

Assisting Functional Assignment for Hypothetical *Haemophilus influenzae* Gene Products through Structural Genomics

Gary L. Gilliland ^{*†}, Alexey Teplyakov[†], Galina Obmolova [†], Maria Tordova [†], Narmada Thanki^{†,‡}, Jane Ladner[†], Osnat Herzberg[†], Kap Lim[†], Hong Zhang[†], Kui Huang[†], Zhong Li[†], Aleksandra Tempczyk [†], Wojciech Krajewski[†], Lisa Parsons [†], Deok Cheon Yeh[†], John Orban[†], Andrew J. Howard[§], Edward Eisenstein^{†,¶}, James F. Parsons [†], Nicklas Bonander^{†,£}, Kathryn E. Fisher[†], John Toedt^{†,€}, Prasad Reddy[¥], C. V. Rao[¥], Eugene Melamud [†] and John Moulton [†]

[†]Center for Advanced Research in Biotechnology of the University of Maryland Biotechnology Institute and the National Institute of Standards and Technology, 9600 Gudelsky Drive, Rockville, MD 20850, USA

[‡]Current Address: Laboratory of Computational Technologies, NCI-Frederick Cancer Research and Development Center, Building 426, Room 158, P.O. Box B, Frederick, MD 21702, USA

[§]Biological, Chemical, and Physical Sciences Department, Illinois Institute of Technology, Chicago, IL 60616, USA

[¶]Department of Chemistry and Biochemistry, University of Maryland Baltimore County, Baltimore, MD, 21228, USA

[£]Current Address: Center for Structural Biology & Department of Molecular Biotechnology, Chalmers University of Technology, Box 462, SE-405 30, Göteborg, Sweden

[€]Current Address: Department of Physical Science, Eastern Connecticut State University, 83 Windham Street, Willimantic, Connecticut 06226-2295, USA

[¥]Biotechnology Division of the National Institute of Standards and Technology, Bureau Drive, Gaithersburg, MD 20899, USA

Abstract: The three-dimensional structures of *Haemophilus influenzae* proteins whose biological functions are unknown are being determined as part of a structural genomics project to ask whether structural information can assist in assigning the functions of proteins. The structures of the hypothetical proteins are being used to guide further studies and narrow the field of such studies for ultimately determining protein function. An outline of the structural genomics methodological approach is provided along with summaries of a number of completed and in progress crystallographic and NMR structure determinations. With more than twenty-five structures determined at this point and with many more in various stages of completion, the results are encouraging in that some level of functional understanding can be deduced from experimentally solved structures. In addition to aiding in functional assignment, this effort is identifying a number of possible new targets for drug development.

INTRODUCTION

In 1995 the first complete genome sequence of a free-living organism, *Haemophilus influenzae* was determined [1]. The initial sequence analysis identified 1703 (originally this number was 1743) proteins for which approximately 65% were assigned specific functions by homology to other known sequences, while the functions of a further 10-20% of the proteins were only generally described [2,3]. Proteins of

the genome lacking functional assignment after annotation were termed *hypothetical*. Numerous approaches have been devised to obtain functional information for hypothetical gene products. One method for functional annotation that has shown potential is through the interpretation and analyses of the three-dimensional structures of hypothetical proteins. Several reports in the literature provide examples of how fold recognition has been used to assign a probable function to a protein (e.g., the studies of Colovos and co-workers [4] and those of Hwang and co-workers [5]), which in one case was later confirmed experimentally [5]. Often, the structure reveals a substrate analogue or cofactors bound to the protein, reducing the number of possible functions to those involved with the particular compound (e.g., Zarembinski and co-workers [6]). With this in mind, a structural genomics project was undertaken in which the

*Address correspondence to this author at the Research in Biotechnology of the University of Maryland Biotechnology Institute and the National Institute of Standards and Technology, 9600 Gudelsky Drive, Rockville, MD 20850; Tel: 301-738-6262; Fax: 301-738-6255; E-mail: gary.gilliland@nist.gov

structures of approximately 50 hypothetical proteins from *Haemophilus influenzae* were to be determined [7] with the long-range goal of using protein structure to facilitate the assignment of function.

Structural genomics refers to the large-scale, high-throughput determination of three-dimensional structures of biological macromolecules. A major goal of such efforts is to determine representative structures of protein families so that homology modeling can be used to provide estimates of the structures of other family members. The accelerated rate of structure determination by x-ray and NMR methods makes structural genomics a feasible undertaking. Key advances include the use of recombinant DNA techniques, the availability of well tested crystallization screens [8], rapid x-ray data collection at synchrotron centers combined with the implementation of multiwavelength anomalous diffraction (MAD) methods [9,10], the availability of NMR cryoprobes for rapid data acquisition [11], new software for automatic structure determination and structure refinement [12], and advances in laboratory automation [13]. As the scope of structural genomics is immense, initial pilot projects, such as the one described here, are focusing on only a subset of proteins.

Presented here is a progress report on a pilot-scale structural genomics effort to use the structures of *Haemophilus influenzae* hypothetical proteins to assist in the assignment of function. The experimental approach is briefly described which includes target selection, gene cloning, expression and purification of the encoded protein, crystallographic and NMR methods, and approaches to functional assignment. The results of a number of the structure determination efforts and their subsequent analysis are then used to illustrate the progress being made in functional assignment using the three-dimensional structure and/or sequence information. Since *Haemophilus influenzae* is a pathogenic organism, the results from a number of the structure analyses also suggest new possible targets for drug development.

METHODOLOGICAL APPROACH¹

Target Selection

Target proteins were selected from the set of predicted coding regions in *Haemophilus influenzae* [1]. ORFs annotated as hypothetical or without well defined molecular function were considered. Relevant information for all ORFs has been assembled in a MYSQL database, with a web front-end (<http://s2f.umbi.umd.edu>). These data are updated monthly. Homologous proteins in the NR database are found using PSI-BLAST [14]. All med-line entries associated with the target or its homologs are identified, and links included in the database. The number of med-line

and where necessary, reading of the associated abstracts, entries provides a convenient and rapid means of assessing the available functional information for each putative target. Pan-genome techniques [15] are also used to obtain clues to possible protein function. Targets were eliminated if a structural homolog could be identified using IMPALA [16]; or if there are one or more transmembrane segments, as determined by TopPhred [17]; or if work by another group is known to be well advanced [<http://www.rcsb.org/pdb/strucgen.html>]. Links to a large number of other information resources are also provided. See the web site for details.

Cloning, Expression and Protein Purification

Heterologous expression of the *H. influenzae* targets in *E. coli* was initially carried out by amplifying selected target ORFs using conventional PCR with the Pfu polymerase and cloning them into plasmid vectors with either *trc* or T7 promoter systems [18,19]. The targets were expressed as either the native or fusion proteins with purification tags such as a thrombin-cleavable polyhistidine sequence. Optimum conditions for the expression of each protein were determined. Protein expression was also assessed for both rich medium and minimal medium containing either ¹⁵N and ¹³C-labeled nutrients or selenomethionine (SeMet) for structural studies. These studies showed that ~25% of the proteins have little or no expression from either the *trc* or T7 promoters and that about half of the selected targets readily express as a histidine-tagged fusion protein yielding between 1-100 mg of purified, native-like protein per liter of culture suitable for the structural studies. The remaining gene products had problems with expression, affinity purification or cleavage of the polyhistidine affinity tag. These became candidates for native protein expression and purification.

The his-tagged target proteins from *H. influenzae* expressed in *E. coli* were purified using metal ion affinity chromatography. After cell lyses the extracts were clarified, and passed over nickel ion affinity columns, which were washed with high salt, and eluted with a metal chelator such as imidazole. The purified fusion proteins were subjected to thrombin cleavage. The target proteins were separated from the his-tag peptide by passing the extracts again over a nickel ion affinity column. A benzamidine affinity column was used to remove thrombin from the protein solutions. The purified proteins were dialyzed and prepared for crystallization trials, NMR analysis, and biochemical characterization.

The native form of the target proteins were purified by classical protocols. Homogeneous protein solutions were obtained using a variety of fractionation steps, including ammonium sulfate precipitation, anion, cation, and hydrophobic interaction chromatography, and gel filtration. Most native proteins were purified with only two columns, and the yields were adequate for the needs of the structural studies.

Crystallography

Crystallization conditions were usually found quickly with fast screens [8], and then optimized with additional

¹Certain commercial materials, instruments, and equipment are identified in this manuscript in order to specify the experimental procedure as completely as possible. In no case does such identification imply a recommendation or endorsement by the National Institute of Standards and Technology nor does it imply that the materials, instruments, or equipment identified is necessarily the best available for the purpose.

experiments. Approximately 80% of the crystals used in the x-ray work required less than 400 experiments, which, in most cases, required less than 5 mg of purified protein. About 70% of the purified proteins have yielded some form of crystals, and of these more than 50% diffracted to better than 3.5-Å resolution, warranting a structure determination. Several of the proteins produced multiple crystal forms, and it was also observed that different expression systems produced differing results. For example, in the case of YacE (HI0890), crystals were not obtained from the His-tag-cleaved version of the protein that contained an additional 3 amino acid residues at the N-terminus, but the native protein produced crystals. In a few cases, the native and SeMet proteins crystallized in different forms, despite the similarity of the crystallization conditions. Typically, less than 2 mg of SeMet-protein samples were used, since the crystallization conditions usually required only optimization.

Flash-cooling protocols were developed for all protein crystals used in the crystallographic studies. Moreover, it was discovered during the course of the studies described here that the presence of high concentrations (>2.0 M) of highly soluble salts such as lithium formate and lithium chloride form ice-ring-free aqueous glasses upon flash cooling to 100 K [20]. These cryosalts are a new class of cryoprotectants that have been shown to be effective with a variety of commonly used crystallization solutions and have been effectively used with a number of crystals of several of the structures described here. Procedures were employed for characterizing, archiving, and shipping frozen crystals to and from the synchrotron facilities. X-ray data were collected at beamlines at both the APS and NSLS. Diffraction resolution typically improved by 0.2 to 1.0 Å when compared with that obtained using the home source.

The phase determination and refinement calculations were done with a variety of software tools. The automated software package SOLVE [21] was used to identify heavy atom sites, calculates phases and electron density maps. The "Halfbaked" direct method of SHELX [22] was also used successfully for identifying large number of Se sites, and the related program XPREP aided in the statistical analyses of MAD data required for determining the optimal resolution for phase calculations. Another software package, SnB [23], was also used for identification and conformation of Se sites. The phase refinement calculations used, in addition to the software mentioned above, CCP4's [24] MLPHARE [25], PHASES [26], and SHARP [27].

The electron density maps were interpreted using fully automated model building procedures of ARP/warp [28] and RESOLVE [29,30] when warranted by the quality and resolution of the data. These methods enabled the interpretation of 50-95% of the structure. Further map interpretation, correcting of the model during refinement, and verification of solvent placement relied on the program O [31].

The initial models built by ARP/wARP and RESOLVE were typically of high quality reducing the number of refinement cycles. Whether the model was built automatically or semi-automatically, fast initial refinement was facilitated by the use of algorithms capable of large

radius of convergence such as molecular dynamics methods. This was followed by positional and crystallographic temperature factor refinement. The use of real-space R-factor and automated solvent assignment helped reduce the time required for map inspection. Different computer programs were used for different problems, depending on the quality of the diffraction data, the quality of the model, and the stage of the refinement. CNS [32], TNT [33], REFMAC [34], and SHELXL [35] were all used successfully.

NMR Spectroscopy

All NMR spectra were recorded on Bruker DRX-500 and DRX-600 spectrometers equipped with triple resonance 3-axis gradient probes. Spectra were generally recorded in States-TPPI mode with the use of pulsed-field gradients for coherence selection and solvent suppression [36]. Sequence-specific backbone assignments were made using ¹⁵N HSQC [37], HNCO [38], HNCACB [39], and CBCA(CO)NH [40] experiments in combination with the automated assignment algorithm AUTOASSIGN [41]. Side-chain assignments were obtained from ¹³C CT-HSQC [42], HBHA(CO)NH [40], HCCH-COSY [43], HCCH-TOCSY [44], and ¹⁵N-edited TOCSY [45] spectra. Aromatic side chain resonances were assigned using a combination of 2D TOCSY [46] and 2D NOESY spectra [47]. Backbone dihedral angle restraints (ϕ) were identified from ³J_{HNHα} coupling constants derived from an HNHA experiment [48]. Distance information was collected from 2D NOESY (τ_m 100-150 ms), ¹⁵N-edited NOESY [49] (τ_m 100-150 ms), and ¹³C-edited NOESY [50] (τ_m 100-150 ms) spectra. Data were processed on a Dell Precision 410 workstation using nmrPipe/nmrDraw [51] and analyzed with SPARKY [52].

Initial structures were calculated with the CNS 1.0 software package [32] starting from an extended polypeptide chain. Standard simulated annealing protocols with torsion angle dynamics were used. Refinement from intermediate to final structures was carried out in an automated fashion using ARIA 1.0 [53]. Distance restraints were classified based on peak intensities as strong (1.8-2.7 Å), medium strong (1.8-3.1 Å), medium (1.8-3.5 Å), medium-weak (2.3-4.2 Å), weak (2.8-5.0 Å) and very weak (2.8-6.0 Å). Very weak NOE restraints were only used in cases where the reciprocal NOE could be observed. Two restraints, 2.3-3.2 Å for r_{N-O} and 1.5-3.0 Å for r_{HN-O}, were used for each hydrogen bond. These were only included in the final stages of refinement and were based on a combination of NOE patterns and H-D exchange data. The final values used for the force constants were 1,000 kcal mol⁻¹ Å⁻² for bond lengths, 500 kcal mol⁻¹ rad⁻² for angles and improper torsions, 40 kcal mol⁻¹ Å⁻² for experimental distance restraints, 200 kcal mol⁻¹ rad⁻² for dihedral restraints, and 4.0 kcal mol⁻¹ Å⁻⁴ for the van der Waals repulsion term. Structures were analyzed with PROCHECK-NMR [54] and QUANTA (Molecular Simulations, Inc.).

Interpretation of Function

The approach to gaining functional insight used in these studies has been through structural comparisons combined

with genome analyses that include detection of sequence homologies and conserved residues, of clustering of genes on DNA into operons and of phylogenetic patterns [55]. The structural analysis of a protein begins with comparing the structures with previously determined structures using resources such as CATH [56], SCOP [57,58], VAST [59] and Dali [60]. If the fold of the structure is recognized, this prompted detailed sequence analyses that would in many cases lead to the discovery of low-level sequence homologies that identified active site residues or other features that relate to function. The three-dimensional structure reveals the spatial distribution of charges, disposition of key functional groups and the shape of binding site cavities, all of which are indicators of protein function [61]. When the new structure does not resemble any known structure, assessment of function may prove to be more difficult, but the discovery of a new fold does expand the set of known protein folds that is of general use in understanding how protein structure may relate to function.

STRUCTURAL STUDIES

In the structural genomics pilot study reported on here, more than twenty-five structures of hypothetical proteins have been determined at this point. Examples listed in Table

1 are described below that illustrate the progress that has been made in establishing how structures of hypothetical proteins can assist in functional assignments. The individual summaries presented below reveal that the three-dimensional structures may provide very little, to a very rich level of detail that can be used to assist the functional assignment of the hypothetical proteins. Details of all proteins under investigation and their current status can be found at <http://s2f.carb.nist.gov/>.

HI0065 – YjeE, Nucleotide-Binding Protein

The YjeE gene (HI0065) encodes an 18-kDa protein that has homologues in virtually all bacteria, but not in archaea and eukaryotes. The amino acid sequence shows 30% to 60% identity within the family, but no homology with other proteins. The presence of the Walker A motif GXXXXGKT [62] indicates the possibility of a nucleotide-binding protein. The protein was cloned and expressed, and its structure determined at 1.7-Å resolution by the MAD method using the SeMet protein [63]. The protein structure shown in Fig. (1) revealed a nucleotide-binding fold and confirms that it belongs to the P-loop family of nucleotide-binding proteins, which includes nucleotide and nucleoside kinases, various ATPases, and G-proteins, although the topology of the β -sheet is unique. YjeE was a target for structure prediction at

Table 1. Hypothetical *H. Influenzae* Gene Product Structure Determinations

Protein	No. aa	x-ray/ nmr	Resolution (Å)	Method x-ray	Quality factor x-ray: R, R _{free} Nmr: rms (backbone, all heavy atoms)	Fold	Function
YjeE HI0065	158	x-ray	1.70	SeMet MAD	0.199, 0.235	P-loop NTP hydrolase	ATPase regulator of cell-wall synthesis
YfiA HI0257	107	nmr	n/a	n/a	0.36±0.07Å 0.91±0.22Å	Double-stranded RNA binding motif (c.f. staufer)	Binds to the 30S ribosomal subunit
YecO HI0319	241	x-ray	2.20	SeMet MAD	0.192; 0.254	Methyltransferase + cofactor	(1) Weak homology to SAM- dependent methyltransferase (2) Structure indicates specificity for small, unknown molecule
YchF HI0393	362	x-ray	2.40	MIR	0.239; 0.311	P-loop NTP hydrolase + Ubiquitin domain + Helix- loop-helix domain	RNA-binding GTPase translation factor
YjgF HI0719	130	nmr	n/a	n/a	0.85±0.22Å 1.46±0.24Å	Similar to chorismate mutase	Unknown
YacE HI0890	206	x-ray	2.0	Hg MAD	0.165; 0.218	NTP kinase	Dephospho-coenzyme A kinase
RbfA HI1288	128	x-ray	1.50	SeMet MAD	0.248; 0.302 in progress	RNA-binding motif of ERA	Ribosome-binding factor A
YbaK HI1434	158	x-ray	1.80	SeMet +Hg MAD	0.183; 0.259	New fold	(1) Homology to an insertion domain in prokaryotic Pro-tRNA synthetase; unknown function (2) Binds ATP in the NMR
YrbI HI1679	180	x-ray	2.30	SeMet MAD	0.184 ; 0.256	P-type ATPase	(1) Homology to part of CMP-N- acetylneuraminic acid synthetase (2) Phosphatase activity
ORN HI1715	182	x-ray	2.10	SeMet MAD	0.220 ; 0.265	RNA polymerase I domain	Homology to 3'-to-5' exoribonuclease

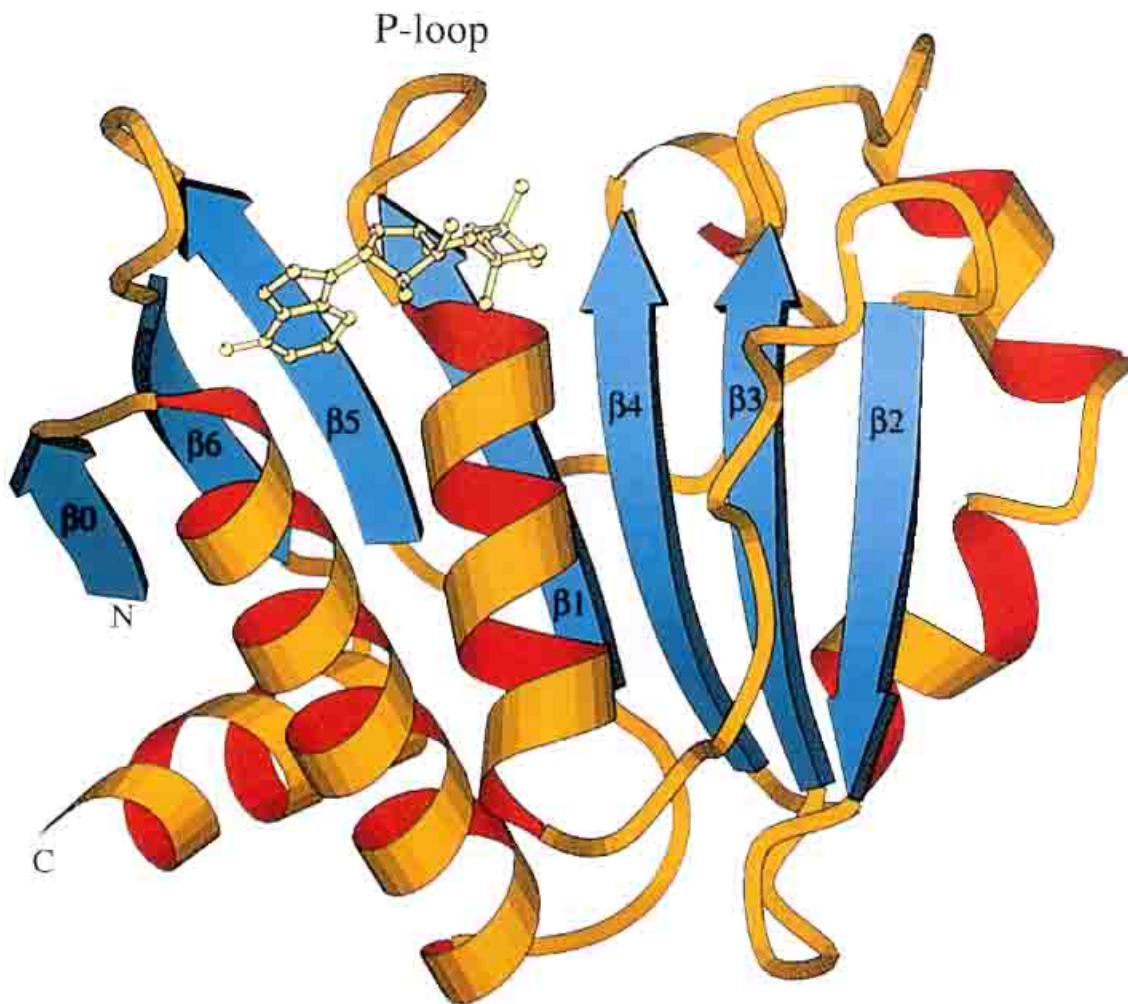


Fig. (1). Ribbon presentation of the polypeptide fold of YjeE (HI0065). The ADP molecule is represented by a ball-and-stick model. The figure was produced with MOLSCRIPT [115].

the CASP-4 experiment [64,65]. It should be noted that in spite of the apparent ease in identifying an NTP binding fold, all 85 attempts failed because of the unique, and unexpected arrangement of the β -sheet strands [66].

The unique topology of YjeE among P-loop proteins and certain similarities to both ATPases and GTPases suggest that it fills a topological niche and is likely to represent a separate group of P-loop proteins. Crystallization experiments and nucleotide modeling suggest the preference of YjeE to ATP rather than to GTP. The observation of a hydrolyzed nucleotide (ADP) in the active site implies ATPase activity of YjeE. The three-dimensional structure combined with the genomic analysis support the contention that the protein is as an ATPase regulator of cell wall synthesis. As such and because it is unique to prokaryotes, it may be a promising target for new antibiotics.

HI0257, YfiA, Ribosome Binding Protein

The three-dimensional solution structure of the 107 amino acid residue YfiA protein (HI0257) was determined

by NMR spectroscopic methods [67]. The sequence analysis revealed a 64% sequence identity to protein Y (also known as YfiA), a bacterial ribosome binding protein. This protein was recently shown to bind at the 30S/50S subunit interface stabilizing the ribosomal 70S complex against dissociation at low magnesium ion concentration [68,69]. The YfiA protein structure shown in Fig. (2) has a β - α - β - β - α folding topology with two parallel α -helices packed against the same side of a four-stranded β -sheet. The structural analysis did not identify any proteins in the Protein Data Bank [70] with exactly the same secondary structural motif, but it did identify a number of structures with some similarities. The closest structural homologues are proteins with the double-stranded RNA-binding domain (dsRBD) motif although there is little (<10%) sequence homology [71-74]. Although the dsRBD and YfiA structures are quite similar, a number of differences occur, including: the YfiA β -sheet has an extra β -strand at the N-terminus; the helices are parallel in YfiA but in dsRBD they are at an angle of about 30° to each other; and the extended loop commonly seen between the first and second β -strands of the dsRBD are absent in the YfiA structure. Further, an analysis of the surface electrostatic potential in YfiA and the dsRBD family

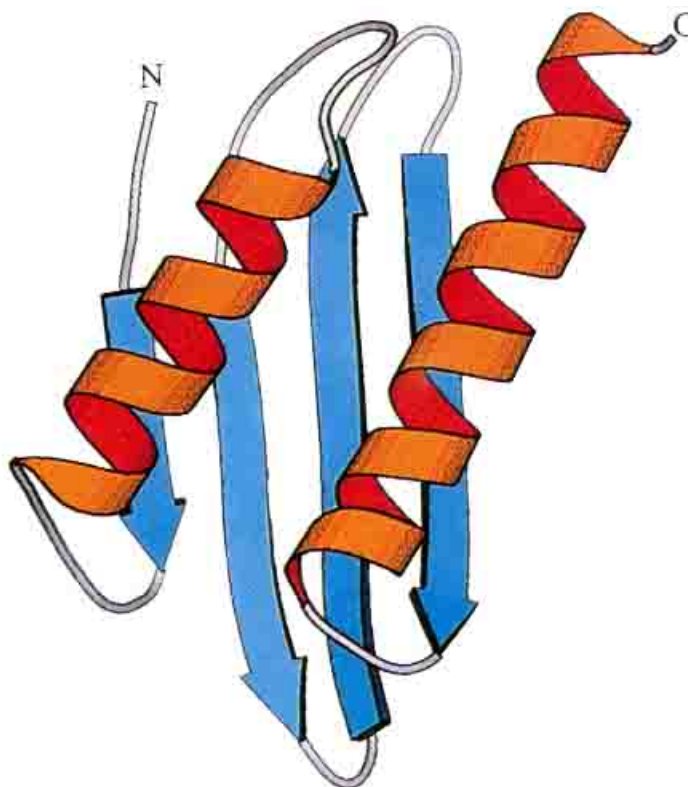


Fig. (2). A ribbon representation of the structure of YfiA (HI0257). The figure was produced with MOLSCRIPT [115].

reveals significant differences in the locations of contiguous positively and negatively charged regions. The structural data, in combination with sequence analysis of YfiA and its homologues, suggest that the most likely mode of RNA recognition for YfiA may be distinct from that of the dsRBD family of proteins.

HI0319 – YecO, S-Adenosyl-L-Methionine-Dependent Methyltransferase

The crystal structure of YecO (HI0319) was determined at 2.2-Å resolution [75]. The YecO structure shown in Fig. (3) revealed the classical fold of a SAM-dependent methyltransferase, with a bound co-factor, S-adenosyl-L-homocysteine (AdoHcy). The structure analyses identified methyltransferases as the best structural analogs followed by structures exhibiting a related α/β Rossmann fold [76].

The active site of the enzyme was identified by the location of the AdoHcy. The bound AdoHcy and nearby protein residues define a small solvent-excluded substrate binding cavity. The environment surrounding the cavity suggests that the substrate molecule contains an anionic group and a hydrophobic moiety. Many of the residues that define the cavity are invariant in the YecO sequence family but are not conserved in other methyltransferases indicating that YecO has a unique substrate specificity. An examination of the substrates of known methyltransferases was carried out using molecular modeling techniques.

The modeling studies indicated that the products of two candidate enzymes, anthranilate N-methyltransferase (EC

2.1.1.111) [77] and nicotinate N-methyltransferase (EC 2.1.1.7) [78], fit within the active site. Of the two, N-methylanthranilate, the product of anthranilate N-methyltransferase fits best in the active site cavity. The substrate can be docked such that the transferred methyl group is oriented towards the sulfur of AdoHcy, while the carboxylate group interacts with Arg198. The possibility remains that the substrate specificity of YecO methyltransferase has not yet been identified. Metabolites such as 2-hydroxy-benzoic acid, 2-carboxy-benzoic acid, and other analogs, are similar in shape to anthranilate, but have not been shown to be targets of any methyltransferase. These or other unidentified compounds are potential substrates.

HI0393 – YchF, GTPase

The YchF gene encodes a 40-kDa protein that has homologues in all organisms, from archae to mammals, and it is considered to be essential for cell survival. The crystal structure of YchF was solved to 2.4-Å resolution [79]. Based on the amino acid sequence analysis, YchF was ascribed to the superfamily of 11 universally conserved GTPases regulating functions such as protein synthesis, cell cycling and differentiation [80]. These proteins are characterized by the four sequence motifs, G1 to G4, that are related to the nucleotide binding in the frame of the Rossmann fold [76]. Motif G1 (GXXXXXGKS/T), also known as the Walker A motif, is the P-loop that binds the nucleotide phosphates. Motif G2, or switch I region, is conserved in each of the 11 protein families but not in the entire superfamily. In the presence of GTP, it is involved in

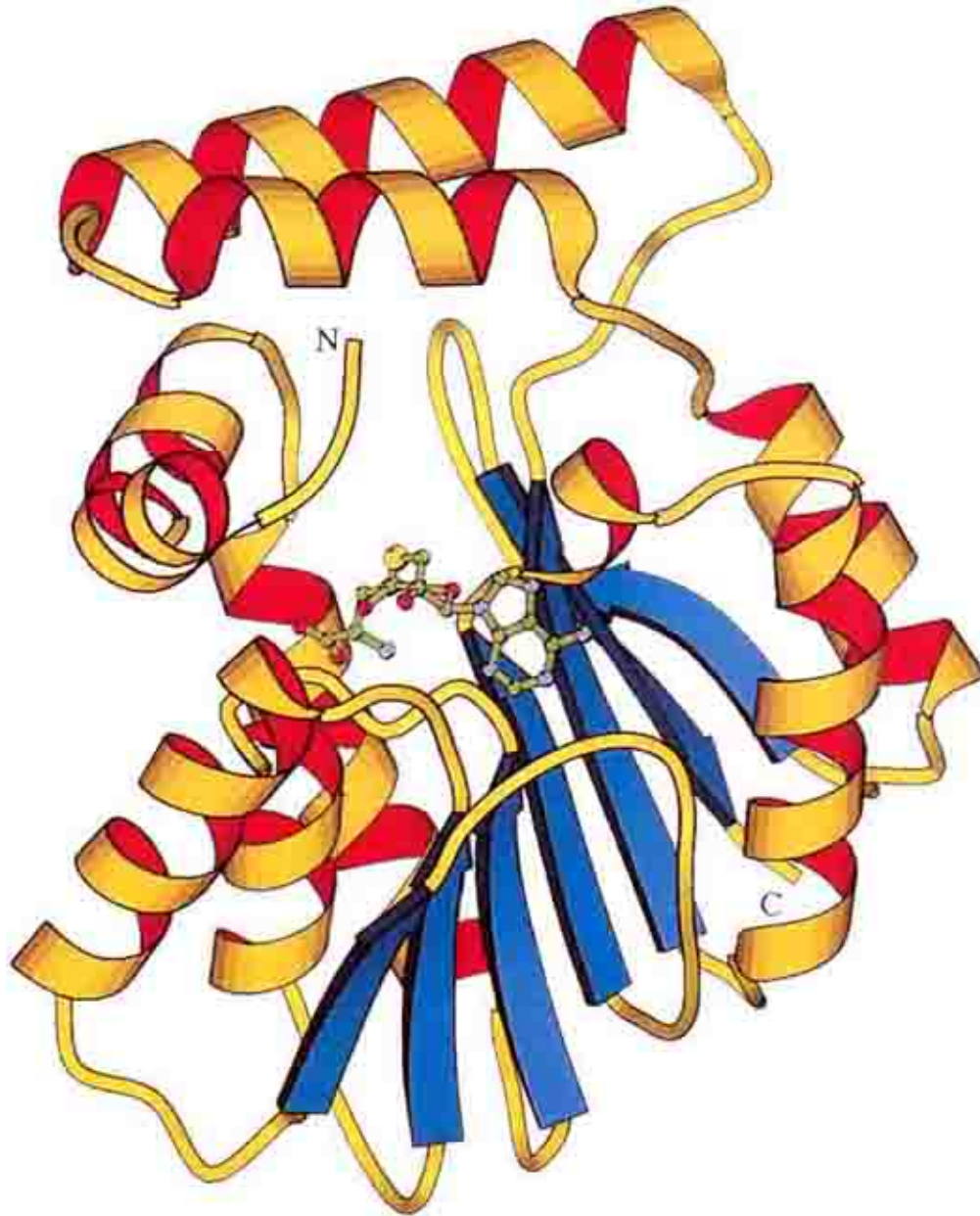


Fig. (3). Ribbon presentation of the three-dimensional structure of YecO (HI0319). The S-adenosyl-L-homocysteine molecule is represented by a ball-and-stick model. The figure was produced with MOLSCRIPT [115].

coordination of Mg^{2+} (often through threonine), and moves several Ångströms upon hydrolysis. Motif G3 (DXXG) or Walker B motif contains an aspartic acid that coordinates Mg^{2+} at the GTP hydrolysis site [62]. Finally, motif G4 with the sequence NKXD is the base recognition site.

A peculiar feature of YchF is that it lacks the G4 consensus sequence (N₁₁₁PAD₁₁₄ of the *E. coli* protein was erroneously taken into account although it is not conserved in the YchF family). The G4 motif is the only base discrimination feature that defines the specificity of NTP hydrolases. GTPases are more specific than ATPases, and therefore the G4 motif is used as a fingerprint in the functional assignment. Thus, in the absence of the consensus

NKXD sequence, YchF is more likely to be an ATPase. No experimental data are available yet to support either of the hypotheses. It should be noted, that molecular switches typically show very low levels of intrinsic nucleotide hydrolysis unless an effector or receptor is present.

The crystal structure of YchF (HI0393) resolves this question in favor of GTPase. The core of the structure is indeed a typical NTP binding fold with the Walker A and B motifs located as expected. However, the base recognition site shows some deviation from canonical structure. The role of the invariant lysine in the NKXD sequence is taken by Phe107, which stacks with the base. A glutamic acid, Glu210, is spatially equivalent to the canonical aspartate and

provides an H-bond acceptor for N1 and N2 of guanine. This is illustrated by the comparison of the active sites of YchF with H-Ras P21 [81] shown in Fig. (4). Glu210 is part of the sequence NXXE conserved in the YchF family with the exception of archaea where Glu is replaced by Asp. Asn207

of this sequence is equivalent to the canonical Asn, and forms H-bonds to O6 and N7 of guanine. Thus, the base recognition elements of YchF are located in two different regions of the sequence separated by 100 residues, whereas in other known GTPases they are grouped in a tetrapeptide.

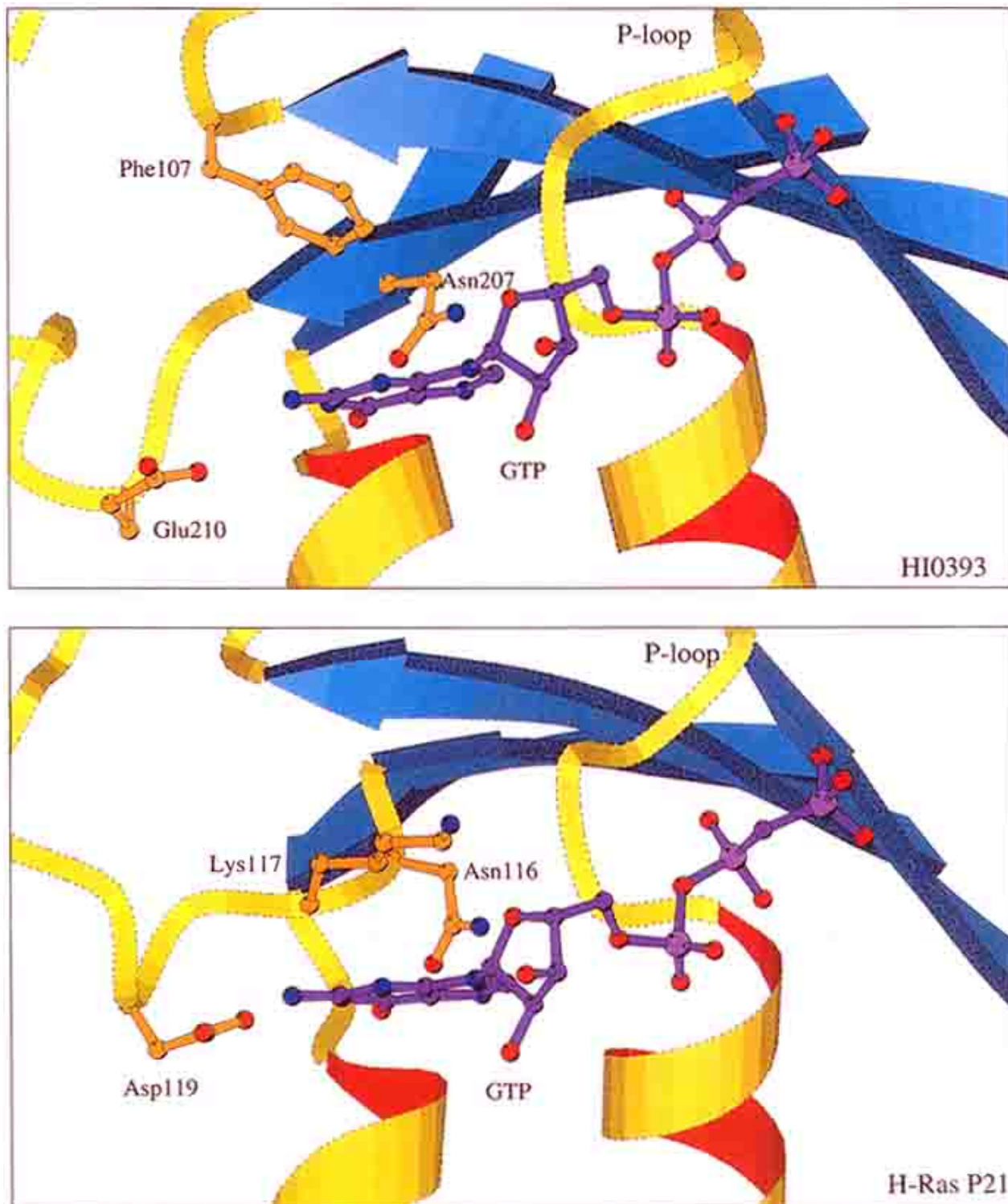


Fig. (4). A comparison of the active sites of YchF (HI0393) and H-Ras P21 [85]. Conserved amino acids and ATP are shown as ball-and-stick models. The figure was produced with MOLSCRIPT [115].

HI0719 – YjgF, Unknown Enzyme

The YjgF protein belongs to a family of highly conserved proteins found in bacteria to humans. These proteins have no strong sequence homology to any proteins of known function. The NMR structure of YjgF was determined and is the first solution structure for a member of this family [82]. The fold of the protein is similar to that of chorismate mutase [83] but there is little sequence homology and no apparent functional connection. YjgF is a homotrimer (verified by native gel analysis) with a distinct cavity located at the subunit interface. Six of the seven invariant residues in the high identity group of proteins are located in this cavity suggesting an enzymatic function for YjgF. Beyond this, the structure does not appear to give any further clues to a function.

A significant amount of work on YjgF family members has appeared in the literature in recent years, describing a variety of biological roles for this protein class. The crystal structures of two YjgF homologues have been determined recently, defining a new fold [84-85]. The NMR structure is consistent with these. Earlier work on a rat homologue of YjgF (43% identical [86]) suggested ribonuclease activity but we have not observed this for YjgF. Sinha and co-workers [84] suggested a role in purine regulation. In particular, YjgF-type proteins appear to be implicated in regulation of isoleucine biosynthesis and related pathways in a number of organisms [87-89]. NMR ligand screening of intermediates and structural analogues in these pathways, as well as screening of numerous commonly occurring small molecules in the cell, is currently underway in an effort to narrow down the potential biochemical role of this family of proteins.

HI0890, YacE, Dephospho-Coenzyme A Kinase

The crystal structure HI0890 from *Haemophilus influenzae* was determined at 2.0-Å resolution in complex with ATP [90]. When the X-ray analysis was in progress, functional assignment of YacE as dephosphocoenzyme A (dCoA) kinase (DCK) was reported [91]. The structural analysis verified that HI0890 was a kinase. According to the SCOP classification [57,58], HI0890 belongs to the family of nucleotide and nucleoside kinases. The molecule has a three-domain structure typical for nucleotide and nucleoside kinases [92]. The core domain consists of a 5-stranded parallel β -sheet flanked on both sides by α -helices. This is a canonical nucleotide-binding fold of the P-loop-containing proteins [93] with the sequence of β -strands 5-4-1-3-2. The second domain is an α -helical bundle inserted after strand β 2. It includes residues 32-102 folded into four α -helices (α 2, α 3, α 4, and α 5). In kinases, this domain has been designated the CoA domain since it forms the binding site for the acceptor substrate. The third domain is an insertion after strand β 4. It comprises residues 133-160 that forms a lid over the active site of the protein. Both the CoA and the lid domains have higher atomic temperature factors than other regions of the protein indicating their mobility. In fact, residues 59-64 and 143-147 are not visible at all in the electron density map.

HI1288 – rbfA, Ribosome Binding Factor A

The high-resolution crystal structure of HI1288 has been determined to 1.7-Å resolution [94]. The α/β type fold is composed of a three-strand anti-parallel β -sheet, one face of which interacts with two α -helices (see Fig. (5)). The native sequence contains 128 amino acid residues, 100 of which are found in the refined structure. All conserved amino acids in HI1288 are positively charged or hydrophobic. The surface formed by, or close, to the α -helices contains nearly all the conserved lysine and arginine residues clustered such that a protein-RNA complex can easily be formed. The open face of the β -sheet is more neutral and has to some extent a negative electrostatic potential. The conserved hydrophobic residues are found on the interior of the protein.

HI1288 has a 70% sequence identity to an *Escherichia coli* protein that has been identified as ribosome binding factor A (rbfA) [95]. It has been established that rbfA is associated with free 30S ribosomes, but not with the mature 70S ribosomes. Such a protein, that is needed for proper development and growth of a bacterium, and especially one involved in the fundamental processes of either translation or transcription, is a highly attractive target for drug development. The rbfA protein is an especially attractive target since the rbfA sequence is highly conserved in bacteria, and no protein sequence with high homology is present in eukaryotes, including humans. From the structure of rbfA, a highly conserved sequence pattern Lys31-Xxx-Pro33-Arg34 is found on the surface of the protein. This region contains the conserved residues that are surface exposed, and it is likely that this area is important for the function *in vivo*. The arrangement of these residues form an indentation on the surface making the site a highly attractive target for developing a drug-candidate to inhibit critical protein-RNA interactions. Pathogenic bacteria would therefore have a highly reduced capacity for protein translation.

HI1434, YbaK, Nucleotide-Binding Protein

The crystal structure of YbaK (HI1434) was determined at 1.8Å resolution [96]. The fold is only remotely related to the C-lectin fold, in particular to endostatin [97], and thus is not sufficiently similar to imply that YbaK proteins are saccharide binding proteins. However, a crevice that may accommodate a small ligand is evident. The putative binding site contains only one invariant residue, Lys46, which carries a functional group that could play a role in catalysis, indicating that YbaK is probably not an enzyme. Detailed sequence analysis, highlights sequence homology to an insertion domain in prolyl-tRNA synthetases (proRS) from prokaryote, a domain whose function is unknown. A YbaK-based model of the insertion domain shows that it should also contain the putative binding site. Being part of a tRNA synthetases, the insertion domain is likely to be involved in oligonucleotide binding, with possible roles in recognition/discrimination or editing of prolyl-tRNA. By analogy, YbaK may also play a role in nucleotide or oligonucleotide binding, the nature of which is yet to be determined. Preliminary ligand-screening by NMR (^{15}N HSQC methods), initially using mixtures of

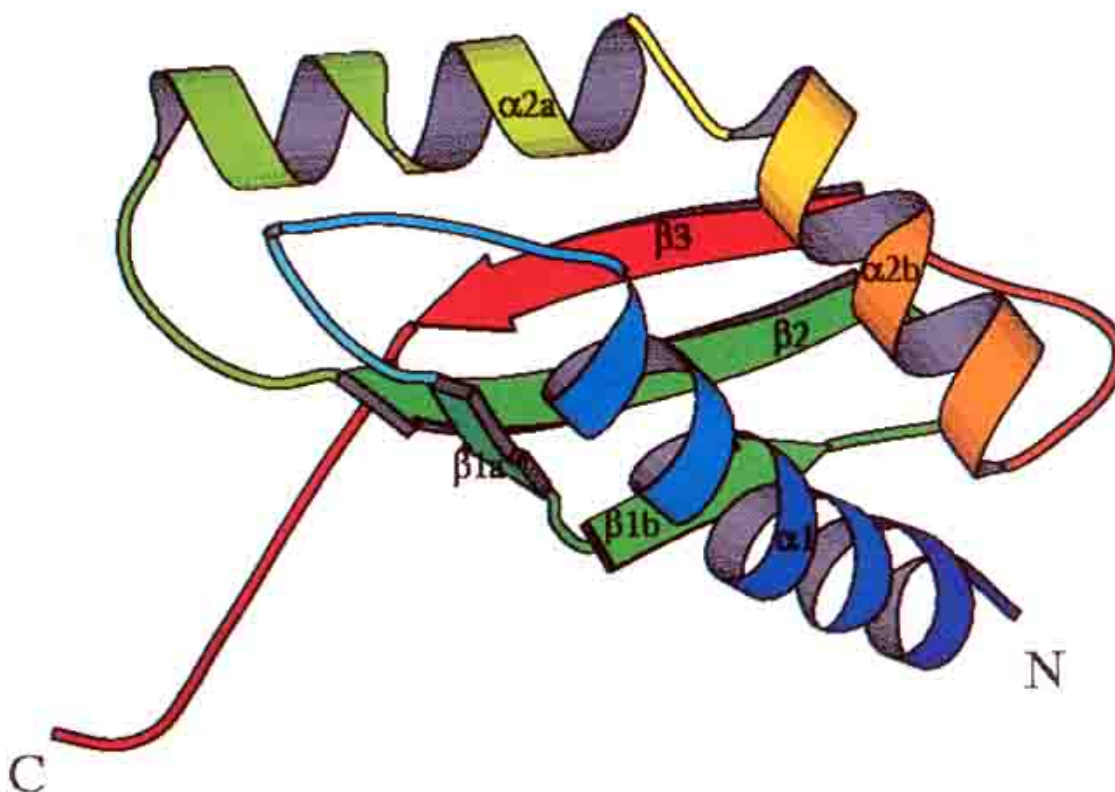


Fig. (5). The secondary fold components of *rbfA* (HI1288) are shown as a ribbon representation, colored from blue to red, from N- to C-terminus. The figure was produced with MOLSCRIPT [115].

nucleotides/nucleosides/nucleobases, and then deconvoluting the mixtures, showed that HI1434 binds ATP (and to a less extent ADP and AMP), but not any of the other compounds tested.

HI1679 – YrbI, Phosphatase

The crystal structure of the 19 kDa YrbI protein (HI1679) has been determined at 1.67 Å resolution [98]. The protein has an α/β Rossmann fold [76]), and the structural analysis found the fold to most closely resemble that of the L-2-haloacid dehalogenase (HAD) superfamily [99-102]. Once the structural similarity was observed, a detailed sequence analysis suggested a remote homology with two members of the HAD superfamily, phosphoserine phosphatase [103] and the P domain of Ca^{2+} ATPase [104]. The YrbI protein is a tetramer both in the crystal (see Fig. (6)) and in solution. The four subunits form a ring with an eight-stranded β -barrel at the center of the assembly. Each monomer contributes two β -strands formed by a $_{-}$ hairpin loop, inserted after the first β -strand of the core α/β fold. The four active sites are located at the subunit interfaces. Within each active site is a bound cobalt ion, an ion used in the crystallization of the protein. The cobalt ion is octahedrally coordinated, bound to two aspartate side chains, a backbone oxygen, and three solvent molecules, indicating that the physiological metal may be magnesium. Experiments found that YrbI hydrolyzes a number of phosphates, including 6-phosphogluconate and phosphotyrosine, suggesting that it functions as a phosphatase *in vivo*. The physiological substrate is yet to be

identified; however, the location of the gene on the *yrb* operon suggests involvement in sugar metabolism.

HI1715 – ORN, Oligoribonuclease

The HI1715 oligoribonuclease, composed of 182 amino acid residues, was selected as a target for structural studies because initially it was thought to be a hypothetical protein. During the course of the structural investigation, it was discovered that earlier reports had identified a corresponding gene in *Escherichia coli* as an oligoribonuclease [105,106]. Of the ribonucleases found in bacteria, oligoribonuclease is suggested to catalyze the final stages of polyribonucleotide degradation. In *Escherichia coli* this enzyme is the only ribonuclease that has been shown to be essential for viability [107]. The homologous protein from *Haemophilus influenzae*, HI1715, has a 73% sequence identity to the *E. coli* enzyme.

The X-ray crystal structure of the dimeric *H. influenzae* oligoribonuclease (HI1715) has been determined at 2.5-Å resolution [108]. The structural fold shown in Fig. (7) is of the mixed α/β type, with a central mixed β -sheet; the C-terminal region is primarily α -helical. The *H. influenzae* oligoribonuclease represents a new structural class of ribonucleases. The analysis of the structure and sequence identify the location of the active site and implicate a number of residues in catalysis. Active site homology between oligoribonuclease, *E. coli* exonuclease I [109] and the proofreading domain of *E. coli* DNA polymerase I [110-

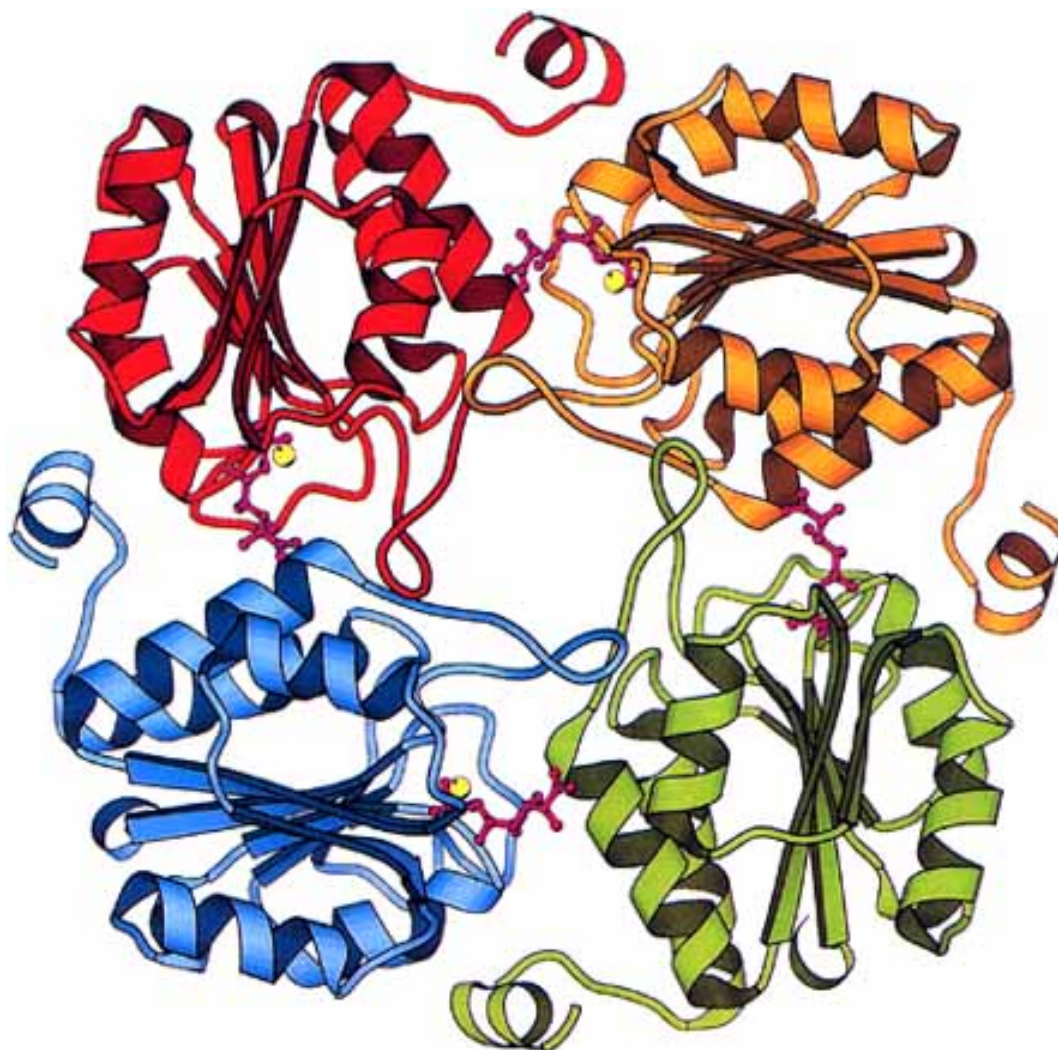


Fig. (6). The crystal structure of tetrameric YrBI (HI1679) is shown as a ribbon representation with each of the four subunit colored separately. The modeled 6-phosphogluconate molecules are shown as ball and stick figures. The figure was produced with MOLSCRIPT [115].

112] suggests that this oligoribonuclease hydrolyzes phosphodiester bonds in a two-metal ion catalyzed reaction. The functional annotation of the protein as a ribonuclease was initially based on the high sequence identity to the *E. coli* homologue on which a number of biochemical studies have been performed. Spectroscopic measurements with ORN indicate that it has activity for cytidine 2':3'-phosphate [113], supporting the evidence that it is a ribonuclease.

SUMMARY

The three-dimensional structures of biological macromolecules determined by x-ray crystallographic or NMR spectroscopic methods provide a wealth of

information. Fold recognition may give the first clues to function. This was the case for nearly all of the structures summarized in this report (see Table 1). Despite the fact that the selection criteria for target gene products included no homology to proteins with known structures, the majority of the structural results yielded folds that are similar if not identical to those of known structures. This is consistent with the fact that the majority of new family representatives entering the Protein Data bank [70] have known folds, even though there is still a significant fraction of new ones [114]. For example in the studies reported here, the structure of YjeE (HI0065) revealed a nucleotide-binding fold. This information coupled with a genome analysis suggests a functional role in cell wall synthesis. In another example, the crystal structure of YecO (HI0319) has the classic fold of

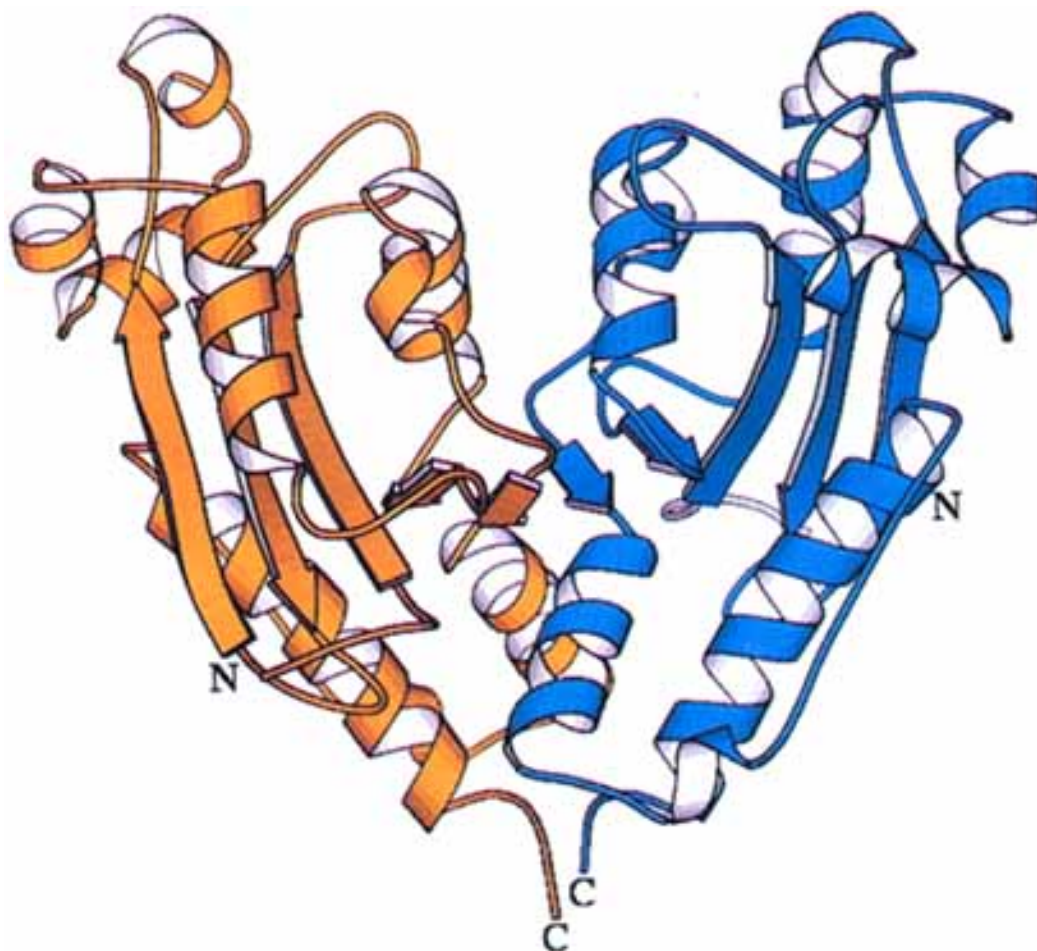


Fig. (7). The active form of the *H. influenzae* oligoribonuclease is a dimer. The two monomers of the dimer are shown as blue and yellow ribbons. The figure was produced with MOLSCRIPT [115].

a SAM-dependent methyltransferase, but in this case the substrate of the enzyme is not known, and hence its role in metabolism has yet to be deduced. In both of these studies, testable hypotheses of function have been developed that will have to be confirmed.

The structural results provided further surprises when the discovery of bound ligands and ions are found. The presence of these ions and compounds limit the possible functions. The structures of YecO (HI0319) and YrbI (HI1679) are good examples of this principle. For YecO the presence of S-adenosyl-L-homocysteine provided additional evidence that the protein is a SAM-dependent methyltransferase, and it pinpointed the location of the active site. For YrbI the presence of a cobalt ion also confirmed the location of the active site derived from the structural similarity with two HAD superfamily members, phosphoserine phosphatase [103] and the P domain of Ca^{2+} ATPase [104]. Cobalt appears to be substituted for magnesium, and subsequent experiments confirmed that the protein has phosphatase activity.

Biological research moves at an accelerating pace, providing much new information on hypothetical proteins that in some cases can establish function. During the course of the studies reported here, new annotation of several of the genes emerged. These included YfiA (HI0257), YacE (HI0890), rbfA (HI1288), and ORN (HI1715). In these examples two of the proteins YfiA and rbfA are both involved with ribosome function [68,69,95], YacE was reported to be a dephosphocoenzyme A (dCoA) kinase [91], and ORN as oligoribonuclease [105,106]. The structural results complement the functional annotation with three of the four cases and provide a structural basis for interpreting their biological activities.

As the structures of hypothetical proteins emerge and function assignment is made, many new targets for drug development will emerge. The structural genomics efforts described above, which focus on *H. influenzae* whose proteins are related to the pathogenic strain, illustrate this point. Several of the proteins described above (YjeE, HI0065; rbfA, HI1288; ORN, HI1715) that have been assigned function, that have important roles in metabolism,

and that are restricted to prokaryotes based on genomic profiling, are good candidates as drug targets. For example, the structure of YjeE (HI0065) along with the genomic analyses strongly suggests that it is as an ATPase regulator of cell wall synthesis. It is also limited to prokaryotes making it a potential drug target.

Finally, the structural genomics efforts reported here illustrate that the three dimensional architecture of proteins of unknown function can provide critical insight into the assignment of biochemical activity when the new fold resembles that of proteins whose functions are known. On the other hand, a major challenge is to assign biological function as the hypothetical protein structure diverges and if no ligands are found associated with the molecule. Despite this, when a new fold is revealed the universe of known folds is enriched, and once the function is confirmed or determined by other means, novel structure-function relationships are established.

ACKNOWLEDGEMENTS

This work was supported in part by the National Institutes of Health grant No. P01-GM57890. Use of the APS was supported by the U.S. Department of Energy, Basic Energy Sciences, Office of Science, under Contract No. W-31-109-Eng-38. This work was also supported in part by an award from W.M. Keck Foundation.

REFERENCES

[1] Fleischmann, R.D.; Adams, M.D.; White, O.; Clayton, R.A.; Kirkness, E.F.; Kerlavage, A.R.; Bult, C.J.; Tomb, J.-F.; Dougherty, B.A.; Merrick, J.M.; McKenney, K.; Sutton, G.; FitzHugh, W.; Fields, C.; Gocayne, J.D.; Scott, J.; Shirley, R.; Liu, L.-I.; Glodek, A.; Kelley, J.M.; Weidman, J.F.; Phillips, C.A.; Spriggs, T.; Hedblom, E.; Cotton, M.D.; Utterback, T.R.; Hanna, M.C.; Nguyen, D.T.; Saudek, D.M.; Brandon, R.C.; Fine, L.D.; Fritchman, J.L.; Fuhrmann, J.L.; Geoghagen, N.S.M.; Gnehm, C.L.; McDonald, L.A.; Small, K.V.; Fraser, C.M.; Smith, H.O.; Venter, J.C. *Science*, **1995**, 269, 496.

[2] Casari, G.; Andrade, M.A.; Bork, P.; Boyle, J.; Daruvar, A.; Ouzounis, C.; Schneider, R.; Tamames, J.; Valencia, A.; Sander, C. *Nature*, **1995**, 376, 647.

[3] Tatusov, R.L.; Mushegian, A.R.; Bork, P.; Brown, N.P.; Hayes, W.S.; Borodovsky, M.; Rudd, K.E.; Koonin, E.V. *Curr. Biol.*, **1996**, 6, 279.

[4] Colovos, C.; Cascio, D.; Yeates, T.O. *Structure*, **1998**, 6, 1329.

[5] Hwang, K.Y.; Chung, J.H.; Kim, S.H.; Han, Y.S.; Cho, Y. *Nature Struct. Biol.*, **1999**, 6, 691.

[6] Zarembinski, T.I.; Hung, L.W.; Mueller-Dieckmann, H.J.; Kim, K.K.; Yokota, H.; Kim, R.; Kim, S.H. *Proc. Natl. Acad. Sci. U.S.A.*, **1998**, 95, 15189.

[7] Eisenstein, E.; Gilliland, G.L.; Herzberg, O.; Moul, J.; Orban, J.; Poljak, R.J.; Banerjee, L.; Richardson, D.; Howard, A.J. *Curr. Opin. Biotechnol.*, **2000**, 11, 25.

[8] Jancarik, J.; Kim, S.-H. *J. Appl. Crystallogr.*, **1991**, 24, 409.

[9] Hendrickson, W.A. *Science*, **1991**, 254, 51.

[10] Smith, J.L. *Curr. Opin. Struct. Biol.*, **1991**, 1, 1002.

[11] Black, R.D.; Early, T.A.; Roemer, P.B.; Mueller, O.M.; Mobro-Campero, A.; Turner, L.G.; Johnson, G.A. *Science*, **1993**, 259, 793.

[12] Lamzin, V.S.; Perrakis, A. *Nature Struct. Biol.*, **2000**, 7 suppl., 978.

[13] Abola, E.; Kuhn, P.; Earnest, T.; Stevens, R.C. *Nature Struct. Biol.*, **2000**, 7 suppl., 973.

[14] Altschul, S.F.; Madden, T.L.; Schäffer, A.A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D.J. *Nucleic Acids Res.*, **1997**, 25, 3389.

[15] Marcotte, E.M.; Pellegrini, M.; Thompson, M.J.; Yeates, T.O.; Eisenberg, D. *Nature*, **1999**, 402, 83.

[16] Schaffer, A.A.; Wolf, Y.I.; Ponting, C.P.; Koonin, E.V.; Aravind, L.; Altschul, S.F. *Bioinformatics*, **1999**, 15, 1000.

[17] Claros, M.G.; von Heijn, G. *Comput. Appl. Biosci.*, **1994**, 10, 685.

[18] Amann, E.; Brosius, J. *Gene*, **1985**, 40, 183.

[19] Studier, F.W.; Rosenberg, A.H.; Dunn, J.J.; Dubendorff, J.W. *Methods Enzymol.*, **1990**, 185, 60.

[20] Rubinson, K.A.; Ladner, J.E.; Tordova, M.; Gilliland, G.L. *Acta Crystallogr. Sect. D*, **2000**, 56, 996.

[21] Terwilliger, T.C.; Berendzen, J. *Amer. Crystallogr. Assoc.*, **1997**, 24, ThB02.

[22] Sheldrick, G.M. *Proceedings of the CCP4 Study Weekend*; Wilson, K.S.; Davies, G.D. Eds.; SERC Daresbury Laboratory: Warrington, **1997**, pp. 147-158.

[23] Weeks, C.; DeTitta, G.T.; Hauptmann, H.A.; Thuman, P.; Miller, R. *Acta Crystallogr. Sect. A*, **1994**, 50, 211.

[24] Collaborative Computational Project, Number 4. *Acta Crystallogr. Sect. D*, **1994**, 50, 760.

[25] Otwinowski, Z. Maximum likelihood refinement of heavy atom parameters, in *Isomorphous Replacement and Anomalous Scattering*, Proceedings of the CCP4 Study Weekend 25-26 January, **1991**. Wolf, W.; Evans, P.R.; Leslie A.G.W. Eds.; Daresbury Laboratory, Warrington WA4 4AD, England. pp. 80-86.

[26] Furey, W.; Swaminathan, S. *Methods in Enzymol.*, **1997**, 277, 590.

[27] La Fortelle, E.; Bricogne, G. *Methods in Enzymol.*, **1997**, 276, 472.

[28] Perrakis, R.; Morris, R.; Lamzin, V.S. *Nat. Struct. Biol.*, **1999**, 6, 458.

[29] Terwilliger, T.C. *Acta Crystallogr. Sect. D*, **2001**, 57, 1755.

- [30] Terwilliger, T.C. *Acta Crystallogr. Sect. D*, **2001**, *57*, 1763.
- [31] Jones, T.A.; Zou, J.-Y.; Cowan, S.W.; Kjeldgaard, M. *Acta Crystallogr. Sect. A*, **1991**, *47*, 110.
- [32] Brunger, A.T.; Adams, P.D.; Clore, G.M.; DeLano, W.L.; Gros, P.; Grosse, K.R.; Jiang, J.S.; Kuszewski, J.; Nilges, M.; Pannu, N.S.; Read, R.J.; Rice, L.M.; Simonson, T.; Warren, G.L. *Acta Crystallogr. Sect. D*, **1998**, *54*, 905.
- [33] Tronrud, D.E.; TenEyck, L.F.; Matthews, B.W. *Acta Crystallogr. Sect. A*, **1987**, *43*, 489.
- [34] Murshudov, G.N.; Vagin, A.A.; Dodson, E.J. *Acta Crystallogr. Sect. D*, **1997**, *53*, 240.
- [35] Sheldrick, G.M. *SHELXL-97*, University of Göttingen, Germany, **1997**.
- [36] Bax, A.; Pochapsky, S.S. *J. Magn. Reson.*, **1992**, *99*, 638.
- [37] Mori, S.; Abeygunawardana, C.; Johnson, M.O.; van Zijl, P.C. *J. Magn. Reson. B*, **1995**, *108*, 94.
- [38] Grzesiek, S.; Bax, A. *J. Magn. Reson.*, **1992**, *96*, 432.
- [39] Wittekind, M.; Mueller, L. *J. Magn. Reson. Ser. B*, **1993**, *101*, 201.
- [40] Grzesiek, S.; Bax, A. *J. Am. Chem. Soc.*, **1992**, *114*, 6291.
- [41] Zimmerman, D.E.; Kulikowski, C.A.; Huang, Y.; Feng W.; Tashiro, M.; Shimotakahara, S.; Chien, C.; Powers, R.; Montelione, G.T. *J. Mol. Biol.*, **1997**, *269*, 592.
- [42] Vuister, G.W.; Bax, A. *J. Magn. Reson.*, **1992**, *98*, 428.
- [43] Bax, A.; Clore, G.M.; Driscoll, P.C.; Gronenborn, A.M.; Ikura, M.; Kay, L.E. *J. Magn. Reson.*, **1990**, *87*, 620.
- [44] Kay, L.E.; Ikura, M.; Bax, A. *J. Am. Chem. Soc.*, **1990**, *112*, 888.
- [45] Marion, D.; Driscoll, P.C.; Kay, L.E.; Wingfield, P.T.; Bax, A.; Gronenborn, A.M.; Clore, G.M. *Biochemistry*, **1989**, *29*, 6150.
- [46] Braunschweiler, L.; Ernst, R.R. *J. Magn. Reson.*, **1983**, *53*, 521.
- [47] Kumar, A.; Ernst, R.R.; Wuthrich, K. *Biochem. Biophys. Res. Commun.*, **1980**, *95*, 1.
- [48] Vuister, G.W.; Bax, A. *J. Am. Chem. Soc.*, **1993**, *115*, 7772.
- [49] Fesik, S.W.; Zuiderweg, E.R.P. *J. Magn. Reson.*, **1988**, *78*, 588.
- [50] Ikura, M.; Kay, L.E.; Tschudin, R.; Bax, A. *J. Magn. Reson.*, **1990**, *86*, 204.
- [51] Delaglio, F.; Grzesiek, S.; Vuister, G.W.; Zhu, G.; Pfeifer, J.; Bax, A. *J. Biomol. NMR*, **1995**, *6*, 277.
- [52] Goddard, T.D.; Kneller, D.G. *SPARKY 3*, University of California, San Francisco.
- [53] Nilges, M.; Macias, M.J.; O'Donoghue, S.I.; Oschkinat, H. *J. Mol. Biol.*, **1997**, *269*, 408.
- [54] Laskowski, R.A.; Rullmann, J.A.; MacArthur, M.W.; Kaptein, R.; Thornton, J.M. *J. Biomol. NMR*, **1996**, *8*, 477.
- [55] Moulton, J.; Melamud, E. *Curr. Opin. Struct. Biol.*, **2000**, *10*, 384.
- [56] Orengo, C.A.; Michie, A.D.; Jones, S.; Jones, D.T.; Swindells, M.B.; Thornton, J.M. *Structure*, **1997**, *5*, 1093.
- [57] Murzin, A.G.; Brenner, S.E.; Hubbard, T.; Chothia, C. *J. Mol. Biol.*, **1995**, *247*, 536.
- [58] Lo Conte, L.; Ailey, B.; Hubbard, T.J.; Brenner, S.E.; Murzin, A.G.; Chothia, C. *Nucleic Acids Res.*, **2000**, *28*, 257.
- [59] Gilbrat, J.F.; Madej, T.; Bryant, S.H. *Curr. Opin. Struct. Biol.*, **1996**, *6*, 377.
- [60] Holm, L.; Sander, C. *J. Mol. Biol.*, **1993**, *233*, 123.
- [61] Kasuya, A.; Thornton, J.M. *J. Mol. Biol.*, **1999**, *286*, 1673.
- [62] Saraste, M.; Sibbald, P.R.; Wittinghofer, A. *Trends Biochem. Sci.*, **1990**, *15*, 430.
- [63] Teplyakov, A.; Tordova, M.; Obmolova, G.; Thanki, N.; Howard, A.J.; Bonander, N.; Eisenstein, E.; Gilliland, G.L. *Proteins: Struct. Funct. Genet.*, **2002**, in press.
- [64] Moulton, J.; Fidelis, K.; Zemla, A.; Hubbard, T. *Proteins: Struct. Funct. Genet.*, **2001**, *Suppl. 5*, 2.
- [65] Murzin, A.; Hubbard, T. *Proteins: Struct. Funct. Genet.*, **2001**, *Suppl. 5*, 8.
- [66] Sippl, M.; Lackner, P.; Domingues, F.S.; Prlic, A.; Malik, R.; Andreeva, A.; Wiederstein, M. *Proteins: Struct. Funct. Genet.*, **2001**, *Suppl. 5*, 55.
- [67] Parsons, L.; Eisenstein, E.; Orban, J. *Biochemistry*, **2001**, *40*, 10979.
- [68] Agafonov, D.E.; Kolb, V.A.; Spirin, A.S. *Proc. Natl. Acad. Sci. U.S.A.*, **1997**, *94*, 12892.
- [69] Agafonov, D.E.; Kolb, V.A.; Nazimov, I.V.; Spirin, A.S. *Proc. Natl. Acad. Sci. U.S.A.*, **1999**, *96*, 12345.
- [70] Berman, H.M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T.N.; Weissig, H.; Shindyalov, I.N.; Bourne P.E. *Nucleic Acids Res.*, **2000**, *28*, 235.
- [71] Ryter, J.M.; Schultz, S.C. *EMBO J.*, **1998**, *17*, 7505.
- [72] Nanduri, S.; Carpick, B.W.; Yang, Y.; Williams, B.R.; Qin, J. *EMBO J.*, **1998**, *17*, 5458.
- [73] Bycroft, M.; Grunert, S.; Murzin, A.; Proctor, M.; St. Johnston, D. *EMBO J.*, **1995**, *14*, 3563.
- [74] Ramos, A.; Grunert, S.; Adams, J.; Micklem, D.R.; Proctor, M.R.; Freund, S.; Bycroft, M.; St. Johnston, D.; Varani, G. *EMBO J.*, **2000**, *19*, 997.
- [75] Lim, K.; Zhang, H.; Tempczyk, A.; Bonander, N.; Toedt, J.; Howard, A.; Eisenstein, E.; Herzberg, O. *Proteins: Struct. Funct. Genet.*, **2001**, *45*, 397

- [76] Rossmann, M.G.; Moras, D.; Olsen, K.W. *Nature*, **1974**, 250, 194.
- [77] Eilert, U.; Wolters, B. *Plant Cell, Tissue Organ Cult.*, **1989**, 18, 1.
- [78] Joshi, J.G.; Handler, P. *J. Biol. Chem.*, **1960**, 235, 2981.
- [79] Teplyakov, A.; Chu, S.; Reddy, P.; Gilliland, G.L. manuscript in preparation.
- [80] Caldon, C.E.; Yoong, P.; March, P.E. *Mol. Microbiol.*, **2001**, 41, 289.
- [81] Pai, E. F.; Krengel, U.; Petsko, G. A.; Goody, R. S.; Kabsch, W.; Wittinghofer, A. *EMBO J.*, **1990**, 9, 2351.
- [82] Parsons, L.; Eisenstein, E.; Orban, J. manuscript in preparation.
- [83] Chook, Y.M.; Gray, J.V.; Ke, H.; Lipscomb, W.N. *J. Mol. Biol.*, **1994**, 240, 476.
- [84] Sinha, S.; Rappu, P.; Lange, S. C.; Mantsala, P.; Zalkin, H.; Smith, J.L. *Proc. Natl. Acad. Sci. U.S.A.*, **1999**, 96, 13074.
- [85] Volz, K. *Protein Sci.*, **1999**, 8, 2428.
- [86] Melloni, E.; Michetti, M.; Salamino, F.; Pontremoli, S. *J. Biol. Chem.*, **1998**, 273, 12827.
- [87] Goupil-Feuillerat, N.; Cocaign-Bousquet, M.; Godon, J.J.; Ehrlich, S.D.; Renault, P. *J. Bacteriol.*, **1997**, 179, 6285.
- [88] Hesslinger, C.; Fairhurst, S.A.; Sawers, G. *Mol. Microbiol.*, **1998**, 27, 477.
- [89] Kim, J.M.; Yoshikawa, H.; Shirahige, K. *Genes Cells*, **2001**, 6, 507.
- [90] Obmolova, G.; Teplyakov, A.; Howard, A.J.; Gilliland, G.L. *J. Structural Biol.*, **2001**, 136, 119.
- [91] Mishra, P.; Park, P.K.; Drueckhammer, D.G. *J. Bacteriol.*, **2001**, 183, 2774.
- [92] Schulz, G.E.; Mueller, C.W.; Diederichs, K. *J. Mol. Biol.*, **1990**, 213, 627.
- [93] Schulz, G.E. *Curr. Opin. Struct. Biol.*, **1992**, 2, 61.
- [94] Bonander, N.; Tordova, M.; Howard, A.J.; Eisenstein, E.; Gilliland, G.L. manuscript in preparation.
- [95] Dammal, C.S.; Noller, H.F. *Genes Dev.*, **1995**, 9, 626.
- [96] Zhang, H.; Huang, K.; Li, Z.; Banerjee, L.; Fisher, K.E.; Grishin, N.V.; Eisenstein, E.; Herzberg, O. *Proteins: Struct. Funct. Genet.*, **2000**, 40, 86.
- [97] Hohenester, E.; Sasaki, T.; Olsen, B.R.; Timpl, R. *EMBO J.*, **1998**, 17, 1656.
- [98] Parsons, J.F.; Lim, K.; Tempczyk, A.; Krajewski, W.; Eisenstein, E.; Herzberg, O. *Proteins: Struct. Funct. Genet.*, **2002**, 46, 393.
- [99] Selengut, J.D.; Levine, R.L. *Biochemistry*, **2000**, 39, 8315.
- [100] Collet, J.F.; van Schaftingen, E.; Stroobant, V. *Trends Biochem. Sci.*, **1998**, 23, 284.
- [101] Collet, J.F.; Stroobant, V.; Pirard, M.; Delpierre, G.; Van Schaftingen, E. *J. Biol. Chem.*, **1998**, 273, 14107.
- [102] Collet, J.F.; Stroobant, V.; Van Schaftingen, E. *J. Biol. Chem.*, **1999**, 274, 33985.
- [103] Wang, W.; Kim, R.; Jancarik, J.; Yokota, H.; Kim, S. H. *Structure*, **2001**, 9, 65.
- [104] Toyoshima, C.; Nakasako, M.; Nomura, H.; Ogawa, H. *Nature*, **2000**, 405, 647.
- [105] Datta, A.K.; Niyogi, S.K. *J. Biol. Chem.*, **1975**, 250, 7313.
- [106] Yu, D.; Deutscher, M.P. *J. Bacteriol.*, **1995**, 177, 4137.
- [107] Ghosh, S.; Deutscher, M.P. *Proc. Natl. Acad. Sci. U.S.A.*, **1999**, 96, 4372.
- [108] Bonander, N.; Tordova, M.; Ladner, J.E.; Howard, A.J.; Eisenstein, E.; Gilliland, G.L. manuscript in preparation.
- [109] Breyer, W.A.; Matthews, B.W. *Nat. Struct. Biol.*, **2000**, 7, 1125.
- [110] Brautigam, C.A.; Sun, S.; Piccirilli, J.A.; Steitz, T.A. *Biochemistry*, **1999**, 38, 696.
- [111] Doublet, S.; Tabor, S.; Long, A.M.; Richardson, C.C.; Ellenberger, T. *Nature*, **1998**, 391, 231.
- [112] Shamoo, Y.; Steitz, T.A. *Cell*, **1999**, 99, 155.
- [113] Crook, E.M.; Mathias, A.P.; Rabin, B.R. *Biochem. J.*, **1960**, 74, 234.
- [114] Coulson, A.W.F.; Moulton, J. *Proteins: Struct. Funct. Genet.*, **2002**, 46, 61.
- [115] Kraulis, P.J. *J. Appl. Crystallogr.*, **1991**, 24, 946.

