Title

NMR in structural genomics to increase structural coverage of the protein universe

Permalink

https://escholarship.org/uc/item/5sp7z583

Journal

The FEBS Journal, 283(21)

ISSN

1742-464X

Authors

Serrano, Pedro
Dutta, Samit K
Proudfoot, Andrew
et al.

Publication Date

2016-11-01

DOI

10.1111/febs.13751

Peer reviewed

NMR in structural genomics to increase structural coverage of the protein universe

Pedro Serrano[1,2], **Samit K. Dutta**[1,2,†], **Andrew Proudfoot**[1,2,§], **Biswaranjan Mohanty**[1,2,3,#], **Lukas Susac**[1,2], **Bryan Martin**[1,2], **Michael Geralt**[1,2], **Lukasz Jaroszewski**[1,4], **Adam Godzik**[1,4], **Marc Elsliger**[1,2], **Ian A. Wilson**[1,2,3], and **Kurt Wüthrich**[1,2,3,*]

[1] Joint Center for Structural Genomics (http://www.jcsg.org)

[2] Department of Integrative Structural and Computational Biology, The Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, CA 92037, USA

[3] Skaggs Institute for Chemical Biology, The Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, CA 92037, USA.

[4] Program on Bioinformatics and Systems Biology, Sanford Burnham Prebys Medical Discovery Institute, La Jolla, CA 92037, USA.

Abstract

For more than a decade, the Joint Center for Structural Genomics (JCSG; www.jcsg.org) worked toward increased three-dimensional structure coverage of the protein universe. This coordinated quest was one of the main goals of the four high-throughput (HT) structure determination centers of the Protein Structure Initiative (PSI; www.nigms.nih.gov/Research/specificareas/PSI). To achieve the goals of the PSI, the JCSG made use of the complementarity of structure determination by X-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy to increase and diversify the range of targets entering the HT structure determination pipeline. The overall strategy, for both techniques, was to determine atomic resolution structures for representatives of large protein families, as defined by the Pfam database, which had no structural coverage and could make significant contributions to biological and biomedical research. Furthermore, the experimental structures could be leveraged by homology modeling to further expand the structural coverage of the protein universe and increase biological insights. Here, we describe what could be achieved by this structural genomics approach, using as an illustration the contributions from 20 NMR structure determinations out of a total of 98 JCSG NMR structures, which were selected because they are the first three-dimensional structure representations of the

[*] **Correspondence** Kurt Wüthrich, Department of Integrative Structural and Computational Biology, The Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, CA, USA. Fax: +1.858.784.8014, Tel: +1.858.784.8011, wuthrich@scripps.edu.

[†] **Present address**: Sanford Burnham Prebys Medical Discovery Institute, 10901 North Torrey Pines Road, La Jolla, CA 92037, USA.

[§] **Present address**: Novartis Institute for BioMedical Research, 5300 Chiron Way, Emeryville, CA 94608, USA.

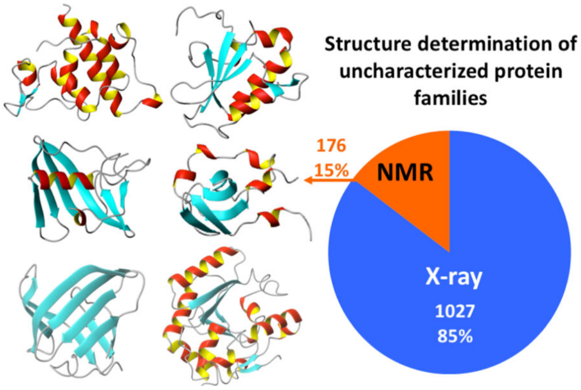
[#] **Present address**: Medicinal Chemistry, Monash Institute of Pharmaceutical Sciences. Monash University, 381 Royal Parade, Parkville, Victoria 3052, Australia.

Author contributions

PS, SKD, AP, BMo, LS, BMa, and MG prepared protein solutions for NMR structure determination, recorded and analyzed NMR data, and determined NMR structures. LJ and AG performed the bioinformatics analyses. PS, ME, IAW and KW designed the research strategy. PS and KW wrote the first draft of the paper, all the authors submitted criticisms and additions to the early manuscript versions, which were then included in the final manuscript by PS and KW.

respective Pfam protein families. The information from this small sample is representative for the overall results from crystal and NMR structure determination in the JCSG. There are five new folds, which were classified as domains of unknown functions (DUF), three of the proteins could be functionally annotated based on three-dimensional structure similarity with previously characterized proteins, and twelve proteins showed only limited similarly with previous deposits in the protein data bank (PDB) and were classified as DUFs.

Graphical Abstract



To increase structural coverage of the protein universe, the Joint Center for Structural Genomics (JCSG, www.jcsg.org) developed a strategy for exploiting complementarities between X-ray crystallography and NMR spectroscopy in solution. A sample of 20 NMR structure determinations is used to illustrate the target selection strategy and the different outcomes in terms of fold novelty and structure-based functional annotation.

Keywords

protein structures; structural coverage of Pfam protein families; domains of unknown function (DUF); Joint Center for Structural Genomics (JCSG); Protein Structure Initiative (PSI)

Introduction

The human genome project [1] and other genome sequencing efforts, focused either on individual species [2, 3] or on groups of microorganisms in specific environments (microbiomes) [4-7], had a huge impact on many different fields, including biomedical research and healthcare [6, 8, 9] , nutrition [10, 11] and agriculture [3, 12]. However, in most cases such advances are not expected to translate directly from the genomic sequences themselves, but rather to depend on acquiring detailed knowledge about newly uncovered gene products in the organisms of interest, i.e., primarily their proteomes. To capitalize on the genomics sequencing efforts, structural genomics was therefore initiated at the turn of the 21st century, with one of the main goals to increase the coverage of the genomic protein universe with three-dimensional structures of novel, previously not investigated gene products [13-15]. To meet the challenge presented by the continued and rapid growth of the pool of genomic sequences, the pilot phase (PSI-1; 2000 to 2005) and production phase (PSI-2; 2005 to 2010) of the NIGMS Protein Structure Initiative (PSI; https://

www.nigms.nih.gov/Research/specificareas/PSI/background/) focused on assembling and operating high-throughput (HT) structure determination pipelines, using X-ray diffraction in protein single crystals [16, 17] and nuclear magnetic resonance (NMR) spectroscopy with proteins in solution [18-27] (Fig. 1). For the target selection, criteria were implemented that supported investigations of a wide array of previously uncharacterized proteins ("domains of unknown function"; DUF) as candidates for structure determination [15-18, 28]. Today, the DUFs deposited in the Protein Data Bank (PDB) represent a large, untapped resource for investigating novel physiological processes involving the complete range of protein structures and functions in known organisms, as well as to address evolutionary aspects and differentiate one species from another. About 26% (4286 of 16,317) of the protein families defined at the time in the Pfam data base of protein families were used to guide the JCSG effort (the latest version used by the JCSG, Pfam 28.0, was released on May 20, 2015; see also the caption to Fig. 4) are DUFs, and about 20% (823) of these have at least one structurally characterized representative. More than 60% of these initial structural representatives were determined under the auspices of the PSI. As they explore uncharted regions of the protein universe [28], the DUFs represent valuable additions to the ever-increasing compendium of protein structures available as templates for molecular modeling, as a foundation for rational drug design, and for obtaining novel insights and supporting potential applications in many other fields [7-12].

NMR spectroscopy in solution has for many years been one of the principal techniques in the structural biology of proteins and nucleic acids [29]. It is unique in that atomic resolution structures and additional function-related data can be obtained at near-physiological solution conditions. One of the roles of NMR within the research program of the Joint Center of Structural Genomics (JCSG: www.jcsg.org) was to complement crystallography with protein structure determinations in solution, thus further contributing to increased structural coverage of the protein universe. NMR worked on targets which were prioritized by the JCSG Bioinformatics Core with special criteria for structure determination in solution, as well as on targets for which the HT pipeline did not yield a crystal structure (Fig. 1). In practice, for the sake of high efficiency, structural genomics avoided parallel efforts focused on the same target, so that NMR was used to large extent as the salvage method for crystallographic studies and the overlap between novel structures determined by crystallography and NMR was intentionally minimized. The XtalPred method [30] used for target identification was primarily developed to prioritize proteins for structure determination by crystallography. However, many protein families targeted by the JCSG did not contain proteins predicted by XtalPred to be optimal for crystallization. Many of these structures were then determined by NMR. The PSI had also imposed a criterion that targets for structure determination have less than 30% sequence identity with proteins that had previously been structurally characterized [15], and additional criteria were used to ensure high probability of success with a structure determination. For NMR targets this included an upper size limit of 200 amino acids.

To illustrate the structural genomics approach to increasing the structural coverage of the genomic protein universe, this paper discusses 20 of the 98 NMR structure determinations completed under the auspices of the JCSG, which were selected because they represented the first three-dimensional structures of their respective Pfam protein families. These numbers reflect an important detail of the strategy for target selection by the JSCG, which focused on increasing structural coverage of the protein universe by pursuing two tiers of targets. The first tier included targets with very low or no detectable sequence similarity to proteins with known structures. The structures resulting from this tier were most often the first structurally characterized members of their respective protein families. The second tier of targets ("fine-grained coverage targets") were selected from branches of large protein families whose general folds were already known (these targets still shared less than 30% conserved residues with proteins for which experimentally determined structures had been deposited in the PDB). In this study we focus on structures from the first tier.

Results and Discussion

NMR structures

The protein structures presented in this paper were determined with in-house standard solution conditions (50 mM NaCl, 5 mM NaN_3, 25 mM Na_2HPO_4 at pH 6.0; for proteins containing free cysteines, 2 mM deuterated DTT was added), using the J-UNIO protocol [27]. The experimental details of this approach have been described elsewhere [31-36], and statistics of the structure determinations discussed in this paper are given in Table 1. Stereo views of common representations for structures determined by NMR in solution are shown for the protein YP_399305.1 (PDB ID 2l1n) (Fig. 2). Panel A shows a bundle of 20 conformers superimposed for best fit to the mean atom coordinates; for each conformer, the polypeptide backbone is shown as a spline function through the α-carbon atoms. A tight fit of this bundle of conformers indicates high precision of the structure determination [29, 37]. Panel B shows an all-heavy-atom presentation of the conformer in the bundle for which the backbone was closest to the mean coordinates of the backbone heavy atoms. The color code identifies structurally well-defined amino acid side chains (blue), with a value for the displacement, D [38], of < 0.8 Å, with all other amino acids being in red. Compared to crystal structures of proteins, NMR has the advantage of being able to measure the intrinsic flexibility of proteins in solution at ambient temperature. In solution structures, the increase of dynamics when going from the protein core toward the solvent-exposed surface is typically more pronounced than in protein crystals [29]. Panel C shows a ribbon representation of the backbone for the conformer in B. For all 20 proteins discussed in this manuscript, corresponding ribbon presentations are shown in Fig. 3. Some of these proteins have previously been discussed in different contexts, where additional details of the structure determinations were given [39-42].

Structural coverage of the protein universe

To gain insights into evolutionary and structural relationships among protein sequences, they are typically grouped into families. Pfam is the most widely used database of conserved protein domain families and has been central to PSI efforts towards increasing structural coverage of the protein universe [15-18, 28]. The NMR structures in Fig. 3 represent the first structural representatives of 20 such Pfam families. At the time of the structure depositions, these families contained between 9438 and 73 members (Fig. 4; see also the comments on Pfam in the figure caption). These structures (Fig. 3) thus provide a platform for generating three-dimensional molecular models for over 32,000 protein sequences (Fig. 4) by homology modeling [43,44], which would yield a mean leverage of 671 for these experimentally determined protein structures.

An interesting aspect of the larger Pfam families (Fig. 4) is that they are typically represented in the genomes of a variety or important groups of different species. For example, PF03724 contains proteins encoded in archaea, bacteria and eukaryotes, PF11776 includes proteins from gram-positive and gram-negative bacteria, and PF06042 includes sequences found in bacteria and fungi. Proteins in smaller Pfam families, such as PF14466 and PF13642 (Figs. 3 and 4), tend to belong to organisms populating specific niches, such as the human gut microbiome [45] or deep-sea anaerobic habitats [46]. While characterizing larger families can yield higher homology modeling leverage per experimental structure determination, smaller families may represent recent adaptations to unique environments with novel biochemistry and function [47,48], potentially giving rise to important new biotechnological applications.

At the time of deposition, each of the 20 proteins in Fig. 3 shared less than 15% sequence identity with any previously deposited protein in the PDB. Interestingly, based on structure comparison with the algorithms DALI [49] and FATCAT [50], only the five proteins YP_926445.1, YP_399305.1, YP_001336205.1, YP_546394.1 and YP_001302112.1 (top row in Fig. 3) were at the time classified as new folds. Thus, a key conclusion is that many of these proteins with low sequence identity to known proteins adopt extensive variations of known folds, and are difficult to detect a *priori* by sequence-based algorithms. It is now a general trend that discovery of novel folds has greatly decreased over the years, indicating that, in 2015, the PDB represents a much more complete fold repertoire than when the PSI was started in 2000. Due to the aforementioned target selection criteria used by the PSI, this program has been the largest contributor to the discovery of new folds, as well as of extreme variants of known folds since the start of the 21st century [13]. Given the absence of similarities with functionally characterized proteins, proteins with new folds have typically come from structure determination of DUFs [51].

In addition to the discovery of new protein architectures [52-54], expanding the structural coverage of the genomic protein universe has revealed striking structural similarities between proteins with very low, and at times undetectable, sequence identity. The proteins YP_510488.1, NP_888769.1 and NP_344798.1 (three left-most entries in the second row of Fig. 3), which shared less than 15% sequence identity with structural counterparts in the PDB, illustrate this situation. Structure determination of these three proteins revealed close three-dimensional structure similarity with functionally annotated proteins in the PDB [39-42]. In these scenarios, without obtaining additional information on the individual proteins, we were able to provide functional annotations for the newly studied proteins and their respective Pfam families [39-42].

The remaining proteins in Fig. 3 were not recognized as new folds, but their similarity with known structures was limited, as it often extended over only small polypeptide segments of larger proteins. For these 12 proteins, attempts at functional annotation remained ambiguous, so that they were classified as DUFs [51]. Quite generally, it turned out that the functional annotation of proteins based on their three-dimensional structure alone is difficult, especially when dealing with little explored regions of the protein universe, which represents the majority of the proteins targeted by the PSI selection criteria [13, 15]. Nonetheless, structure determination can severely limit the array of possible functions of a particular protein and thus aid in the discovery of the actual biological role.

Conclusions

NMR has contributed about 10% of the structures deposited in the PDB by PSI-associated research teams (Fig. 5A). Although the number of structures is much smaller than what was achieved with X-ray crystallography, the data presented in this paper provide a representative illustration of how the PSI NMR effort, with its much more limited resources, has been successfully integrated into the structural genomics approach to increase the three-dimensional coverage of the protein universe. The homology modeling leverage of each structure solved is of course tied to the selection of the protein structures considered, such as those shown in Fig. 3. Nonetheless, while this leverage from homology building varies for different samples of PSI structures determined either by NMR or by X-ray crystallography, the data in Fig. 4 are representative of the average leverage for structures determined in PSI-1 and PSI-2 (http://sbkb.org/metrics).

When comparing the relative contributions from X-ray crystallography and NMR in solution either to the PSI deposits or all deposits in the PDB or to first representations of new Pfam families (Fig. 5, A and B), it is apparent that NMR has contributed more than its share of first structural representatives. The complementarity between X-ray crystallography and NMR in solution in providing first representative structures for new Pfam families is also readily apparent from the increased contributions of NMR structures for proteins that have been classified as difficult to handle with crystallographic methods (Fig. 5C). Moreover, most of the remaining Pfam families with no structural coverage are predicted to be “difficult” for characterization by X-ray crystallography, mostly because of predicted structural disorder, indicating that many potential “NMR-accessible” targets remain to be explored. Overall, these data indicate that continued efforts by NMR in solution should be encouraged in the interest of increasing structural coverage of new Pfam families, particularly where they include proteins with sizes up to about 250 residues.

For the coming years, it is an interesting task to develop methods for efficient functional annotation of the resource of DUFs deposited in the PDB. Success in such efforts would be an ultimate validation of the systematic PSI approach to increase structural coverage of the genomic protein universe and enhance its impact on biological and biomedical research.

Acknowledgements

This work was supported by the Joint Center for Structural Genomics through the NIH Protein Structure Initiative (PSI) grant U54-GM094586 from the National Institute of General Medical Sciences (www.nigms.nih.gov). Kurt Wüthrich is the Cecil H. and Ida M. Green Professor of Structural Biology at TSRI. We thank the JCSG Crystallomics Core for providing us with clones of the presently discussed proteins.

Abbreviations

JCSG	Joint Center for Structural Genomics

PSI	Protein Structure Initiative
NMR	nuclear magnetic resonance
HT	high-throughput
Pfam	a data base for identifying and classifying protein families
DUF	domain of unknown function
PDB	protein data bank

References

1. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, et al. The sequence of the human genome. Science. 2001; 291:1304–1351. [PubMed: 11181995]
2. Vodovar N, Vallenet D, Cruveiller S, Rouy Z, Barbe V, et al. Complete genome sequence of the entomopathogenic and metabolically versatile soil bacterium Pseudomonas entomophila. Nat Biotech. 2006; 24:673–679.
3. Mace ES, Tai S, Gilding EK, Li Y, Prentis PJ, et al. Whole-genome sequencing reveals untapped genetic potential in Africa's indigenous cereal crop sorghum. Nat Commun. 2013; 4 doi:10.1038/ncomms3320.
4. Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, et al. Environmental Genome Shotgun Sequencing of the Sargasso Sea. Science. 2004; 304:66–74. [PubMed: 15001713]
5. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, et al. Genome sequencing in microfabricated high-density picolitre reactors. Nature. 2005; 437:376–380. [PubMed: 16056220]
6. Turnbaugh JP, Ley RE, Hamady M, Fraser-Liggett C, Knight R, Gordon JI. The human microbiome project: exploring the microbial part of ourselves in a changing world. Nature. 2007; 449:804–810. [PubMed: 17943116]
7. Ellrott K, Jaroszewski L, Li W, Wooley JC, Godzik A. Expansion of the protein repertoire in newly explored environments: human gut microbiome specific protein families. PLOS Comp Biol. 2010 DOI: 10.1371/journal.pcbi.1000798.
8. Green ED, Guyer MS, National human genome research institute. Charting a course for genomic medicine from base pairs to bedside. Nature. 2011; 470:204–213. [PubMed: 21307933]
9. Weigelt J. Structural genomics—Impact on biomedicine and drug discovery. Exp Cell Res. 2010; 316:1332–1338. [PubMed: 20211166]
10. Kau AL, Ahern PP, Griffin NW, Goodman AL, Gordon JI. Human nutrition, the gut microbiome and the immune system. Nature. 2011; 474:327–336. [PubMed: 21677749]
11. Mutch DM, Wahli W, Williamson G. Nutrigenomics and nutrigenetics: the emerging faces of nutrition. FASEB J. 2005; 19:1602–1616. [PubMed: 16195369]
12. Chia JM, Ware D. Sequencing for the cream of the crop. Nat Biotech. 2011; 29:138–139.
13. Khafizova K, Madrid-Aliste C, Almo SC, Fiser A. Trends in structural coverage of the protein universe and the impact of the protein structure initiative. Proc Natl Acad Sci USA. 2014; 111:3733–3738. [PubMed: 24567391]
14. Levitt M. Nature of the protein universe. Proc Natl Acad Sci USA. 2009; 106:11079–11084. [PubMed: 19541617]
15. Dessailly BH, Nair R, Jaroszewski L, Fajardo JE, Kouranov A, et al. PSI-2: structural genomics to cover protein domain family space. Structure. 2009; 17:869–881. [PubMed: 19523904]
16. Elsliger MA, Deacon AM, Godzik A, Lesley SA, Wooley J, Wüthrich K, Wilson IA. The JCSG high-throughput structural biology pipeline. Acta Crystal F. 2010; 66:1137–1142.
17. Lesley SA, Kuhn P, Godzik A, Deacon AM, Mathews I, et al. Structural genomics of the Thermotoga maritima proteome implemented in a high-throughput structure determination pipeline. Proc Natl Acad Sci USA. 2002; 99:11664–11669. [PubMed: 12193646]

18. Page R, Peti W, Wilson IA, Stevens RC, Wüthrich K. NMR screening and crystal quality of bacterially expressed prokaryotic and eukaryotic proteins in a structural genomics pipeline. Proc Nat Acad Sci USA. 2005; 102:1901–1905. [PubMed: 15677718]

19. Peti W, Etezady-Esfarjani T, Herrmann T, Klock HE, Lesley SA, Wüthrich K. NMR for structural proteomics of Thermotoga maritima: screening and structure determination. J Struct Funct Genomics. 2004; 5:205–215. [PubMed: 15263836]

20. Peti W, Page R, Moy K, O'Neil-Johnson M, Wilson IA, Stevens RC, Wüthrich K. Towards miniaturization of a structural genomics pipeline using mocro-expression and microcoil NMR. J Struct Funct Genomics. 2005; 6:259–267. [PubMed: 16283429]

21. Rosato A, Aramini JM, Arrowsmith C, Bagaria A, Baker D, et al. Blind testing of routine, fully automated determination of protein structures from NMR data. Structure. 2012; 2:227–236.

22. Montelione GT, Yuanpeng DZ, Huang J, Gunsalus KC, Szyperski T. Protein NMR spectroscopy in structural genomics. Nat Struct Biol. 2000; 7:982–985. [PubMed: 11104006]

23. Yee A, Chang X, Pineda-Lucena A, Wu B, Semesi A, et al. An NMR approach to structural proteomics. Proc Natl Acad Sci USA. 2001; 99:1825–1830.

24. Xiao R, Anderson S, Aramini J, Belote R, Buchwald WA, et al. The high-throughput protein sample production platform of the northeast structural genomics consortium. J Struct Biol. 2010; 172:21–33. [PubMed: 20688167]

25. Dutta S, Serrano P, Proudfoot A, Geralt M, Pedrini B, Herrmann T, Wüthrich K. APSY-NMR for protein backbone assignment in high-throughput structural biology. J Biomol NMR. 2015; 61:47–53. [PubMed: 25428764]

26. Pedrini B, Serrano P, Mohanty B, Geralt M, Wüthrich K. NMR-profiles of protein solutions. Biopolymers. 2013; 99:825–831. [PubMed: 23839514]

27. Serrano P, Pedrini B, Mohanty B, Geralt M, Herrmann T, Wüthrich K. The J-UNIO protocol for automated protein structure determination by NMR in solution. J Biomol NMR. 2012; 53:341–354. [PubMed: 22752932]

28. Jaroszewski L, Li Z, Sri Krishna S, Bakolitsa C, Wooley J, et al. Exploration of uncharted regions of the protein universe. Plos Biol. 2009; 7:e10000205.

29. Wüthrich, K. NMR of Proteins and Nucleic Acids. Wiley; New York: 1986.

30. Slabinski L, Jaroszewski L, Rychlewski L, Wilson IA, Lesley S, Godzik A. XtalPred: a web server for prediction of protein crystallizability. Bioinformatics. 2007; 23:3403–3405. [PubMed: 17921170]

31. Volk J, Herrmann T, Wüthrich K. Automated sequence-specific protein NMR assignment using the memetic algorithm MATCH. J Biomol NMR. 2008; 41:127–138. [PubMed: 18512031]

32. Fiorito F, Herrmann T, Damberger FF, Wüthrich K. Automated amino acid side-chain NMR assignment of proteins using 13C- and 15N-resolved 3D [1H,1H]-NOESY. J Biomol NMR. 2008; 42:23–33. [PubMed: 18709333]

33. Hiller S, Fiorito F, Wüthrich K, Wider G. Automated projection spectroscopy (APSY). Proc Nat Acad Sci USA. 2005; 102:10876–10881. [PubMed: 16043707]

34. Herrmann T, Güntert P, Wüthrich K. Protein NMR structure determination with automated NOE-identification in the NOESY spectra using the new software ATNOS. J Biomol NMR. 2002; 24:171–189. [PubMed: 12522306]

35. Herrmann T, Güntert P, Wüthrich K. Protein NMR structure determination with automated NOE assignment using the new software CANDID and the torsion angle dynamics algorithm DYANA. J Mol Biol. 2002; 319:209–227. [PubMed: 12051947]

36. Güntert P, Mumenthaler C, Wüthrich K. Torsion angle dynamics for NMR structure calculation with the new program DYANA. J Mol Biol. 1997; 273:283–298. [PubMed: 9367762]

37. Wüthrich K. NMR studies of structure and function of biological macromolecules. Angew Chem. 2003; 42:3340–3363. [PubMed: 12888958]

38. Billeter M, Kline A, Braun W, Huber R, Wüthrich K. Comparison of the high-resolution structures of the a-amylase inhibitor Tendamistat determined by nuclear magnetic resonance in solution and by X-ray diffraction in single crystals. J Mol Biol. 1989; 206:677–687. [PubMed: 2786963]

39. tul-Wahab A, Serrano P, Geralt M, Wüthrich K. NMR structure of the Bordetella bronchiseptica protein NP_888769.1 establishes a new phage-related protein family PF13554. Prot Sci. 2011; 20:1137–1144.

40. Serrano P, Geralt M, Mohanty B, Wüthrich K. NMR structures of the α-proteobacterial ATPase-regulating ζ-subunits. J Mol Biol. 2014; 15:2547–2453.

41. Mohanty B, Serrano P, Geralt M, Wüthrich K. NMR structure determination of the protein NP_344798.1 as the first representative of Pfam PF06042. J Biomol NMR. 2014; 61:83–87. [PubMed: 25430057]

42. Mohanty B, Geralt M, Wüthrich K, Serrano P. NMR reveals structural rearrangements associated to substrate insertion in nucleotide-adding enzymes. Prot Sci. 2016 DOI: 10.1002/pro.2872.

43. Yura K, Yamaguchi A, Go M. Coverage of whole proteome by structural genomics observed through protein homology modeling database. J Struct Funct Genomics. 2006; 7:65–76. [PubMed: 17146617]

44. Bordoli L, Kiefer F, Arnold K, Benkert P, Battey J, Schwede T. Protein structure homology modeling using SWISS-MODEL workspace. Nature Protocols. 2009; 4:1–13. [PubMed: 19131951]

45. Ellrott K, Jaroszewski L, Li W, Wooley JC, Godzik A. Expansion of the protein repertoire in newly explored environments: human gut microbiome specific protein families. PLoS Comput Biol. 2010; 6:e1000798. [PubMed: 20532204]

46. Sogin M, Morrison H, Huber J, Welch D, Huse S, et al. Microbial diversity in the deep sea and the underexplored “rare biosphere”. Proc Natl Acad Sci USA. 2006; 103:12115–12120. [PubMed: 16880384]

47. Harron M, Hu SH, Shi Y, Imelfort M, Keller J, et al. Anaerobic oxidation of methane coupled to nitrate reduction in a novel archaeal lineage. Nature. 2013; 500:567–570. [PubMed: 23892779]

48. Dekas A, Poretsky R, Orphan V. Deep-sea archaea fix and share nitrogen in methane-consuming microbial consortia. Science. 2009; 326:422–426. [PubMed: 19833965]

49. Holm L, Sander C. Dali: a network tool for protein structure comparison. Trends Biochem Sci. 1995; 20:478–480. [PubMed: 8578593]

50. Ye Y, Godzik A. Flexible structure alignment by chaining aligned fragment pairs allowing twists. Bioinformatics. 2003; 19:ii246–ii255. [PubMed: 14534198]

51. Bateman A, Coggill P, Finn RD. DUFs: families in search of function. Acta Cryst F. 2010; 66:1148–1152.

52. Fernandez-Fuentes N, Dybas J, Fiser A. Structural characteristics of novel protein folds. PLoS Comput Biol. 2010; 6:e1000750. [PubMed: 20421995]

53. Andreeva A, Murzin AG. Structural classification of proteins and structural genomics: new insights into protein folding and evolution. Acta Cryst F. 2010; F66:1190–1197.

54. Kolodny R, Pereyaslavets L, Samson A, Levitt M. On the universe of protein folds. Biophysics. 2013; 42:559–582.

55. Finn R, Coggill P, Eberhardt R, Eddy S, Mistry J, et al. The Pfam protein families database: towards a more sustainable future. Nucl Acids Res. 2016; 44:D279–D285. [PubMed: 26673716]

56. Wüthrich K. NMR in a crystallography-based high-throughput protein structure-determination environment. Acta Cryst F. 2010; 66:1365–1366.

57. Jaudzems K, Geralt M, Serrano P, Mohanty B, Horst R, Pedrini B, Elsliger M-A, Wilson IA, Wüthrich K. NMR structure of the protein NP_247299.1: comparison with the crystal structure. Acta Cryst F. 2010; 66:1367–1380.

58. Mohanty B, Serrano P, Pedrini B, Jaudzems K, Geralt M, Horst R, Herrmann T, Elsliger M-A, Wilson IA, Wüthrich K. Comparison of NMR and crystal structures for the proteins TM1112 and TM1367. Acta Cryst. F. 2010; 66:1381–1392.

59. Serrano P, Pedrini B, Geralt M, Jaudzems K, Mohanty B, Horst R, Herrmann T, Elsliger M-A, Wilson IA, Wüthrich K. Comparison of NMR and crystal structures highlights conformational isomerism in protein active sites. Acta Cryst. F. 2010; 66:1393–1405.

60. Braun W, Epp O, Wüthrich K, Huber R. Solution of the phase problem in the X-ray diffraction method for proteins with the nuclear magnetic resonance solution structure as initial model. J Mol Biol. 1989; 206:669–676. [PubMed: 2786962]

61. Laskowski RA, Macarthur MW, Moss DS, Thornton JM. PROCHECK - A program to check the stereochemical quality of protein structures. J Appl Cryst. 1993; 26:283–291.

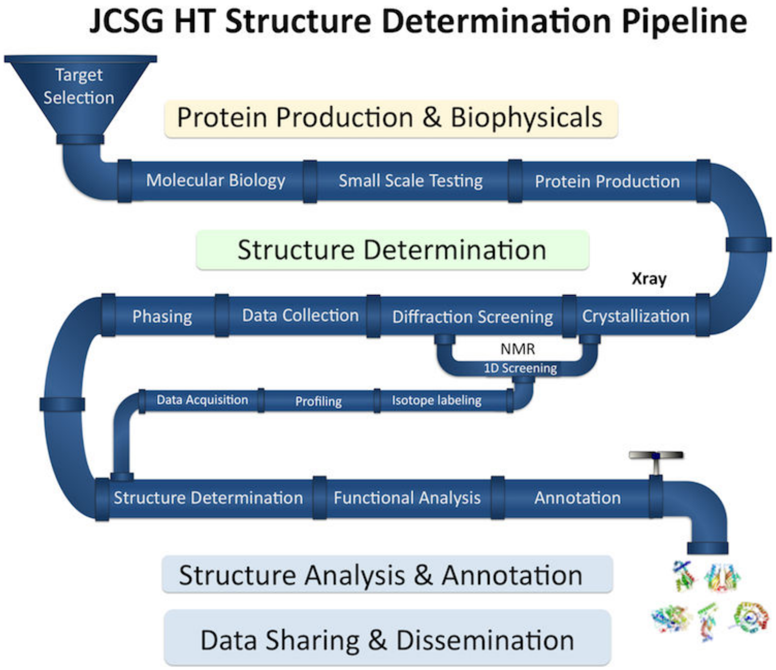


Fig. 1.

JSCG high-throughput (HT) structure determination pipeline. Targets for NMR structure determination included primarily proteins for which the HT crystallography pipeline did not yield a structure but additional proteins were prioritized for NMR structure determination (reproduced from [16]).

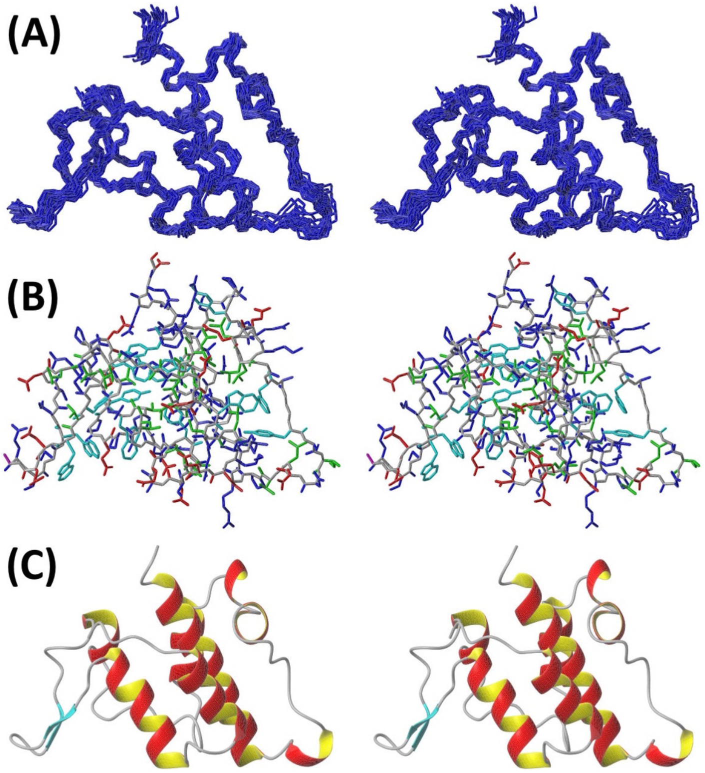


Fig. 2.

NMR structure of the protein YP_399305.1 (PDB ID 2l1n). (A) Bundle of 20 conformers selected to represent the NMR structure, superimposed for minimal RMSD to the mean atom coordinates. Only the polypeptide backbone is shown. For the selection of the 20 conformers, see [27] and Table 1. (B) All-heavy-atom presentation of the conformer from (A), for which the polypeptide backbone is closest to the mean atom coordinates of the bundle. The color code represents the displacements, D [37], calculated for the amino acid side chains, where small values indicate that the residue is structurally well defined; blue, D < 0.8 Å; red, D ≥ 0.8 Å; the backbone and the prolines are gray. (C) Ribbon presentation of the backbone in the structure of panel B.

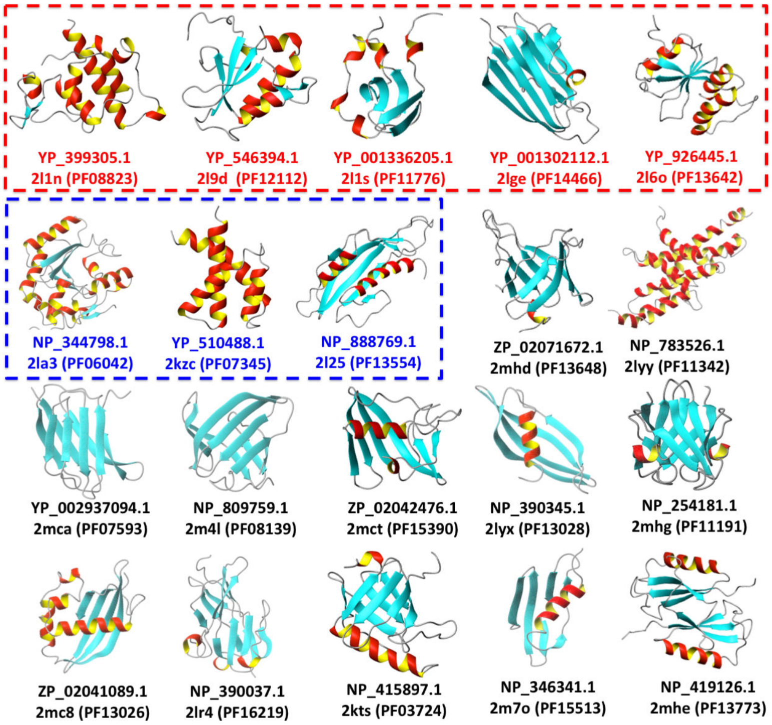


Fig. 3.

20 first representatives of Pfam protein families determined by NMR. For each protein, a ribbon presentation is shown of the conformer closest to the mean coordinates of a bundle of 20 conformers used to represent the NMR structure (see Fig. 2). Below each structure, the NCBI accession name of the protein, the Protein Data Bank identification number and, in parentheses, the Pfam family code, are given. The top row (red box) shows five proteins for which a new fold was discovered. The first three entries in the second row (blue box) were functionally annotated based on near-coincidence of their three-dimensional structures with PDB entries of proteins with known functions.

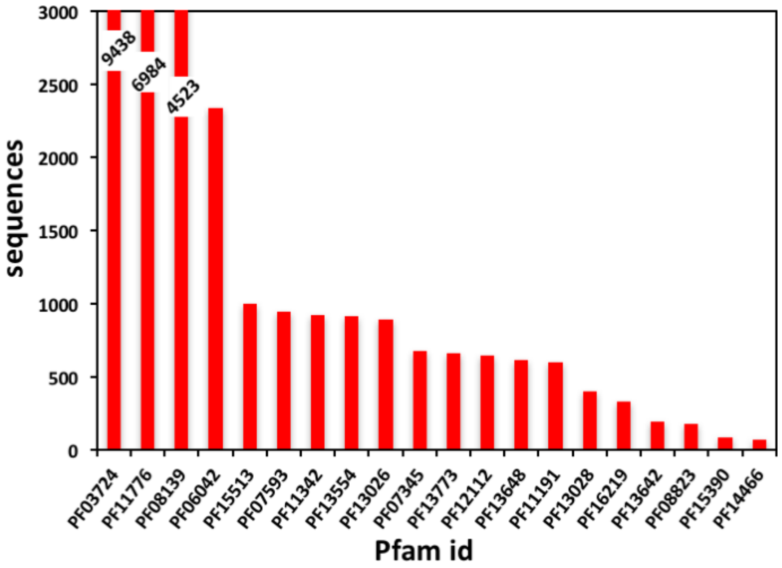


Fig. 4.

Size distribution of the 20 Pfam families covered in Fig. 3. A histogram shows the number of protein sequences belonging to the individual Pfams. For the largest families, the off-scale size is indicated (family size is based on the Pfam release 28.0 of May 20, 2015). After the end of the PSI on June 30, 2015, the Pfam protein families database has been fundamentally revised, as described by Finn et al. [54], so that the numbers shown in this figure do no longer have counterparts in the current Pfam database. For the present discussion it is important that Fig. 4 represents the number of genomic protein sequences which are accessible to structural coverage by homology modeling.

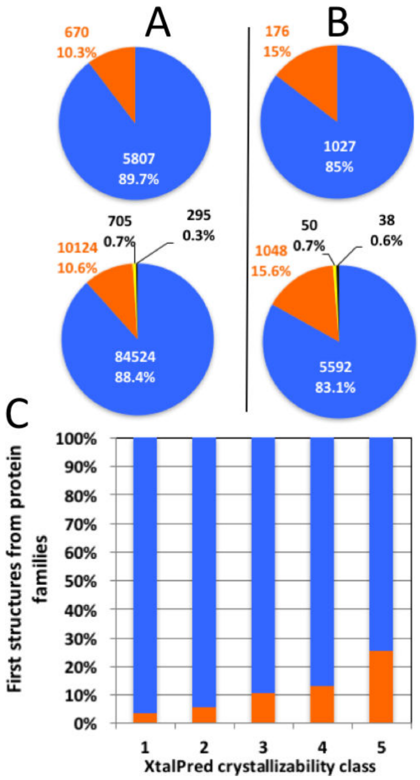


Fig. 5.

Contributions from X-ray crystallography (blue), NMR in solution (orange), electron microscopy (yellow), and other methods (black) to protein structure determinations. (A) Top: All PSI deposits in the PDB; bottom: all PDB structures. (B) Top: PSI deposits representing first structures of Pfam families; bottom. PDB deposits representing first structures of Pfam families. (C) Contributions by X-ray crystallography and NMR to the five XtalPred classes of proteins; increasing difficulty for X-ray structure determination is predicted when going from classes 1 to 5 [30]. We also want to briefly comment on requests by the referees which otherwise are beyond the scope of this article. Firstly, it was not a primary aim of the JCSG to compare structures of the same proteins determined by X-ray crystallography and by NMR. To the contrary, in the interest of efficiency, overlap between crystal and NMR structure determination was minimized. Nonetheless, such comparisons have been investigated for a small sample of proteins [56-59]. Secondly, the use of NMR structures for solving crystal structures by molecular replacement has long been established [38,60] In the context of follow-up studies, for example for complexes of DUF structures with macromolecular partners, this approach may be of special interest, considering that the PSI tapped new proteins in otherwise sparsely explored parts of the genomic protein universe [13-15].

Table 1

Input for the structure calculations and characterization of the ensembles of 20 energy-minimized CYANA conformers (Fig. 2A) used to represent the NMR structures presented in Fig. 3.

Protein[a]	Value[b]									
Quantity	**YP_399305.1 (2L1N, 120)**	**YP_546394.1 (2L9D, 108)**	**YP_001336205.1 (2L1S 83)**	**YP_001302112.1 (2LGE, 129)**	**YP_926445.1 (2L6O, 114)**	**NP_344798.1 (2LA3, 191)**	**YP_510488.1 (2KZC, 85)**	**NP_888769.1 (2L25, 141)**	**ZP_02071672.1 (2MHD, 110)**	**NP_783526 (2LYY, 96)**
NOE upper distance										
limits	2331	2416	1736	1850	1910	4090	1507	1972	1733	3686
intraresidual	604	535	330	500	507	993	403	526	469	934
short-range	611	673	439	552	506	1055	347	605	493	855
medium-range	640	547	361	183	343	904	462	275	168	891
long-range	476	661	606	605	554	1138	295	566	603	1006
Dihedral angle										
constraints	476	438	348	528	462	799	310	559	560	1974
Residual target function value [Å^2]	1.87 ± 0.25	1.64 ± 0.16	1.17 ± 0.23	1.68 ± 0.26	1.52 ± 0.24	3.9 ± 0.43	1.05 ± 0.19	1.95 ± 0.37	3.02 ± 0.76	6.64 ± 0.63
Residual NOE violations										
number ≥ 0.1 Å	26 ± 3	16 ± 4	12 ± 4	12 ± 4	13 ± 5	41 ± 6	11 ± 4	24 ± 5	19 ± 5	40 ± 6
maximum [Å]	0.14 ± 0.02	0.15 ± 0.09	0.16 ± 0.06	0.11 ± 0.08	0.15 ± 0.09	0.20 ± 0.19	0.14 ± 0.02	0.15 ± 0.04	0.18 ± 0.12	0.17 ± 0.09
Residual dihedral angle violations										
number ≥ 2.5°	1 ± 1	1 ± 1	0 ± 1	1 ± 1	1 ± 1	1 ± 1	1 ± 1	2 ± 1	3 ± 2	2 ± 1
maximum [°]	3.72 ± 0.71	1.65 ± 0.8	1.67± 1.23	3.0 ± 0.5	2.98 ±1.09	2.79 ± 1.5	1.7 ± 0.96	3.14 ± 1.35	0.29 ± 0.07	1.89 ± 0.2
AMBER energies [kcal/mol]										
total	−4666 ± 89	−4097 ± 82	−4264 ± 44	−4607 ± 152	−4231 ± 130	−7843 ± 144	−3373 ± 79	−5064 ± 97	−4048 ± 365	−7343 ± 132
van der Waals	−413 ± 13	−306 ± 19	−276 ± 9	−322 ± 36	−401 ± 79	−705 ± 35	−271 ± 14	−375 ± 18	−294 ± 76	−831 ± 23
electrostatic	−5145 ± 88	−4602 ± 68	−3603 ± 37	−5238 ± 176	−4791 ± 85	−8779 ± 112	−3678 ± 84	−5836 ± 90	−4762 ± 172	−7981 ± 138
RMSD from mean coordinates [Å][c]										
backbone	0.53 ± 0.06	0.51± 0.09	0.38 ± 0.10	0.59 ± 0.11	0.62 ± 0.08	0.53 ± 0.09	0.46 ± 0.08	0.71 ± 0.09	0.73 ± 0.21	0.73 ± 0.10
all heavy atoms	0.91 ± 0.09	0.96 ± 0.08	0.74 ± 0.09	1.10 ± 0.12	1.09 ± 0.07	0.91 ± 0.09	0.87 ± 0.08	1.18 ± 0.11	1.21 ± 0.20	1.07 ± 0.10
Ramachandran plot statistics[d]										
most favored regions [%]	80.6	80.5	76.9	71.7	67.5	76.7	87.1	77.4	81.4	91.0
additional allowed regions [%]	18.1	17.2	21.3	27.6	29.5	21.4	10.9	19.8	12.5	8.4
generously allowed regions [%]	1.1	1.5	1.7	0.5	1.6	1.6	1.3	1.5	5.1	0.4

Protein[a]	Value[b]									
Quantity	**YP_399305.1 (2L1N, 120)**	**YP_546394.1 (2L9D, 108)**	**YP_001336205.1 (2L1S 83)**	**YP_001302112.1 (2LGE, 129)**	**YP_926445.1 (2L6O, 114)**	**NP_344798.1 (2LA3, 191)**	**YP_510488.1 (2KZC, 85)**	**NP_888769.1 (2L25, 141)**	**ZP_02071672.1 (2MHD, 110)**	**NP_783526 (2LYY, 96)**
disallowed regions (%)	0.2	0.8	0.1	0.2	1.4	0.3	0.7	1.3	1.0	0.2
NOE upper distance										
limits	2498	1966	1969	1528	2729	2091	2563	2276	1568	2605
intraresidual	427	453	503	378	674	566	616	510	333	686
short-range	636	508	537	427	790	527	629	643	355	736
medium-range	263	202	297	263	416	367	329	872	268	380
long-range	890	803	632	460	849	631	989	476	309	803
Dihedral angle										
constraints	282	542	524	389	1260	605	614	492	303	1344
Residual target function value [Å^2]	1.48 ± 0.11	1.58 ± 0.59	2.2 ± 0.41	1.04 ± 0.10	1.98 ± 0.23	3.21 ± 0.99	2.27 ± 0.38	1.93 ± 0.25	1.11 ± 0.14	2.20 ± 0.21
Residual NOE violations										
number ≥ 0.1 Å	20 ± 4	14 ± 4	7 ± 2	11 ± 3	142 ± 11	21 ± 1	27 ± 6	20 ± 4	8 ± 3	13 ± 4
maximum [Å]	0.14 ± 0.01	0.16 ± 0.06	0.13 ± 0.02	0.16 ± 0.06		0.10 ± 0.06	0.17 ± 0.11	0.15 ± 0.04	0.16 ± 0.08	0.13 ± 0.03
Residual dihedral angle violations										
number ≥ 2.5°	0 ± 0	0 ± 0	2 ± 1	0 ± 0	0 ± 0	0 ± 0	2 ± 2	0 ± 1	0 ± 1	1 ± 1
maximum [°]	2.32± 1.53	2.20 ± 1.15	2.39 ± 0.61	2.16 ± 0.73	0.3 ± 0.04	0.12 ± 0.05	2.66 ± 1.65	2.62 ± 0.90	1.77 ± 1.58	2.1 ± 0.74
AMBER energies [kcal/mol]										
total	−3763 ± 92	−3313 ± 402	−3153 ± 91	−3271 ± 73	−5154 ± 121	−3793 ± 125	−4857 ± 99	−4774 ± 31	−2087 ± 57	−5102 ± 99
van der Waals	−298 ± 15	−221 ± 192	−325 ± 15	−260 ± 11	−697 ± 20	−408 ± 18	−425 ± 16	−381 ± 15	−178 ± 9	−573 ± 31
electrostatic	−4281 ± 93	−3838 ± 192	−3611 ± 87	−3745 ± 65	−5465 ± 120	−4749 ± 101	−5440 ± 83	−5349 ± 94	−2455 ± 61	−5565 ± 127
RMSD from mean coordinates [Å][c]										
backbone	0.38 ± 0.10	0.59 ± 0.12	0.47 ± 0.06	0.58 ± 0.10	0.81 ± 0.25	0.71 ± 0.14	0.68 ± 0.11	0.62 ± 0.08	0.64 ± 0.14	0.60 ± 0.09
all heavy atoms	0.74 ± 0.09	1.15 ± 0.15	0.93 ± 0.08	1.09 ± 0.10	1.27 ± 0.24	1.33 ± 0.16	1.24 ± 0.14	1.09 ± 0.07	1.06 ± 0.17	0.88 ± 0.12
Ramachandran plot statistics[d]										
most favored regions [%]	73.3	78.7	82.0	73.8	68.9	82.6	71.7	67.5	76.6	73.9
additional allowed regions [%]	24.4	19.7	15.6	26.2	29.5	17.1	27.2	29.5	21.9	24.3
generously allowed regions [%]	1.1	0.4	2.4	0.0	0.3	0.2	0.1	1.6	0.0	1.1
disallowed regions (%)	1.1	1.2	0.0	0.0	1.3	0.0	1.0	1.4	1.6	0.7

[a] The ncbi accession names of the proteins are indicated, and in parentheses the PDB_id and the size in number of amino acids are given. The first five entries are proteins for which a novel fold was observed, and entries 6 to 8 are proteins which were functionally annotated based on 3D structure similarity with previously characterized proteins (see Fig. 3 and the text).

[b] Except for the top six entries, average values and standard deviations for the bundle of 20 energy-minimized conformers (see Fig. 2A) are given.

[c] The residues for which the RMSDs were calculated are: NP_415897.1 (3–117), YP_399305.1 (4–88, 100–117), YP_001336205.1 (5–37, 45–81), YP_510488.1 (5–40, 47–82), YP_926445.1 (6–39, 45–91, 98–114), NP_390345.1 (12–87), NP_346341.1 (6–70), SP19397A (12–87), NP_390037.1 (6–127), NP_809759.1 (4–96), ZP_02042476.1 (1–102), NP_344798.1 (5–99, 114–172, 178–191), YP_546394.1 (11–40, 47–107), NP_814968.1(10–124, 134–142), NP_687636.1 (3–92), NP_888769.1 (3–50,66–136), ZP_02071672.1 (7–44,52–60,65–108), and YP_001302112.1 (10–79,92–112,125–129).

[d] As determined by *PROCHECK* (61)