance. For the most medically-relevant proteins, it is unclear how this effort would be differentiated from structural biology.

A related approach is to solve all the human protein structures. This will have obvious biological and medical value, and is immensely compelling. Unfortunately, completion of the human proteome structure is unlikely to be tractable in the next phase of structural genomics, and it is unclear how to describe a reasonable endpoint short of completion.

A radically different method of target selection that has been suggested is to develop a mapping of protein families, and to choose the sets of families that will provide homology models structures for the largest number of sequences at some level of reliability [5,16-18]. The biological importance of such an approach is unquestioned; this approach also has implicit medical importance, but it is more broadly dispersed than the pathogen- or human-focused plans. It is the most distinct from structural biology, which it will complement and allow to ensue in parallel. Downsides of this idea are that it is hard to describe, both to the public and even to biologists. The effort would require developing a mapping of the protein universe that will be new and unfamiliar to most researchers. It is unclear whether a new, reliable, and broadly accepted method for defining sequence space could reach currency in time for the second phase of structural genomics to move forward.

The Pfam5000 and data sources

We propose the Pfam5000 approach as one example that may help illuminate the strengths and limitations of a variety of target selection procedures. It is intended to provide a balance between the previously suggested approaches to target selection. Briefly, the Pfam5000 is an regularly updated index of the 5000 most important, tractable families in the Pfam database[19] at a given point in time. The biological value of solving these structures is self-evident, and the medical value will be implicit yet clear. While slightly more complicated to explain that "all of a pathogen," it is relatively succinct and expressive. Biologists are familiar with Pfam and will be able to immediately understand what it describes. The public will need slightly more background, but this should not be unduly difficult to provide. The effort, with its focus on providing structural knowledge for the largest number of protein sequences, is clearly distinct from structural biology. Like the approaches relying on defining new sequence families and a global mapping, it provides good value. Unlike them, it draws upon existing highly curated and well-recognized resources, allowing analysis and plans to be laid in place immediately, with no delay and modest expenditure. Finally the figure of 5000 is intended to ensure tractability.

How does one pick the 5000 most important, tractable families? The simplest definition for importance is size: number of proteins that belong to a family may be taken as a proxy for its significance. Many other primary criteria are also possible, such as first selecting all Pfam families with human proteins and then filling the remainder by size, or emphasizing families with many citations in the literature, as suggested by an anonymous referee. As we show here, so long as size is a secondary criterion in the current Pfam database, the selected set of proteins is relatively insensitive to a wide variety of primary criteria. The 5000 number was chosen to be feasible; it will include roughly 2000 proteins whose structures are known already and perhaps 500 whose structures are solved by groups beyond PSI. The remaining 2500 structures represent 500 per year, a figure that seems plausible given the intended investment. The intent is to continually monitor progress in PSI as well as new Pfam families, to update the Pfam5000 to exclude families that are not tractable and to include new families of great importance.

Fundamental to the Pfam5000 is the Pfam database. Pfam is a collection of protein families manually curated from the Swiss-Prot and TrEMBL sequence databases [20]. Version 10.0 contains 6190 curated families in the Pfam-A collection, which match 86.5% of the proteins in Swiss-Prot 41.0 and 74.5% of the proteins in Pfamseq 10.0, a non-redundant database which includes all sequences in Swiss-Prot 41.0 and TrEMBL 23.15. The Pfam database includes annotations of all sequences in Pfamseq, and these annotations were used in our analysis of coverage of Swiss-Prot and TrEMBL.

Critically, the curators of Pfam now primarily select families based on their size; thus, Pfam represents the roughly 6000 largest families represented in sequence databases.

To evaluate the benefits of target selection based on Pfam, we mapped Pfam families onto Swiss-Prot, TrEMBL, and currently sequenced proteomes. For Swiss-Prot and TrEMBL, we used the mappings included with the Pfam database. We obtained Pfam annotations for complete proteomes from the Proteome Analysis database [21]. These mappings were used to evaluate coverage by Pfam of Swiss-Prot, TrEMBL, and complete proteomes on both a per-protein and per-residue basis, in order to make informed decisions about which targets to prioritize in the next phase of the PSI. The benefits of this strategy are compared to those resulting from solving an entire bacterial proteome, such as that of *Mycoplasma genitalium* or *Mycobacterium tuberculosis*. We also compare this strategy to the strategy of random target selection within the human proteome, or randomly chosen proteins from all currently sequenced genomes.

Methods

Our analysis of the Pfam5000 strategy, choosing targets from roughly the 5000 largest families in Pfam, is currently based on Pfam 10.0. Pfam 10.0 contains 6190 curated families in the Pfam-A database. (More recent versions of Pfam since have been released, but were not included in this analysis.) Pfam includes a mapping of all Pfam families to sequences in Pfamseq 10, a nonredundant database which includes all sequences in Swiss-Prot 41.0 and TrEMBL 23.15. Family size is defined as the number of unique sequences in Pfamseq matching a Pfam family. We calculated statistics separately on the 127,046 sequences from Swiss-Prot, and the full set (denoted SP+TrEMBL), which includes 984,936 sequences. The "seg" program[22] (version dated 5/24/2000) was run on all sequences in Pfamseq 10 to identify putative low complexity regions. The "ccp" program[23] (version dated 6/14/1998) was used to predict coiled coil regions in all sequences, and TMHMM 2.0a [24] was used to predict the locations of transmembrane helices. Default options were used for all programs.

This analysis has two targets: the known universe of sequences represented by Pfamseq, and individual proteomes. The Proteome Analysis database was used to map Pfam domains to protein sequences of 152 complete genomes, including 10 eukaryotes, 16 archaea, and 126 bacteria. The proteome for each organism includes a set of proteins curated from the Swiss-Prot, TrEMBL, and TrEMBL-new databases, and additional eukaryotic proteins are added from the Ensembl [25] database. All proteins except those in TrEMBL-new are annotated with hidden Markov models [26,27] from the InterPro [28] database. Since InterPro includes models from Pfam, we used the supplied InterPro annotations to map Pfam domains onto each protein. The current version of InterPro includes Pfam 9.0. Thus, the 470 families added to Pfam between version 9.0 and version 10.0 were not identified in the proteome sequences, but few of these families are included in the Pfam5000; nonetheless, this means that the coverage numbers for proteomes are slight underestimates. 21 proteomes were excluded from our analysis because 10% or more of their proteins are currently only in TrEMBL-new and thus not yet annotated. Low complexity, coiled coil, and transmembrane regions in proteome sequences were predicted using the same methods as above.

The ultimate goal of structural genomics is to provide structural information for the complete repertoire of biological macromolecules. We measure progress towards that goal as "coverage." Coverage of a proteome is the fraction of its sequences or residues that are covered. Per-sequence coverage is measured as the fraction of sequences that have at least one domain that belongs to a family with a representative whose structure is to be experimentally characterized; this would allow the relevant domain to have its fold assigned. Per-residue coverage by Pfam families was calculated using the beginning and end residues annotated in Pfamseq and the Proteome Analysis databases. All residues between the endpoints were annotated as part of the matching family, ignoring any potential gaps. Three additional variations of per-residue coverage were also calculated, as described in Table II.

To identify Pfam families with currently known structures, we ran all Pfam-A models against our database of sequences of known structure. This database includes sequences of all proteins currently in the PDB [29], as well as sequences of proteins on hold in the PDB where available, as well as sequences of proteins reported as solved by structural genomics centers in the TargetDB database. This database was updated on 9/22/2003.

The simplest Pfam5000 strategy is to choose the largest 5000 families according to family size. Variants of this strategy were also explored, such as "seeding" the Pfam5000 with families with known structures or families appearing in sequenced genomes. In the latter cases, the set of families was biased towards families meeting certain criteria (e.g., families of known structure) by first choosing families meeting the criteria in descending order by size, followed by families which did not meet the criteria in descending order by size. This method enabled the exploration of variants involving other numbers of families besides 5000. Variants of the Pfam5000 are shown in Table I. In cases where multiple criteria were used (e.g., families represented in bacteria, and those of known structure), families meeting any criterion were prioritized over families not meeting the criteria.

Although all members of a single Pfam family are expected to adopt a similar fold, the evolutionary diversity within a family is often too large to allow accurate modeling of all sequences from each other. Current state-of-the-art comparative modeling methods are able to produce models of medium accuracy (about 90% of the main chain modeled to within 1.5 Å RMS error) when sequence identity between the model and the template is at least 30%; below 30% ID, alignment errors increase rapidly and become the major source of modeling error [8]. We have therefore clustered each Pfam family at 30% ID to estimate the number of targets that would have to be solved to provide coverage of structure space at a medium level of accuracy. The clustering algorithm is the greedy clustering algorithm described previously [30] and currently used to create representative subsets at various levels of sequence identity in the ASTRAL database [31]; sequences from each Pfam family are chosen in descending order by length.

To estimate the scope of a target selection strategy focused on single proteomes, or randomly chosen proteins from all proteomes, all proteins in the Proteome database were mapped to the Pfamseq database using Swiss-Prot accession numbers. Pfam-B annotations from Pfamseq were then transferred to the equivalent sequences in Proteome. Each remaining region containing 50 or more consecutive residues bounded by an end of the sequence, an annotated domain from Pfam, or a transmembrane helix, was assumed to be a singleton (having no similar sequences within other proteomes) for purposes of this analysis, or it would have been automatically included in an existing Pfam-B family. These singleton regions were assumed to contain one or more distinct domains.

Results

Pfam5000

Pfam size. Pfam 10.0 contains 6190 families in the Pfam-A database. A graph showing the historic growth of the Pfam database is shown in Figure 1. A histogram of family sizes of families in Pfam 10.0 is shown in Figure 2. Sizes range from 1 to 37,205, with a median size of 33 sequences. Between version 9.0 and version 10.0, 470 families were added, with sizes ranging from 1 to 292; however, only 5 of these families have a size of over 100, and the median family size is 13. Thus, although more families continue to be added to Pfam, they tend to be smaller than the families already included in the database.

A histogram of family sizes for Pfam 4.1, released almost exactly four years prior to Pfam 10.0, is also shown in Figure 2. Family sizes in Pfam 4.1 range from 2 to 15,924, with a median size of 47 sequences. The number of sequences in Pfamseq grew from 257,043 to 984,936, a factor of almost 4. In both versions, the largest Pfam family is GP120, a viral coat protein. Some of the smallest families in Pfam 4.1 increased only slightly in the four years between Pfam 4.1 and Pfam 10.0; e.g., the Diphtheria toxin R domain family only increased from 3 members to 4. 54 of the 1488 families in Pfam 4.1 were merged with other families by version 10.0. The majority of the growth in Pfam has been in new families: of 1,134,710 annotated Pfam-A regions in Pfam 10.0, 575,435 (51%) are in families which were not present in Pfam 4.1; 253,188 (22%) are additions to families which were present in Pfam 4.1, and the remaining proteins were previously annotated in Pfam 4.1.

Diminishing returns in coverage. Pfam coverage of Swiss-Prot proteins in Pfamseq is shown in Figure 3. As the number of Pfam families chosen increases, the coverage of Swiss-Prot by these families also increases; because families are chosen in order by family size, there are diminishing returns as smaller families are considered.

Coverage of known structures. Currently, 2,108 of the 6,190 Pfam-A families (34%) match proteins of known structure. Predictably, larger families have a better chance of having a known structure, as shown in Figure 4. As shown in Figure 4b, the set of families with known structures may be slightly biased towards human proteins, reflecting prior experimental interest in these proteins. Since 1998, structures for approximately 20 new Pfam families have become available every month, based on the release dates of structures from the PDB (Figure 4c); remarkably, this number has not increased even as the number of structures solved per month has grown from about 100 to more than 300 over the same time period. By contrast, Pfam has grown rapidly, using ever-evolving methods of curation. As a result, the fraction of Pfam families with known structure has decreased from 49% in 1999 to 34% today (Figure 4d).

Coverage of proteomes. Coverage of 131 proteomes by Pfam families with known structure, Pfam5000 families (under several bias variations), and by all Pfam-A families is included as supplementary information. Some of the data are shown in Table III, which summarizes percent coverage on a per-protein (the percentage of proteins in the proteome with *any* coverage by the applicable set of Pfam families) and per-residue basis for 10 prominent organisms, as described in the Methods section. Figure 5a shows how the coverage grows with the number of Pfam families characterized.

Several results are apparent from the table. First, that solving the structures of the Pfam5000 families would give almost all the benefits of solving the structures of all 6190 Pfam-A families. Second, that this would provide widespread coverage across a diverse range of organisms.

Only 4905 Pfam families appear in at least one sequenced prokaryotic or eukaryotic genome described in Proteome. The other families in Pfam 9.0 are from viruses or un-sequenced species (the largest family in Pfam is GP120, a viral protein). Of these, 1584 are specific to prokaryotes, and 1729 to eukaryotes. 1592 families appear in both eukaryotes and prokaryotes. Of the 3176 families found in prokaryotes, 1573 are specific to bacteria, 579 to archaea, and 1024 are found in both. These results imply that variants of Pfam5000 may be “seeded” with one or more of these sets and achieve optimal coverage (within the constraints of Pfam-A) with fewer than 5000 families.

Pfam5000 versions biased towards known structures, prokaryotes, or eukaryotes. Table IV summarizes compares the results in Table III to the results for several other variants of Pfam5000 (see Table I). According to this table, the variation of Pfam5000 used makes little difference in the final coverage of each genome. Biasing the families towards prokaryotic families improves prokaryotic coverage by about 1%, at the expense of eukaryotic coverage, and vice versa. Figure 5b shows the growth in structural information using a Pfam5000 biased towards structure; the “bump” at around 2100 families is due to small families of known structure being prioritized over large families without known structure.

Variations of per-residue coverage calculation methods. Several variant methods of calculating per-residue coverage are described in Table II. Coverage in the same 10 organisms using each variation is shown in Table V, using the structure-biased version of Pfam5000.

The various methods of per-residue calculation all give different results, in some cases as much as 10%. The fourth variant probably gives the closest approximation of the tractable portion of the proteome, as regions ignored by this calculation are predicted to represent coiled coil, transmembrane helices, low complexity unstructured regions, and short loops between domains and/or transmembrane helices.

Synergy with structural biology. To date, structural biologists have solved over 24,000 protein structures, from 2108 different Pfam-A families. Current coverage of 10 organisms, as well as Swiss-Prot and TrEMBL, is shown in Table VI. While per-protein coverage of most organisms is currently between 40 and 50%, per-residue coverage is much lower: 37-48% for prokaryotes, and 24-35% for eukaryotes. Coverage is greatest among well-studied model organisms such as *E. coli* and mouse. The estimated coverage of SwissProt+TrEMBL (55.3% of proteins, or 45.6% of residues) is very similar to other current estimates [32].

Incremental benefit of solving structures for the remaining 2892 families in Pfam5000 is also shown in Table VI. This progress would be approximately equivalent to 1/3 - 1/2 of the current progress to date on each genome, or an additional 11-24% more coverage of proteins and 9-20% more coverage of residues. Incremental improvements would be greatest among prokaryotes; targets from these families would also likely be the most tractable and provide the earliest benefits within the 5 year period of the second phase of the PSI.

Extrapolation to future target selection work. Beyond the largest 5000 families, the broad applications of solving structures for a Pfam family rapidly diminish. Incremental improvements to coverage in 10 organisms, Swiss-Prot and TrEMBL by the remaining 1190 Pfam-A families not in Pfam5000, and by all Pfam-B families, are shown in Table VII. In most cases, the additional improvements in coverage by the remaining Pfam-A families are only 1-2%. While individual targets from each family might be of unique biological interest, structures would not be as widely applicable to modeling proteins from other species as the largest 5000 families. In addition, these families would likely be more difficult to solve, as the relative lack of homologs would make it more likely for a single problematic target to present an experimental bottleneck.

Although solution of all families in Pfam-B would provide additional improvements in coverage more comparable to the benefits of completing the Pfam5000 (Table VI), the large number of targets required would make this strategy intractable with current technology and resources.

Single genome target selection strategy

For comparison to the Pfam5000 strategy, we estimate the amount of work required for complete coverage of the *M. tuberculosis* (TB) and *M. genitalium* (MG) proteomes, and the resulting benefits in coverage. An estimate of the number of targets involved, and the coverage provided in the corresponding organisms, is shown in Table VIII. In MG, over 60% of the proteome (47% of the residues) is already covered by 302 Pfam-A families of known structure. Solving one target from each of the remaining 74 Pfam-A families (the MG-A set) would provide coverage for an additional 11.8% of the residues in the MG proteome. Solving 461 additional targets from Pfam-B families (MG-B) would boost residue coverage by an additional 36.4%. The remaining 4.4% of the proteome exists in 101 singleton regions, which would each require at least one target. This procedure sets a minimum bound on the amount of work required to complete the entire MG genome; presumably, the singletons would be harder due to the unavailability of homologs from other species. These estimates exclude the predictably intractable portions of the genome:

predicted coiled coil, low complexity regions, transmembrane helices, and short linker regions cover approximately 20% of the residues in the proteome.

In TB, over 40% of the proteome is already covered by 804 Pfam-A families of known structure. Solving one target from each of the remaining 375 Pfam-A families (the TB-A set) would provide coverage for 57.3% of the residues in the proteome. 2469 Pfam-B families which match the remaining regions (TB-B) are considered next, each as a single target. Finally, the remaining 1636 regions not hit by Pfam-A or Pfam-B families are considered as individual targets. The minimal effort thus required to complete the proteome would involve at least 4480 targets (TB-A + TB-B + singleton regions), more new targets than required for completion of the Pfam5000.

If we examine coverage of the human proteome by the same families, 102 of the families in TB-A match human proteins. Solution of these structures would provide coverage for an additional 1.0% more human proteins, or 1.2% more residues. 141 Pfam-B families from human are included in TB-B; solution of these would yield coverage for only 0.2% more human proteins (0.2% more residues). It was assumed that the singleton proteins from TB would not match any human proteins; if they had, they would probably already be part of Pfam-B.

A full analysis of the structural coverage in other species by the MG and TB single-genome strategies is given in Table IX. While coverage benefits of structural completion of these two prokaryotes are generally higher in prokaryotes than in eukaryotes, the incremental improvement in structural coverage is only higher than for the Pfam5000 in the case of closely related species (e.g., *M. pneumonia* and *M. genitalium*).

Random target selection

We also analyzed the benefits of the strategy of choosing proteins at random from among the 597,532 proteins in the Proteome database. We divided each protein into annotated Pfam-A families, Pfam-B families, and remaining singleton regions. Singleton regions were unannotated regions of 50 or more residues bounded by annotated Pfam families or predicted transmembrane helices. Predicted Pfam families were used to calculate coverage in all proteomes; singleton regions were assumed to not match any other proteins, or they would already be likely to be annotated in Pfam-B.

To compare the amount of work required for this strategy to the Pfam5000 strategy, we assumed that each Pfam-A, Pfam-B, or singleton region would require one target. Random selection of 3197 proteins resulted in 5000 targets under this assumption. Only 1234 of these targets were Pfam-A families; 1562 were Pfam-B families, and the remaining 2204 targets were singleton regions. We also investigated the consequences of selection of 5000 proteins (8376 targets using the previous calculation) under the optimistic assumption that each might be solved as a single target.

We also selected 5000 random targets from the human proteome using the same procedure. Random selection of 2373 proteins resulted in 5000 targets. Of these, 786 were Pfam-A families, 2191 were Pfam-B families, and the remaining 2023 were singletons. We also calculated the coverage resulting from selection of 5000 complete proteins (9981 targets using the previous calculation).

Coverage of several proteomes using these target selection strategies are shown in Figure 8 and Table IX. As expected, random target selection tends to favor larger families, as indicated by the diminishing returns in coverage as more families are chosen. However, coverage using the random strategy is diminished relative to the Pfam5000. Even under the most optimistic assumption that multi-domain proteins will always require only a single target, both per-protein and per-residue coverage are about 10% lower than provided by the Pfam5000 strategy. Selection of random targets from humans rather than all species improves coverage in eukaryotes at the expense of coverage in prokaryotes, but *Homo sapiens* is the only species in which the resulting coverage would be higher than resulting from the Pfam5000.

Domains of unknown function

Some Pfam domains are annotated as domains of unknown function (DUF). In addition to this keyword, we annotated domains described as "hypothetical protein", "unknown function", or "uncharacterized protein family" as having unknown function. Of the 6190 families in Pfam 10.0, 1002 families were annotated as unknown function, and 5188 with some known or predicted function. 951 (95%) of the families with unknown function also have unknown structure, and 565 of these are included in Pfam5000 (biased with known structures). Solution of these protein structures might yield insight as to their function, either through homology which was previously unrecognized by sequence analyses, or because the structure might provide testable hypotheses of functions.

Number of targets required for accurate modeling

As described in the Methods section, we assume that 30% identity between a sequence and structural template is required to produce a reasonably accurate model. Previous estimates [17] based on the same assumption have stated that only 16,000 structures would be sufficient to model 90% of the 300,000 proteins known at that time. However, the number of proteins in Swiss-Prot+TrEMBL has more than tripled since that time, as has the number of Pfam families (from 2000 in version 4.4 of Pfam to 6190 in version 10).

We clustered each Pfam family at 30% identity, and call each cluster a "subfamily." A histogram of subfamily sizes is shown in Figure 6; the median family size is only 8, so most structures would yield relatively few models. Larger families also contain slightly large subfamilies; a cumulative total of the number of subfamilies required to model every sequence in Pfam-A, and the number of resulting models produced, are shown in Figure 7. The number of structures required to accurately model every sequence in Pfam-A is over 90,000. While more sophisticated clustering might reduce this number somewhat, this number of targets is prohibitively large to approach within the scope of PSI phase II.

Another estimate of the number of targets required for accurate modeling was made by Liu and Rost [16]; they identify 18,000 clusters suitable for structural genomics studies in eukaryotes. While this number is closer to becoming tractable, almost as many important targets may be found in prokaryotes as well: our analysis identified over 15,000 subfamilies from families found only in prokaryotes.

Discussion

The families in Pfam5000 represent a tractable yet extremely useful set of targets to study in Phase II of the PSI. If all structures in Pfam5000 were solved, we would know the folds of approximately 68% of prokaryotic proteins (covering 59% of residues) and 61% of eukaryotic proteins (40% of residues). While this goal is feasible within the next five years, this structural knowledge would have a broad impact, allowing a 33% - 50% increase in our ability to assign folds to proteins from all sequenced genomes. If modeling and threading methods enjoy similar advances in the next five years, we will be able to produce accurate models for these proteins as well as fold assignments.

Although we explored several variations of the Pfam5000 strategy, prioritizing different groups of families, final coverage of each proteome differed only by about 1% depending on which variant of the strategy is chosen. As long as secondary consideration is given to large families within Pfam, certain families within the set of particular interest to investigators may be prioritized without compromising the overall impact of the project.

In contrast, focusing the efforts of PSI Phase II on one or more tractable genomes, although possibly of immense medical and biological value, would have a very limited impact in structural knowledge of other proteomes. A significant fraction (about 30-40% of the proteins, and 40-60% of the residues) of each proteome is classified in singletons or small families, which may have only 1% overlap with other species of interest. These would be of limited use for modeling proteins from outside their family without a significant breakthrough in structure prediction methods. The degree of effort required to complete the structural repertoire of a single pathogen could alternatively be invested in work which provides an additional 10-20% structural coverage of all proteomes. On the other hand, devoting a portion of effort to solving representatives of smaller families might result in other benefits, such as discovery of novel methods for identifying previously undiscovered remote evolutionary relationships between the small families.

A random target selection strategy would provide some of the benefits of the Pfam5000 strategy, in that representatives of larger families are more likely to be chosen at random. However, as with the single-genome strategy, approximately 40% of the effort would be spent determining the structure of singletons and smaller families.

We estimate that at least 5-10% of any given proteome is either uninteresting or intractable for high-throughput study: these amino acids are in transmembrane segments, coiled coil, regions of low complexity, or in short interstitial regions between domains and/or transmembrane segments. Other proteins, such as those in large complexes, may prove intractable to high throughput structural genomic methods, and require more focused methodical work to determine their structure.

Solving a single target per Pfam family will result in only a coarse-grained structural coverage of sequence space. The number of targets required for finer grained coverage (e.g., a 30% ID cutoff which would enable accurate structural modeling) of the majority of currently known sequences is intractably large, although improved modeling techniques may improve the situation in the future[33]. However, it might be useful to focus some structural genomics efforts on finer grained coverage of some Pfam families. For example, coverage of families of known medical importance would enable modeling of potential drug targets[34]. Fine coverage of some large Pfam families might improve our understanding of how a single family can evolve to take on a diverse variety of functional roles[13].

Protein domains are not found in isolation, and it is often difficult to determine the conformation of multiple domains from the isolated examples. As Teichmann and colleagues have noted, domain arrangements are not random: certain domain organizations (called superdomains) are far more common than others [35,36]. In order to help extend the structural information beyond single domains, it will likely be very useful to solve the structures of superdomains, as suggested by Orengo and colleagues[37].

In the second 5-year phase of the PSI, the NIH requests that effort be split between coarse grained coverage of sequence space, proteins of known medical interest, and contributions from the scientific community. Stephen Burley has suggested that one strategy that combines the advantages of several of these strategies would be to first spend several years focusing on a coarse-grained coverage of sequence space, solving as many of the largest families as possible. This project could begin immediately at minimal cost, and the overview of sequence space provided by this effort would then enable a more informed decision of which families to cover in more fine-grained detail in the later years of the project. It is also useful to consider possible methods which the PSI target selection committee could use to assign particular families to each large-scale structural genomics center. One possible method would be similar to the NBA draft: each center could take turns picking their favorite family until all are assigned. Conversely, an assignment could be revoked by the committee if no progress were made in an extended period of time. The PSI steering committee would also periodically reevaluate the importance of families in the Pfam5000, adding or removing families in response to new information.

The Pfam5000 strategy would complement existing NIH structural biology initiatives well. Structural biology exploits current knowledge of structures to tactically lead to treatments; structural genomics strategically leads to better understanding of biology as a foundation for the next generation of biomedical research. There are no uninteresting human proteins; we may just not know what their importance is. Therefore, a strategy which aims to provide the broadest possible increase in structural knowledge is most likely to lead to exciting avenues of new research in the long term.

Acknowledgements

We thank Christine Orengo, Janet Thornton, and their collaborators for valuable discussions and sharing results; they developed results similar to those developed here. We are grateful to Dawei Lin for suggesting Figure 4c. Sung-Hou Kim has provided insightful comments and constant support. We thank an anonymous referee for suggesting Figure 4d and the comparison with Pfam 4.1. This work is supported by grants from the NIH (1-P50-GM62412, 1-K22-HG00056) and the Searle Scholars Program (01-L-116), and by the U.S. Department of Energy under contract DE-AC03-76SF00098.

Figure Legends

Figure 1: The Pfam database has been growing exponentially since its inception in 1996.

Figure 2: Distribution of family sizes in Pfam 10.0 and Pfam 4.1, released four years before. Family sizes in Pfam 10.0 range from 1 to 37,205, with a median size of 33 sequences. Family sizes in Pfam 4.1 range from 2 to 15,924, with a median size of 47 sequences.

Figure 3: As the number of Pfam families chosen increases, the coverage of Swiss-Prot by these families also increases; because families are chosen in order by family size, there are diminishing returns as smaller families are considered. 3a) Per-protein coverage is shown as a percentage of the total of 127,046 proteins from Swiss-Prot in Pfamseq which have at least one hit from Pfam. 3b) Per-protein coverage is shown as a percentage of the total of 984,936 proteins in Pfamseq. Percent of residues covered is calculated using method #4 from Table II. A vertical line indicates 5000 families.

Figure 4: How much of Pfam currently has known structure? The number of Pfam families of known structure plotted vs. the total number of families, in order of inclusion into Pfam5000. 4a compares the unbiased set (chosen by size) vs. a set "seeded" with families with already known structure. 4b includes some other possibilities for "seeding" the Pfam5000 set, as described in the Methods section. 4c shows how the coverage of Pfam by known structure has increased over time, based on release dates of PDB entries and reported solution dates by structural genomics centers. 82 current Pfam families had known structure prior to 1990. 4d shows the cumulative number of Pfam families, and the number and fraction with known structure, from release 4.0 in May 1999 until release 10.0 in July 2003.

Figure 5: As the number of Pfam families chosen increases, the coverage of proteomes by these families also increases; because families are chosen in order by family size, there are diminishing returns as smaller families are considered. Per-protein coverage is shown as a percentage of the total number of identified proteins in the proteome which have at least one hit from Pfam. Per-residue coverage is calculated using method #4 from Table II. Coverage of human and *E. coli* proteomes are shown, by the unbiased Pfam5000 (5a) and the version of Pfam5000 seeded with families of known structure (5b). A vertical line indicates 5000 families.

Figure 6: Distribution of subfamily sizes. Subfamilies are created by clustering sequences from each Pfam-A family at 30% identity, as described in the methods section. Subfamily size is defined as the number of sequences in the cluster. A histogram of average subfamily size for each Pfam-A family is shown. The mean subfamily size is 8, and the largest subfamily, from the Pfam family HVC_capsid (hepatitis C virus capsid protein), contains 1236 sequences.

Figure 7: Number of subfamilies and sequences covered by Pfam is plotted vs. the total number of families, in order of inclusion into Pfam5000. In 7a, Pfam families are chosen according to family size; the data indicates that large families contain both more and larger subfamilies. In 7b, Pfam families covering at least one known structure are chosen before families of unknown structure. A vertical line indicates 5000 families.

Figure 8: Proteome coverage by Pfam5000 and random target selection. Per-residue coverage is calculated using method #4 from Table II. Coverage of human and *E. coli* proteomes are shown, by the unbiased Pfam5000 (8a), randomly chosen proteins from all proteomes divided into predicted domains (8b), and several single-genome based strategies (8c). A vertical line indicates 5000 families.

Table I: Variants of the Pfam5000 tested

Variant Name	**Bias**
unbiased	None (ordered only by family size)
structure	Known structures
bacteria_str	Bacterial families & known structures
human_str	Human families & known structures
prokaryote_str	Prokaryotic families & known structures
eukaryote_str	Eukaryotic families & known structures
genomic_str	Prokaryotic & eukaryotic families from currently sequenced genomes, & known structures

Table II: Tested methods of calculating per-residue coverage of Pfam domains.

Variant	Description	Rationale
1	Number of residues in regions matched by Pfam, divided by total length of proteins	Default method of calculating coverage
2	Like #1, but not counting unmatched regions of fewer than 50 consecutive residues in denominator	Short regions unlikely to contain complete domain
3	Like #1, but not counting predicted transmembrane, low complexity, or coiled coil residues in denominator	Regions intractable by high throughput methods, unstructured , or repetitive structure
4	Combination of #2 and #3: does not count regions of fewer than 50 consecutive unmatched residues between transmembrane regions and/or Pfam hits. Does not count transmembrane, low complexity or coiled coil in denominator	Does not count any regions unlikely to include new domains, or intractable

Table III: Coverage of several proteomes by currently known structures, Pfam5000 (variant biased towards known structures), and all Pfam-A families. Percent of residues covered is calculated using method #4 from Table II. The Families column shows the total number of distinct Pfam-A families with hits in each genome.

Organism Name	# of Prot.	Known Struct.		Pfam5000		all Pfam-A		Families
		% Prot.	% Res.	% Prot.	% Res.	% Prot.	% Res.	(Pfam-A)
A. thaliana	26209	47.8	27.5	69.2	42.9	70.5	44.0	2194
C. elegans	22602	36.5	25.0	53.7	37.4	55.4	38.6	2039
D. melanogaster	15908	46.1	27.3	59.9	36.0	61.4	36.9	2084
E. coli	4357	51.0	49.2	74.2	67.3	75.0	67.7	1625
H. sapiens	34560	45.4	29.7	56.7	38.8	57.8	39.6	2509
M. jannaschii	1777	42.7	38.6	64.7	58.3	69.2	62.0	852
M. pneumoniae	687	46.1	38.1	70.0	54.5	71.3	55.5	399
M. tuberculosis	3877	47.9	43.1	66.3	57.0	67.8	58.1	1179
M. musculus	38795	52.5	35.3	64.8	45.1	65.8	45.8	2507
R. norvegicus	27479	52.5	35.9	64.6	45.5	65.7	46.3	2292

Table IV: Coverage of proteomes by Pfam5000 biased towards known structures, bacterial families (with known structures) and eukaryotic families (with known structures). Percent of residues covered is calculated using method #4 from Table II.

Organism	Variant: structure		Variant: bacteria_str		Variant: eukaryote_str	
	% Proteins	% Residues	% Proteins	% Residues	% Proteins	% Residues
A. thaliana	69.2	42.9	68.5	42.4	70.5	44.0
C. elegans	53.7	37.4	53.0	36.9	55.4	38.6
D. melanogaster	59.9	36.0	59.3	35.6	61.4	36.9
E. coli	74.2	67.3	75.0	67.7	73.1	66.6
H. sapiens	56.7	38.8	56.2	38.4	57.8	39.6
M. jannaschii	64.7	58.3	65.0	58.9	62.5	56.6
M. pneumoniae	70.0	54.5	71.3	55.5	63.8	50.1
M. tuberculosis	66.3	57.0	67.8	58.1	65.8	56.4
M. musculus	64.8	45.1	64.2	44.7	65.8	45.8
R. norvegicus	64.6	45.5	64.2	45.2	65.7	46.3

Table V: Coverage of proteomes by Pfam5000 biased towards known structures, using several different methods of calculation of per-residue coverage (described in Table II).

Organism	Coverage by Pfam5000, structure variant				
	% Proteins	% Res., method #1	% Res., method #2	% Res., method #3	% Res., method #4
A. thaliana	69.2	36.6	38.1	40.3	42.9
C. elegans	53.7	30.3	31.4	34.7	37.4
D. melanogaster	59.9	28.9	29.8	34.2	36.0
E. coli	74.2	58.6	62.0	62.1	67.3
H. sapiens	56.7	31.8	33.2	36.4	38.8
M. jannaschii	64.7	49.5	52.3	54.0	58.3
M. pneumoniae	70.0	45.2	47.3	50.1	54.5
M. tuberculosis	66.3	47.0	49.2	52.6	57.0
M. musculus	64.8	37.4	39.4	42.0	45.1
R. norvegicus	64.6	37.7	39.6	42.5	45.5

Table VI: Coverage of proteomes, Swiss-Prot (SP), and TrEMBL by structural biology efforts to date, and incremental benefits of solving all families in Pfam5000 (biased towards currently known structures). Percent of residues covered is calculated using method #4 from Table II. Cost/Benefit is the number of families divided by the incremental percentage increase in residue coverage.

Organism	Current Coverage			Incremental Work and Coverage			
	Families	% Proteins	% Residues	Families	% Proteins	% Residues	Cost/Benefit
A. thaliana	1147	47.8%	27.5%	861	13.4%	15.4%	55.9
C. elegans	1102	36.5%	25.0%	742	17.2%	12.4%	59.9
D. melanogaster	1128	46.1%	27.3%	762	13.8%	8.7%	87.6
E. coli	969	51.0%	49.2%	621	23.2%	18.1%	34.3
H. sapiens	1292	45.4%	29.7%	932	11.3%	9.1%	102.4
M. jannaschii	503	42.7%	38.6%	278	22.0%	19.7%	14.1
M. pneumoniae	319	46.1%	38.1%	78	23.9%	16.4%	4.8
M. tuberculosis	804	47.9%	43.1%	349	18.4%	13.9%	25.1
M. musculus	1288	52.5%	35.3%	937	12.3%	9.8%	95.6
R. norvegicus	1229	52.5%	35.9%	843	12.1%	9.6%	87.8
Swiss-Prot	2090	66.3%	53.5%	2455	18.3%	15.5%	158.4
SP+TrEMBL	2108	55.3%	46.5%	2892	19.5%	16.0%	180.8

Table VII: Incremental work and coverage of proteomes, Swiss-Prot (SP), and TrEMBL if all families in Pfam-A, or all families in Pfam-A+Pfam-B were solved, relative to Pfam5000 biased by existing structures (Table VI)Table VII. Percent of residues covered is calculated using method #4 from Table II. Cost/Benefit is the number of families divided by the incremental percentage increase in residue coverage.

Organism	Incremental Coverage - Pfam-A				Incremental Coverage - Pfam-A+B			
	Families	% Proteins	% Residues	Cost/Benefit	Families	% Proteins	% Residues	Cost/Benefit
A. thaliana	186	1.3%	1.1%	169.1	14797	21.8%	38.0%	389.4
C. elegans	195	1.7%	1.2%	162.5	8103	19.5%	23.5%	344.8
D. melanogaster	194	1.5%	0.9%	215.6	8500	13.2%	19.6%	433.7
E. coli	35	0.8%	0.4%	87.5	3464	20.9%	27.3%	126.9
H. sapiens	285	1.1%	0.8%	356.3	19322	11.0%	23.6%	818.7
M. jannaschii	71	4.5%	3.7%	19.2	801	18.1%	23.5%	34.1
M. pneumoniae	2	1.3%	1.1%	1.8	508	26.2%	39.9%	12.7
M. tuberculosis	26	1.5%	1.1%	23.6	2495	22.9%	30.3%	82.3
M. musculus	282	1.0%	0.7%	402.9	15795	7.0%	15.5%	1019.0
R. norvegicus	220	1.1%	0.8%	275.0	6568	2.4%	8.3%	791.3
Swiss-Prot	818	1.9%	1.3%	629.2	34338	9.2%	21.0%	1635.1
SP+TrEMBL	1190	1.1%	0.9%	1322.2	97740	8.8%	16.8%	5817.9

Table VIII: Structures required for coverage of *M. genitalium* (486 total proteins) and *M. tuberculosis* (3877 total proteins). Percent of residues covered is calculated using method #4 from Table II.

Organism	minimum # of targets	Proteins covered (cumulative %)	Cumulative % residues covered
Target Set			
Mycoplasma genitalium			
Pfam-A (already solved)	302	296 (60.9%)	47.4%
MG-A: new Pfam-A families	74	70 (75.3%)	59.2%
MG-B: Pfam-B families	461	117 (99.4%)	95.6%
singleton regions	101	3 (100%)	100%
Mycobacterium tuberculosis			
Pfam-A (already solved)	804	1858 (47.9%)	43.1%
TB-A: new Pfam-A families	375	770 (67.8%)	58.1%
TB-B: Pfam-B families	2469	832 (89.2%)	87.3%
singleton regions	1636	417 (100%)	100%

Table IX: Incremental increase in coverage (percent of residues) of proteomes by single-genome and random target selection strategies, relative to coverage by currently known structures (Table VI). Percent of residues covered is calculated using method #4 from Table II. MG-A and TB-A refer to the 74 Pfam-A families of unknown structure from *M. genitalium* and the 375 families from *M. tuberculosis* described in Table VIII. "All MG" refers to the entire *M. genitalium* genome, and "All TB" refers to the entire *M. tuberculosis* genome. Pfam5000 refers to the 2892 families of unknown structure from Pfam-A.

Organism	Strategy								
	MG-A	All MG	TB-A	All TB	5000 Random domains	5000 Random proteins	5000 Human domains	5000 Human proteins	Pfam 5000
A. thaliana	0.5	0.8	2.3	2.9	7.8	11.3	4.0	8.2	15.4
C. elegans	0.2	0.5	1.7	2.1	5.0	7.9	5.0	7.4	12.4
D. melanogaster	0.3	0.6	1.3	1.6	3.1	5.6	5.5	8.1	8.7
E. coli	2.3	3.0	8.1	10.9	11.6	15.7	1.7	3.1	18.1
H. sapiens	0.2	0.5	1.0	1.2	3.8	6.0	9.7	14.8	9.1
M. jannaschii	1.9	2.5	8.3	10.2	8.9	14.0	1.9	4.6	19.7
M. pneumoniae	15.5	49.6	6.8	9.3	9.7	13.0	2.2	2.8	16.4
M. tuberculosis	1.3	2.0	15.0	56.9	9.0	12.7	1.2	2.2	13.9
M. musculus	0.3	0.4	1.0	1.2	3.9	5.9	7.7	11.4	9.8
R. norvegicus	0.3	0.3	1.1	1.2	3.5	5.4	6.3	9.1	9.6

References

1. Burley SK, Bonanno JB. Structural genomics. Methods Biochem Anal 2003;44:591-612.
2. Blundell TL, Mizuguchi K. Structural genomics: an overview. Prog Biophys Mol Biol 2000;73(5):289-295.
3. Brenner SE. A tour of structural genomics. Nat Rev Genet 2001;2(10):801-809.
4. Montelione GT. Structural genomics: an approach to the protein folding problem. Proc Natl Acad Sci U S A 2001;98(24):13488-13489.
5. Chance MR, Bresnick AR, Burley SK, Jiang JS, Lima CD, Sali A, Almo SC, Bonanno JB, Buglino JA, Boulton S, Chen H, Eswar N, He G, Huang R, Ilyin V, McMahan L, Pieper U, Ray S, Vidal M, Wang LK. Structural genomics: a pipeline for providing structures for the biologist. Protein Sci 2002;11(4):723-738.
6. Brenner SE, Berry A. A quantitative methodology for the de novo design of proteins. Protein Sci 1994;3(10):1871-1882.
7. Brenner SE. Target selection for structural genomics. Nat Struct Biol 2000;7 Suppl:967-969.
8. Baker D, Sali A. Protein structure prediction and structural genomics. Science 2001;294(5540):93-96.
9. Whisstock JC, Lesk AM. Prediction of protein function from protein sequence and structure. Q Rev Biophys 2003;36(3):307-340.
10. Todd AE, Orengo CA, Thornton JM. Evolution of function in protein superfamilies, from a structural perspective. J Mol Biol 2001;307(4):1113-1143.
11. Goldsmith-Fischman S, Honig B. Structural genomics: computational methods for structure analysis. Protein Sci 2003;12(9):1813-1821.
12. PSI-phase 1 and beyond. Nat Struct Mol Biol 2004;11(3):201.
13. Harrison SC. Whither structural biology? Nat Struct Mol Biol 2004;11(1):12-15.
14. Matte A, Sivaraman J, Ekiel I, Gehring K, Jia Z, Cygler M. Contribution of structural genomics to understanding the biology of Escherichia coli. J Bacteriol 2003;185(14):3994-4002.
15. Goulding CW, Apostol M, Anderson DH, Gill HS, Smith CV, Kuo MR, Yang JK, Waldo GS, Suh SW, Chauhan R, Kale A, Bachhawat N, Mande SC, Johnston JM, Lott JS, Baker EN, Arcus VL, Leys D, McLean KJ, Munro AW, Berendzen J, Sharma V, Park MS, Eisenberg D, Sacchettini J, Alber T, Rupp B, Jacobs W, Jr., Terwilliger TC. The TB structural genomics consortium: providing a structural foundation for drug discovery. Curr Drug Targets Infect Disord 2002;2(2):121-141.
16. Liu J, Rost B. Target space for structural genomics revisited. Bioinformatics 2002;18(7):922-933.
17. Vitkup D, Melamud E, Moult J, Sander C. Completeness in structural genomics. Nat Struct Biol 2001;8(6):559-566.
18. Liu J, Hegyi H, Acton TB, Montelione GT, Rost B. Automatic Target Selection for Structural Genomics on Eukaryotes. Proteins 2004;56(2):188-200.
19. Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, Khanna A, Marshall M, Moxon S, Sonnhammer EL, Studholme DJ, Yeats C, Eddy SR. The

Pfam protein families database. Nucleic Acids Res 2004;32 Database issue:D138-141.

20. 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(1):365-370.
21. Pruess M, Fleischmann W, Kanapin A, Karavidopoulou Y, Kersey P, Kriventseva E, Mittard V, Mulder N, Phan I, Servant F, Apweiler R. The Proteome Analysis database: a tool for the in silico analysis of whole proteomes. Nucleic Acids Res 2003;31(1):414-417.
22. Wootton JC. Non-globular domains in protein sequences: automated segmentation using complexity measures. Comput Chem 1994;18(3):269-285.
23. Lupas A. Prediction and analysis of coiled-coil structures. Methods Enzymol 1996;266:513-525.
24. Krogh A, Larsson B, von Heijne G, Sonnhammer EL. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol 2001;305(3):567-580.
25. Clamp M, Andrews D, Barker D, Bevan P, Cameron G, Chen Y, Clark L, Cox T, Cuff J, Curwen V, Down T, Durbin R, Eyras E, Gilbert J, Hammond M, Hubbard T, Kasprzyk A, Keefe D, Lehvaslaiho H, Iyer V, Melsopp C, Mongin E, Pettett R, Potter S, Rust A, Schmidt E, Searle S, Slater G, Smith J, Spooner W, Stabenau A, Stalker J, Stupka E, Ureta-Vidal A, Vastrik I, Birney E. Ensembl 2002: accommodating comparative genomics. Nucleic Acids Res 2003;31(1):38-42.
26. Krogh A, Brown M, Mian IS, Sjolander K, Haussler D. Hidden Markov models in computational biology. Applications to protein modeling. J Mol Biol 1994;235(5):1501-1531.
27. Eddy SR. Profile hidden Markov models. Bioinformatics 1998;14(9):755-763.
28. Mulder NJ, Apweiler R, Attwood TK, Bairoch A, Barrell D, Bateman A, Binns D, Biswas M, Bradley P, Bork P, Bucher P, Copley RR, Courcelle E, Das U, Durbin R, Falquet L, Fleischmann W, Griffiths-Jones S, Haft D, Harte N, Hulo N, Kahn D, Kanapin A, Krestyaninova M, Lopez R, Letunic I, Lonsdale D, Silventoinen V, Orchard SE, Pagni M, Peyruc D, Ponting CP, Selengut JD, Servant F, Sigrist CJ, Vaughan R, Zdobnov EM. The InterPro Database, 2003 brings increased coverage and new features. Nucleic Acids Res 2003;31(1):315-318.
29. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res 2000;28(1):235-242.
30. Brenner SE, Chothia C, Hubbard TJ. Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. Proc Natl Acad Sci U S A 1998;95(11):6073-6078.
31. Chandonia JM, Hon G, Walker NS, Lo Conte L, Koehl P, Levitt M, Brenner SE. The ASTRAL Compendium in 2004. Nucleic Acids Res 2004;32 Database issue:D189-192.
32. Pieper U, Eswar N, Braberg H, Madhusudhan MS, Davis FP, Stuart AC, Mirkovic N, Rossi A, Marti-Renom MA, Fiser A, Webb B, Greenblatt D, Huang CC, Ferrin TE, Sali A. MODBASE, a database of annotated comparative protein structure

models, and associated resources. Nucleic Acids Res 2004;32 Database issue:D217-222.

33. Heger A, Holm L. More for less in structural genomics. J Struct Funct Genomics 2003;4(2-3):57-66.

34. Burley SK, Bonanno JB. Structural genomics of proteins from conserved biochemical pathways and processes. Curr Opin Struct Biol 2002;12(3):383-391.

35. Apic G, Gough J, Teichmann SA. Domain combinations in archaeal, eubacterial and eukaryotic proteomes. J Mol Biol 2001;310(2):311-325.

36. Apic G, Huber W, Teichmann SA. Multi-domain protein families and domain pairs: comparison with known structures and a random model of domain recombination. J Struct Funct Genomics 2003;4(2-3):67-78.

37. Lee D, Grant A, Buchan D, Orengo C. A structural perspective on genome evolution. Curr Opin Struct Biol 2003;13(3):359-369.

Description of supplementary information

The following files are included as supplementary information.

pfam5k_proteome_all_bacteria_str.txt
pfam5k_proteome_all_eukaryote_str.txt
pfam5k_proteome_all_genomic_str.txt
pfam5k_proteome_all_human_str.txt
pfam5k_proteome_all_prokaryote_str.txt
pfam5k_proteome_all_structure.txt
pfam5k_proteome_all_unbiased.txt

The above files contain summaries of every proteome in the Proteome database with at least 90% annotation. The contents are documented in the files. Each of the files contains one of the 7 variants of seeding Pfam5000 shown in Table I, as indicated by the file name.

pfam5k_proteome_bacteria_str.txt
pfam5k_proteome_eukaryote_str.txt
pfam5k_proteome_genomic_str.txt
pfam5k_proteome_human_str.txt
pfam5k_proteome_prokaryote_str.txt
pfam5k_proteome_structure.txt
pfam5k_proteome_unbiased.txt

The above files contain detailed results for the 7 variants (as indicated by the file names) on the 10 organisms described in Tables III-VII. Each organism is in a separate section of the file. Rows of data after each organism contain the following (space-separated) columns:

1) number of Pfam families (selected in the order implied by the seeding method, as described in the Methods section and Table 1.
2) number of proteins in the proteome hit by at least one of these families
3) total # of proteins in the proteome
4) total # of proteins with at least one predicted transmembrane region
5) # of residues covered by the Pfam hits in (2)
6) total # of residues in the proteome
7) total # of residues in unmatched regions of less than 50 residues bounded by Pfam hits or ends of the sequence
8) total # of residues in unmatched regions of less than 50 residues bounded by Pfam hits, ends of the sequence, or predicted transmembrane regions
9) total number of transmembrane, low complexity, and coiled coil residues predicted in regions unmatched by Pfam hits
10) total number of transmembrane, low complexity, and coiled coil residues predicted in regions matched by Pfam hits
11) total residues in predicted transmembrane regions
12) total residues in predicted low complexity regions
13) total residues in predicted coiled coil regions
14) total residues in predicted transmembrane, low complexity, coiled coil regions (such regions could potentially overlap, so this is not the sum of 11-13)

15 - 27) the same figures as columns 2-14, recalculated when "difficult" proteins are excluded from the proteome. "Difficult" proteins are defined as any with at least one transmembrane region, or a region of predicted coiled coil or low complexity of at least 20 consecutive residues.

pfam5k_sp_bacteria_str.txt
pfam5k_sp_eukaryote_str.txt
pfam5k_sp_genomic_str.txt
pfam5k_sp_human_str.txt
pfam5k_sp_prokaryote_str.txt
pfam5k_sp_structure.txt
pfam5k_sp_unbiased.txt

The above files contain the same stats as above, calculated on Pfamseq (Swiss-Prot + TrEMBL), for the same 7 variations of the Pfam5000 seeding method (as indicated by the file names).

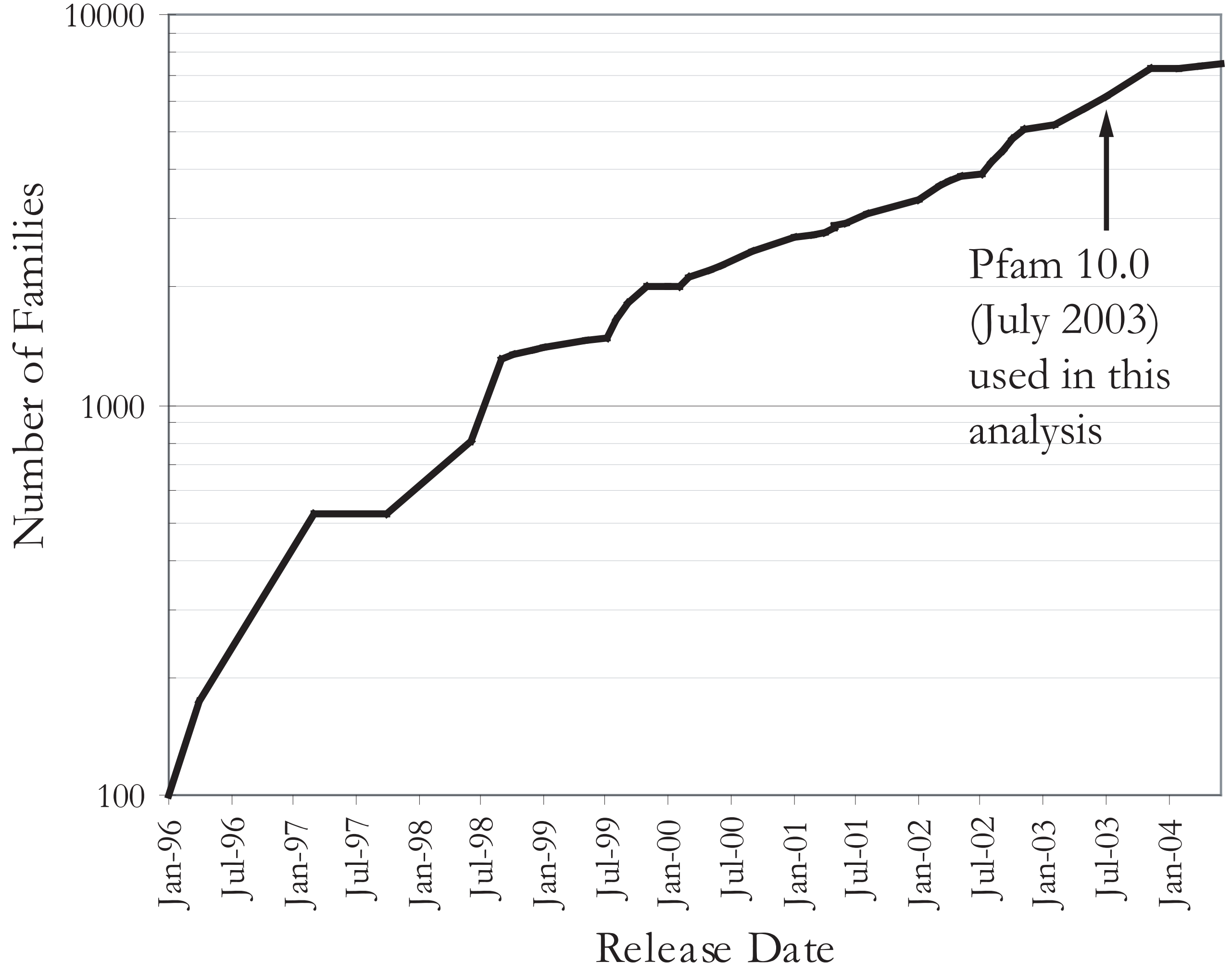
pfam_duf.txt
pfam_notduf.txt

The above files contain the names of Pfam families documented as DUF or not DUF.

pfam5k_proteome_all_mg.txt
pfam5k_proteome_all_tb.txt
pfam5k_proteome_all_random.txt
pfam5k_proteome_all_random_human.txt

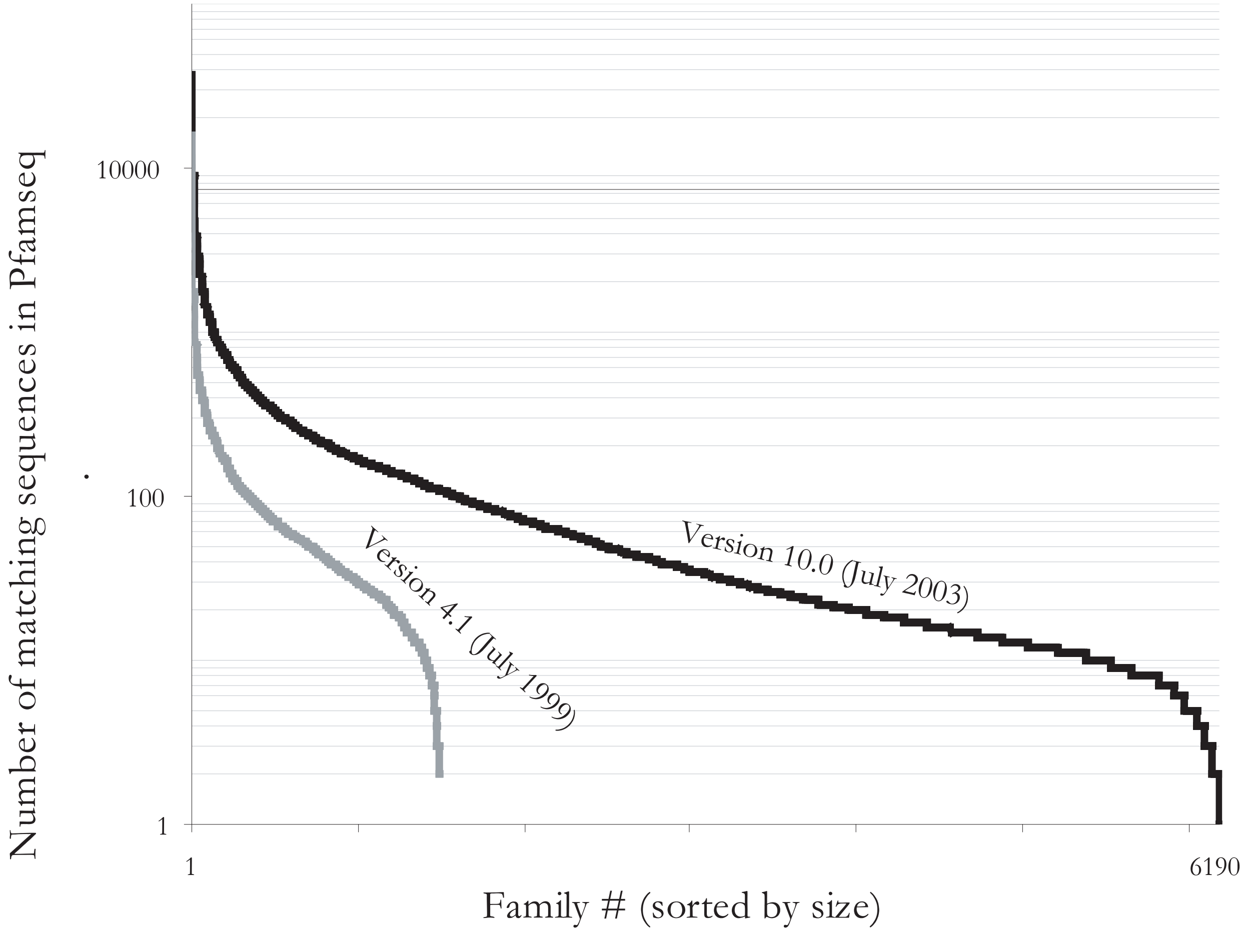
The above file contains summaries of all proteomes in the Proteome database, using the families found in ***M. genitalium*** **(mg),** ***M. tuberculosis*** **(tb), 5000 randomly chosen proteins from all proteomes (random), and 5000 randomly chosen human proteins (random_human)** (as described in Table IX).

Figure 1: Growth of Pfam-A database

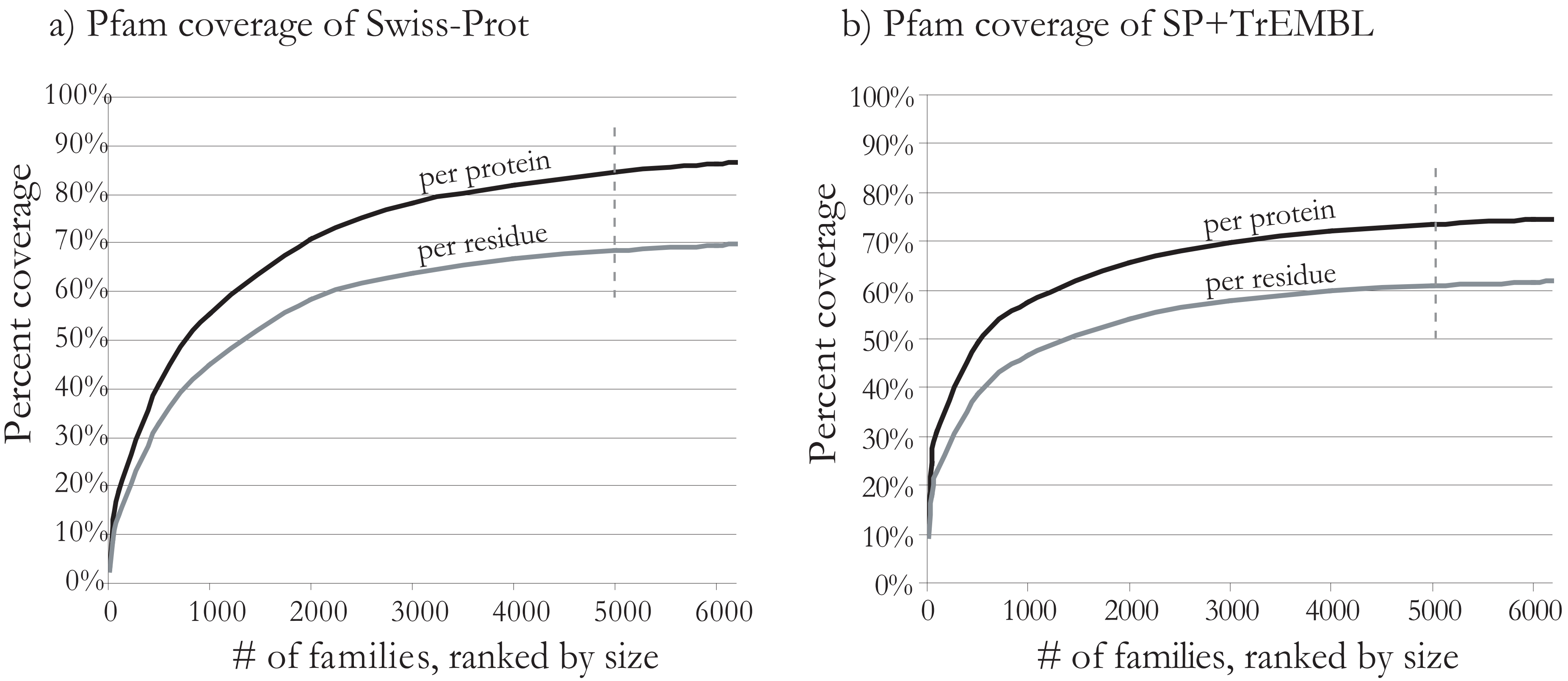


Chandonia & Brenner, Figure 2.

Figure 2: Pfam family size distribution

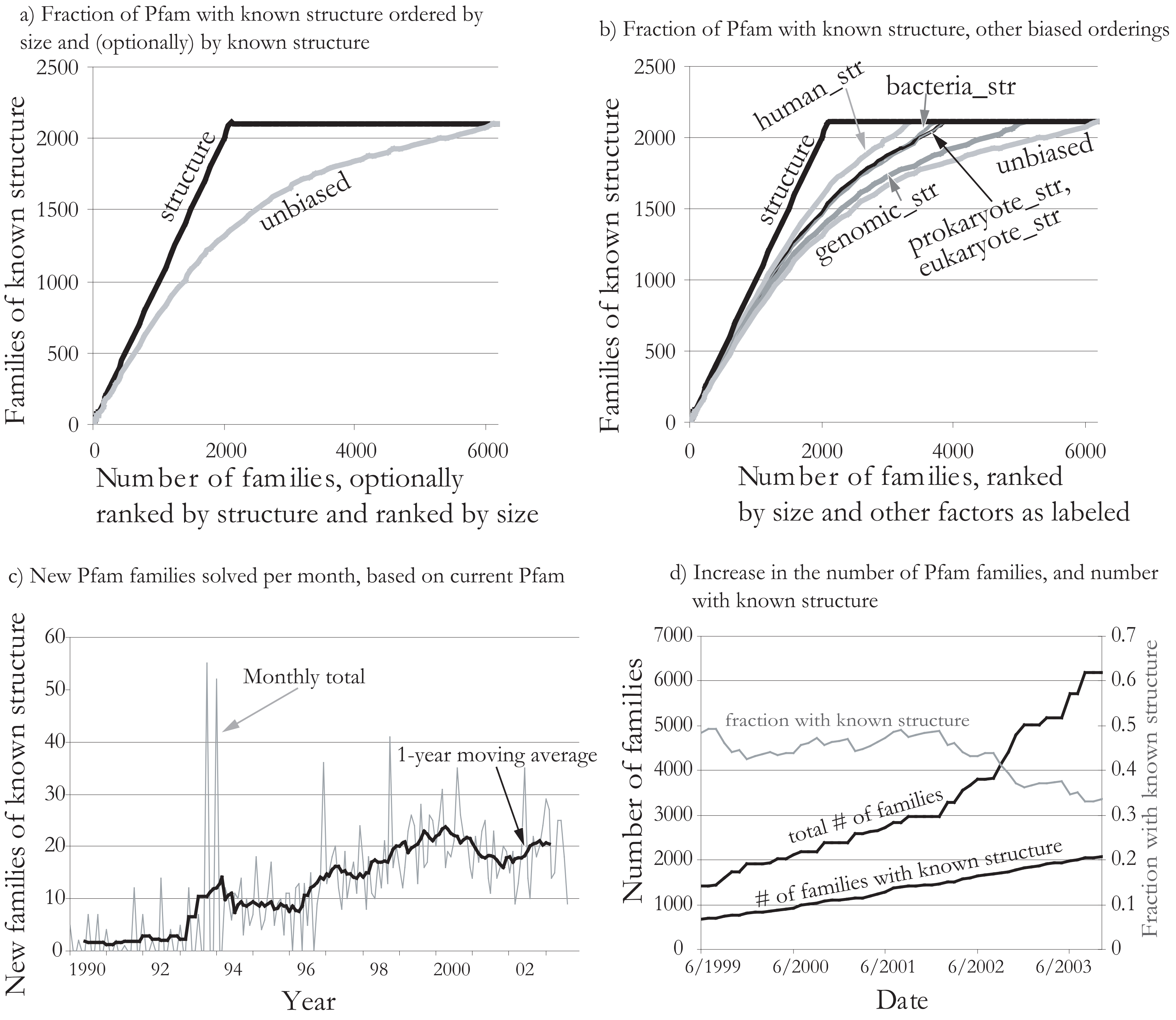


Chandonia & Brenner, Figure 3.


a) Pfam coverage of Swiss-Prot

b) Pfam coverage of SP+TrEMBL

Chandonia & Brenner, Figure 4.



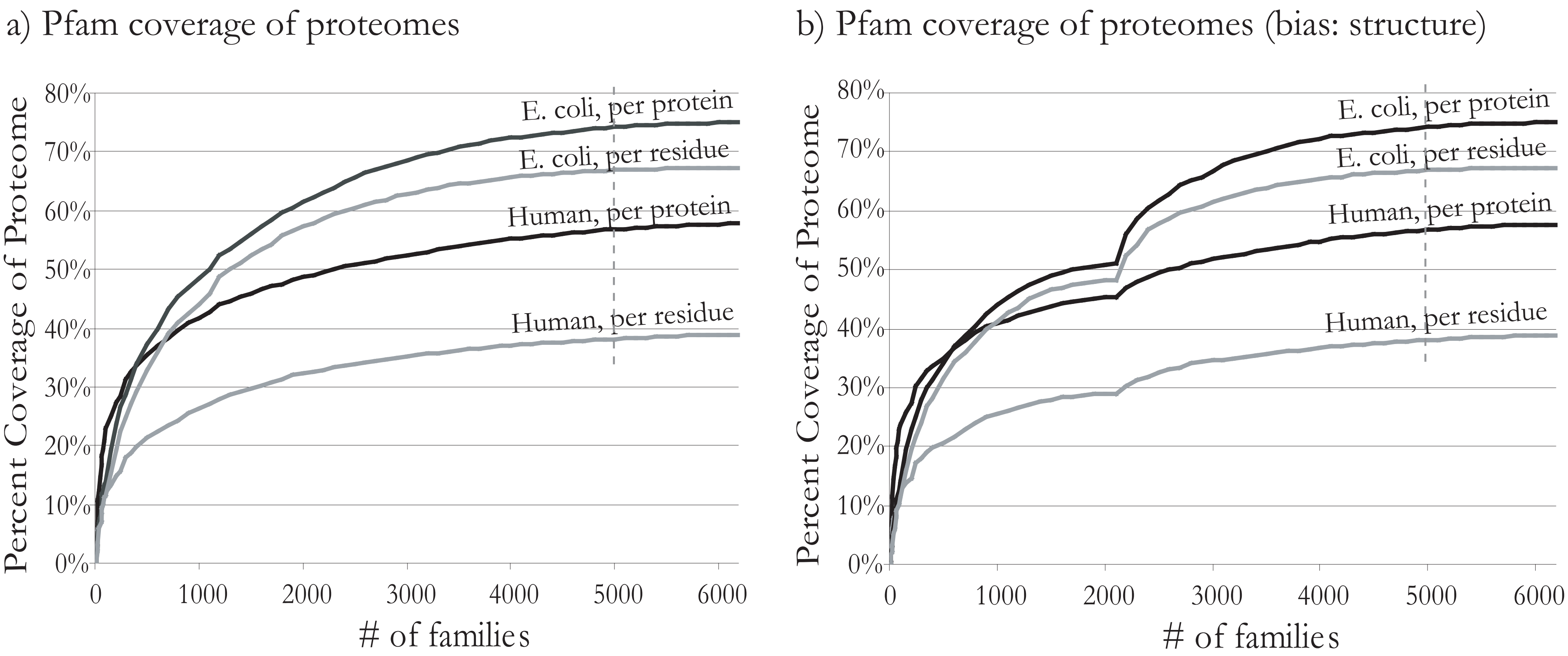
a) Fraction of Pfam with known structure ordered by size and (optionally) by known structure

b) Fraction of Pfam with known structure, other biased orderings

c) New Pfam families solved per month, based on current Pfam

d) Increase in the number of Pfam families, and number with known structure

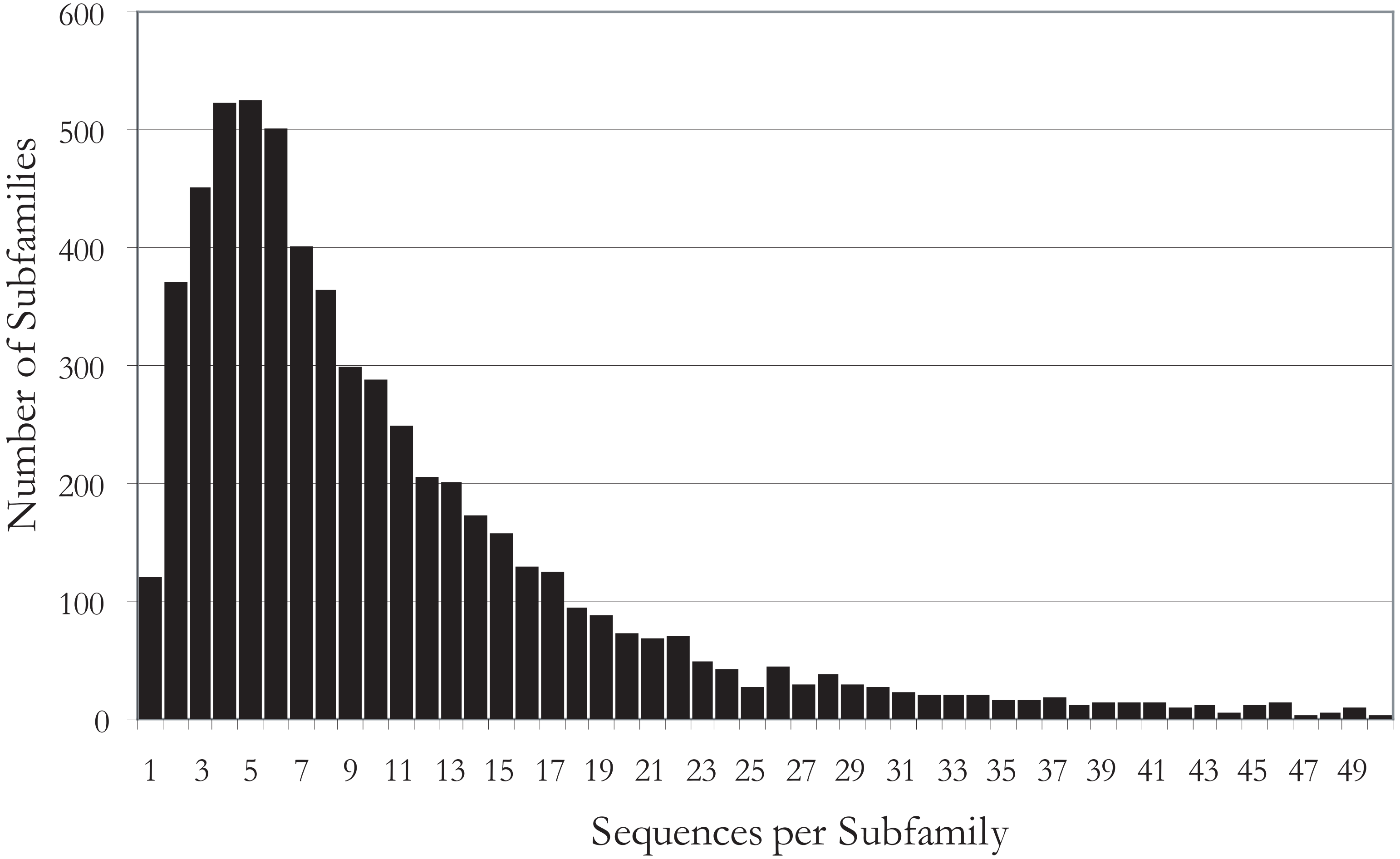
Chandonia & Brenner, Figure 5.



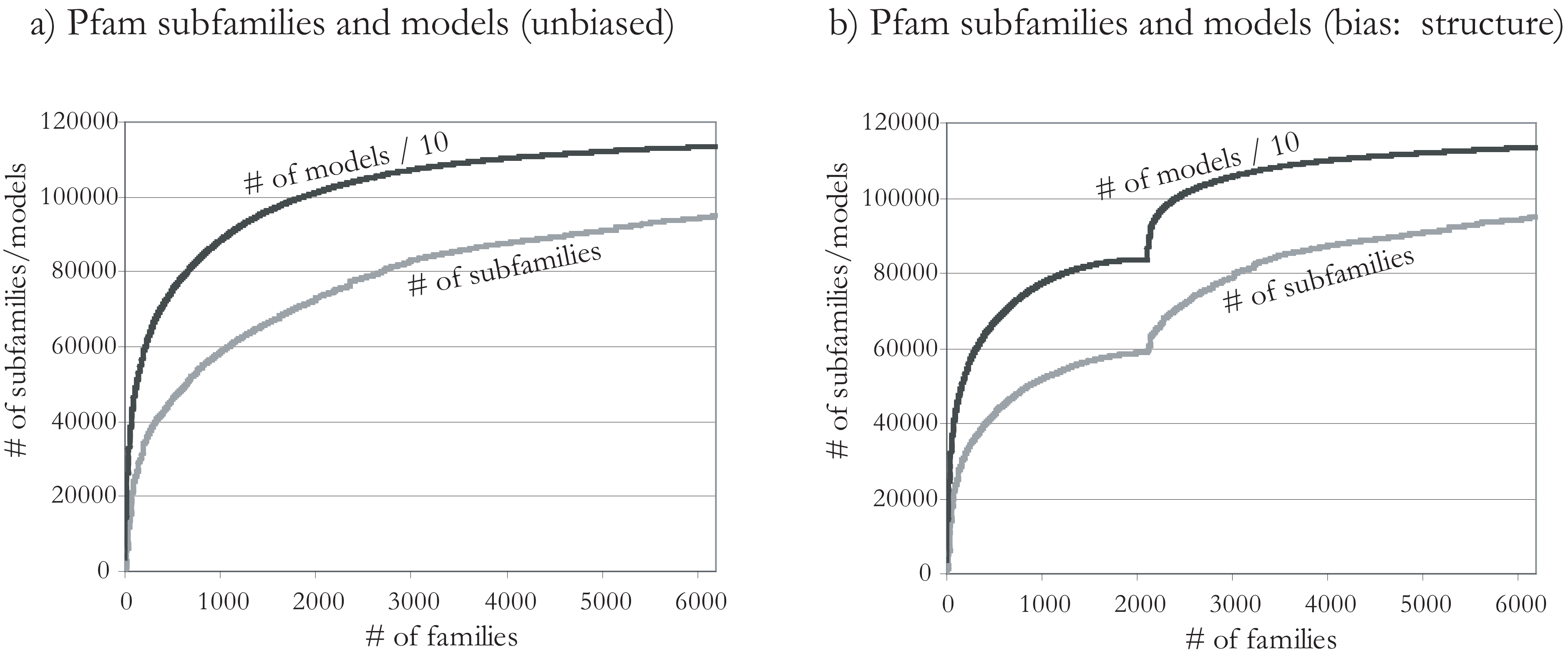
a) Pfam coverage of proteomes

b) Pfam coverage of proteomes (bias: structure)

Chandonia & Brenner, Figure 6.



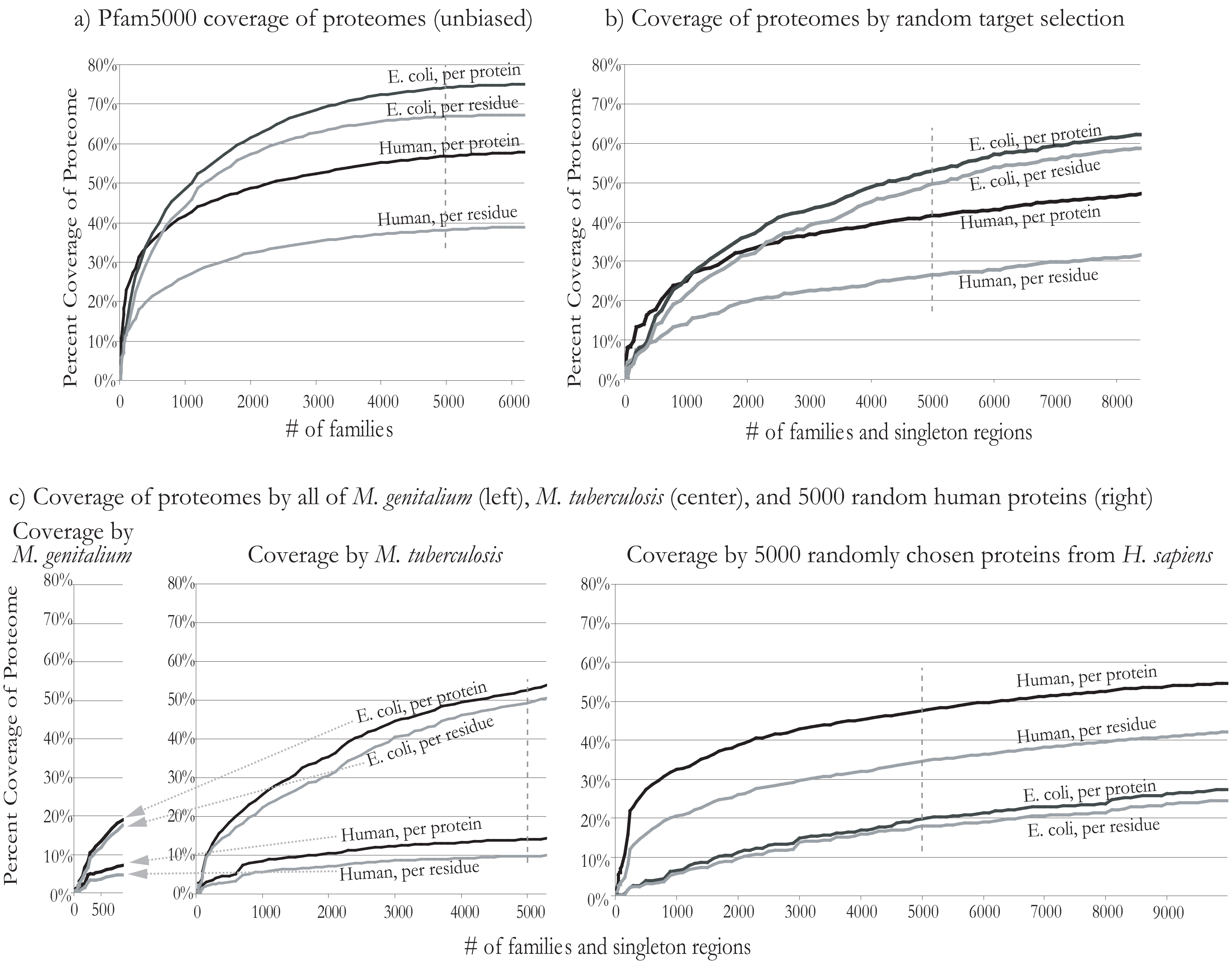
Chandonia & Brenner, Figure 7.



a) Pfam subfamilies and models (unbiased)

b) Pfam subfamilies and models (bias: structure)

Chandonia & Brenner, Figure 8.



a) Pfam5000 coverage of proteomes (unbiased)

b) Coverage of proteomes by random target selection

c) Coverage of proteomes by all of *M. genitalium* (left), *M. tuberculosis* (center), and 5000 random human proteins (right)