Protein Science (1999), *8*:2428–2437. Cambridge University Press. Printed in the USA.

A test case for structure-based functional assignment: The 1.2 Å crystal structure of the yjgF gene product from *Escherichia coli*

KARL VOLZ

Department of Microbiology and Immunology, University of Illinois at Chicago, Chicago, Illinois 60612-7344

(Received May 24, 1999; Accepted July 15, 1999)

Abstract

The YER057c/YIL051c/YjgF protein family is a set of 24 full-length homologs, each ~130 residues in length, and each with no known function or relationship to proteins of known structure. To determine the function of this family, the structure of one member—the YjgF protein from *Escherichia coli*—was solved and refined at a resolution of 1.2 Å. The YjgF molecule is a homotrimer with exact threefold symmetry. Its tertiary and quaternary structures are related to that of *Bacillus subtilis* chorismate mutase, although their active sites are completely different. The YjgF protein has an active site curiously similar to protein tyrosine phosphatases, including a covalently modified cysteine, but it is unlikely to be functionally related. The lessons learned from this attempt to deduce function from structure may be useful to future projects in structural genomics.

Keywords: cysteine modification; structural genomics; X-ray structure; YjgF

Traditionally, determination of a protein's three-dimensional structure was pursued only after the protein's function and biological significance was well established. Genome projects are rapidly altering that approach: initiatives are underway to solve structures of proteins of unknown function based solely on the probability that they will have unique folds—hence, structural genomics: the ultimate mapping of all possible protein families and folds (Kim, 1998; Montelione & Anderson, 1999). But success in structural genomics will depend in part on our ability to deduce a protein's function from its structure alone. How often will this approach be possible? At present, only four examples exist to attest to the feasibility of structure-driven functional analysis (Lima et al., 1997; Colovos et al., 1998; Yang et al., 1998; Zarembinski et al., 1998), with mixed degrees of success.

This paper provides a fifth test case for structure-based functional determination, revealing both what surprises can arise and what limitations persist. The YER057c/YIL051c/YjgF family of proteins consists of ~30 homologs. The family is ubiquitous, with members appearing in species ranging from bacteria to humans. None of the proteins have any known function, and none are homologous to other proteins of known structure. One family member—a hypothetical protein from *Escherichia coli* known as YjgF—was crystallized, and its three-dimensional structure was solved and refined to atomic resolution. The results show that the folding topology of the YjgF protein is not unique, even though the topologically related Bacillus subtilis chorismate mutase has no identifiable sequence similarity. Although the function of YjgF was not entirely determined, the results suggest a number of experiments to complete that goal.

Reprint requests to: Karl Volz, Department of Microbiology and Immunology, University of Illinois at Chicago, Chicago, Illinois 60612-7344; e-mail: karl@e002.mim.uic.edu.

Results and discussion

Properties of the YER057c/YIL051c/YjgF family and members

An open reading frame in the *mgtA-pyrI* intergenic region of *E. coli* codes for a hypothetical 13.6 kDa protein named YjgF (Burland et al., 1995). A PSI-BLAST (Altschul et al., 1997) search with this sequence through the nonredundant Genbank at the National Center for Biotechnology returned 24 sequences with full length homology (Fig. 1), ranging from 28 to 77% identity (five partial sequences with significant homologies were also retrieved, and a few other homologs with multiple loop insertions, but they will not be included here). This group of proteins has been named the YER057c/YIL051c/YjgF family (hereafter simply referred to as the YjgF family). Most of the sequences were of hypothetical protein products of open reading frames identified through unrelated works or by genome projects. The species ranged from bacteria (archea, gram negative, gram positive, cyanobacter, thermophilic, and nitrogen-fixing) to vertebrates, including *Mus musculus*, *Rattus norvegicus*, *Capra hircus* (goat), and *Homo sapiens*. No plants were represented. Some organisms contain multiple paral-

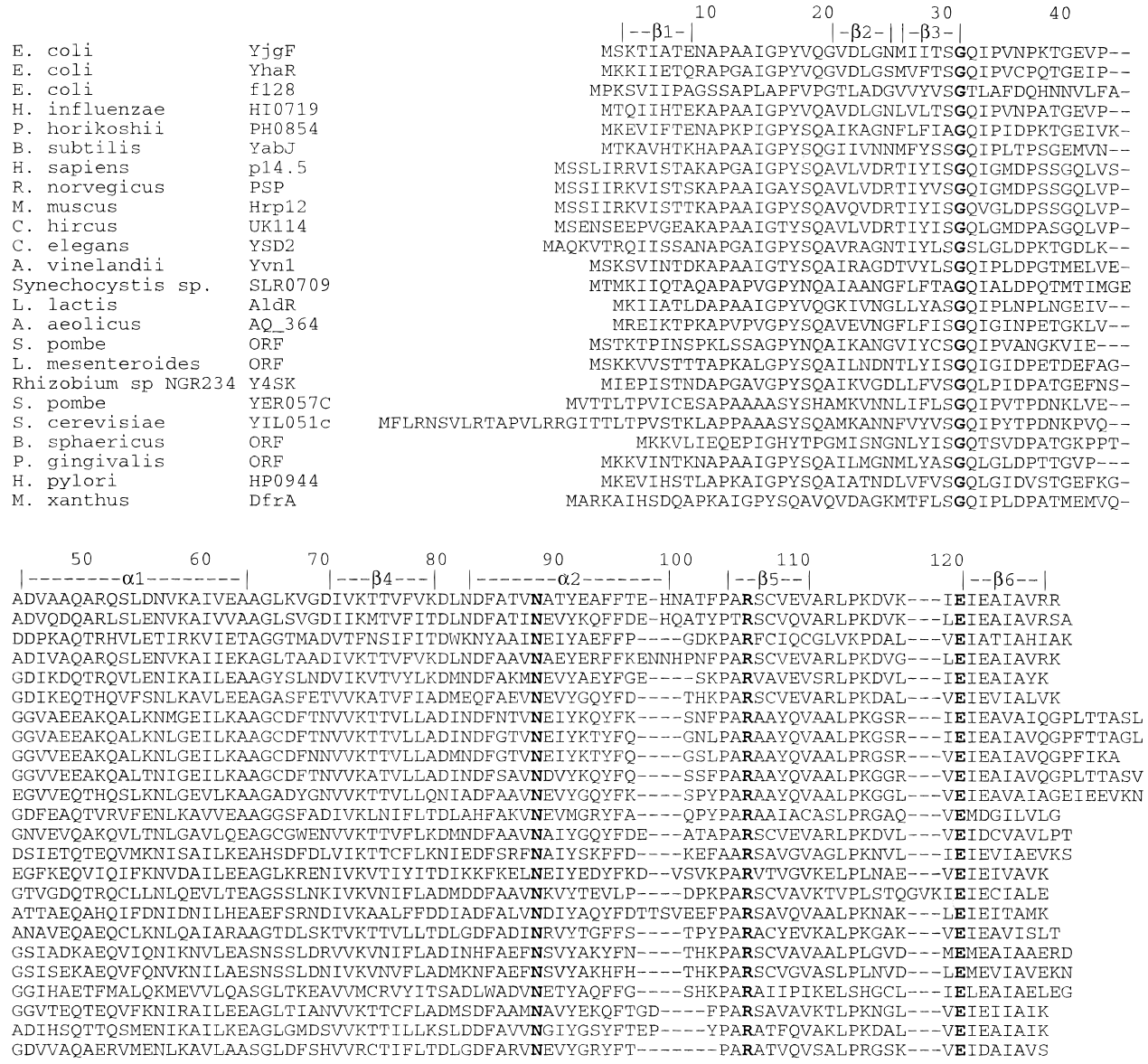

Fig. 1. Multiple sequence alignment of the YjgF family of proteins. Alignment was done manually. Invariant residues are shown in bold. α-Helices and β-strands are represented by brackets above the sequences.

ogs (e.g., *E. coli*, three sequences; *Saccharomyces cerevisiae*, two sequences). Reading frames were found in operons coding for proteins associated with such disparate pathways as pyrimidine biosynthesis (Burland et al., 1995), biotin biosynthesis (Gloeckler et al., 1990), threonine biosynthesis (Han et al., 1990), and nitrogen fixation (Joerger et al., 1989). A few of the protein products have been partially characterized: one homolog, known as the PSP protein from rat liver, was said to be an inhibitor of protein synthesis initiation (Oka et al., 1995). Another, p14.5 from human monocytes, was also described as a translational inhibitor (Schmiedeknecht et al., 1996). Other researchers have isolated a homolog called murine hepatic Hrp12, described as a novel, heat-responsive, tissue-specific, phosphorylated protein (Samuel et al., 1996). One group has reported that another YjgF homolog, named human UK114, is a tumor antigen expressed by various malignant neoplasms (Bartorelli et al., 1996; Ceciliani et al., 1996). A separate group observed that a bovine brain calpain activator is nearly identical to the UK114 protein from goat liver (Melloni et al., 1998). Finally, it has recently been shown that null mutations in the *yjgF* gene of *Salmonella typhimurium* cause multiple, pleotropic phenotypes involving the isoleucine biosynthetic pathway (Enos-Berlage et al., 1998). It is remarkable that a single protein family could have members that exhibit such varied putative functions.

Structure of the YjgF monomer

The YjgF protein was crystallized, and its structure was solved by conventional MIR methods. The monomers of the *E. coli* YjgF protein are 127 amino acids long, single domain, each folded up into a six-stranded mixed β-sheet packed tightly against two parallel, four-turn α-helices. The β-sheet is characterized by a +1,

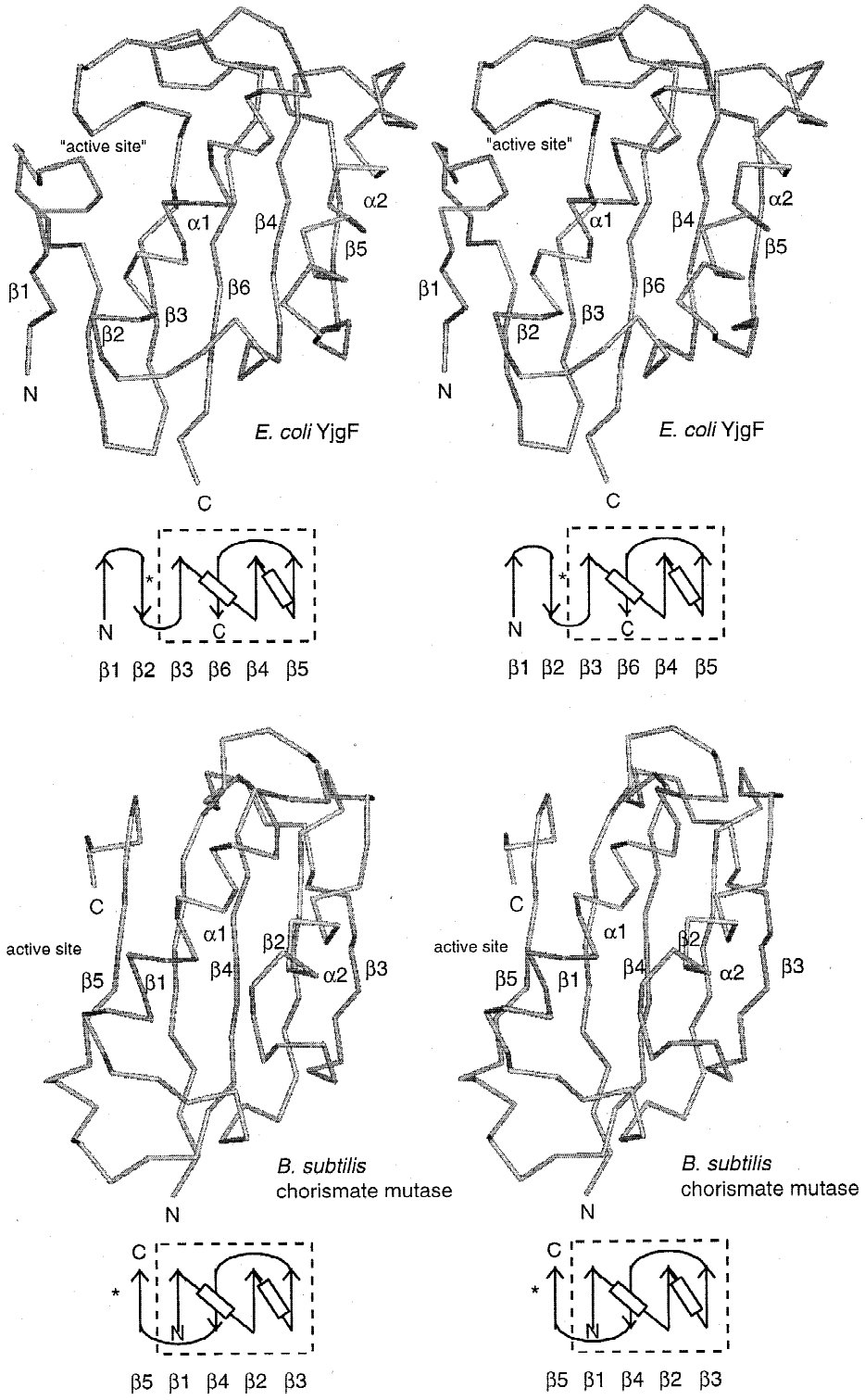

Fig. 2. Stereo diagram of α-carbon backbone of monomers of the E. coli YjgF protein and the *B. subtilis* chorismate mutase. The threefold axes of the trimers run approximately vertical, and the views are from the exterior. The folding topology diagrams are also shown. The asterisks denote the locations of the active sites in the diagrams, and the core β-sheets are enclosed in dashed lines.

$+1, +2, +1, -2$ topology (Fig. 2A). The first two strands, $\beta 1$ and $\beta 2$, are the shortest and are peripheral to the β-sheet, while the remaining four strands are longer, comprising the core of the sheet. In general, the connecting loops at the "tops" of the monomers are long and meandering, while the loops at the "bottoms" are shorter and tighter.

Structure of the YjgF trimer

The YjgF molecule is a homotrimer of perfect threefold symmetry. The core of the trimer is composed of 12 β-strands, four from each monomer, closed into a barrel-type structure, with six α-helices on the outside (Fig. 3). The core has extensive interactions, both hydrophobic and hydrophilic. The center of the core contains a well-ordered hydrogen-bonding network of 16 water molecules, surrounded by triplets of glutamates, lysines, serines, and threonines. Since the tertiary structure of the YjgF protein appears conserved within the sequences of the YjgF family (Fig. 1), it is likely that all other members of the family have this same trimeric quaternary structure.

The YjgF trimer possesses an extensive network of hydrogen bonding and electrostatic interactions throughout the molecule. The number of complementary interactions between oppositely charged ionizable groups in the molecule is noticeably greater than those seen in most other proteins from mesophiles. These stabilizing bonds in YjgF are not only surface interactions: some of the electrostatic bonds are quite buried, especially in the central core.

Similarities to B. subtilis chorismate mutase

E. coli YjgF and *B. subtilis* chorismate mutase have the same quaternary structure and similar tertiary structures. *B. subtilis* chorismate mutase (Protein Data Bank (PDB) codes 1COM, 2CHS, 2CHT) also assembles as a homotrimer (Chook et al., 1993, 1994). The monomers are single domain as well, each folded up into a five-stranded mixed β-sheet packed tightly against one 18-residue α-helix and one two-turn 3_{10} helix (Fig. 2B). The β-sheet is characterized by a $+2$, $+1$, -2, -2 topology.

The YjgF and chorismate mutase monomers have the same core topology in the subset of their four major β-strands, proceeding from right to left, as seen in Figures 2A and 2B. These four strands constitute the structural cores of the monomers and also serve as the scaffold for the trimers in both proteins. However, the two proteins are completely different in the rest of their structures. Each active site of chorismate mutase is formed by a C-terminal β-strand, β5, extending from the core sheet, positioned proximal to β3 of the adjacent monomer. In contrast, the polypeptide of the YjgF monomer terminates in the core sheet, so YjgF is missing that one loop and extra C-terminal β-strand that chorismate mutase possesses. Instead, the YjgF monomer is longer at its N-terminus: it has a loop and two extra β-strands at the N-terminus, extending over to the right-most β-strand of the sheet of the adjacent monomer, β5. Thus the cores of the two proteins are the same, but their additional β-strands come from the opposite termini of the polypeptide chains.

This similarity between YjgF and *B. subtilis* chorismate mutase was confirmed with the program DALI (Holm & Sander, 1993). A comparison of the YjgF structure with all representative folds in the PDB (Abola et al., 1997) gave the highest match with chorismate mutase, with a Z-score of 6.2, a 10% identity over the 82 equivalenced residues of the cores, and a positional RMSD of α-carbon atoms of 2.7 Å. The two other significant matches found

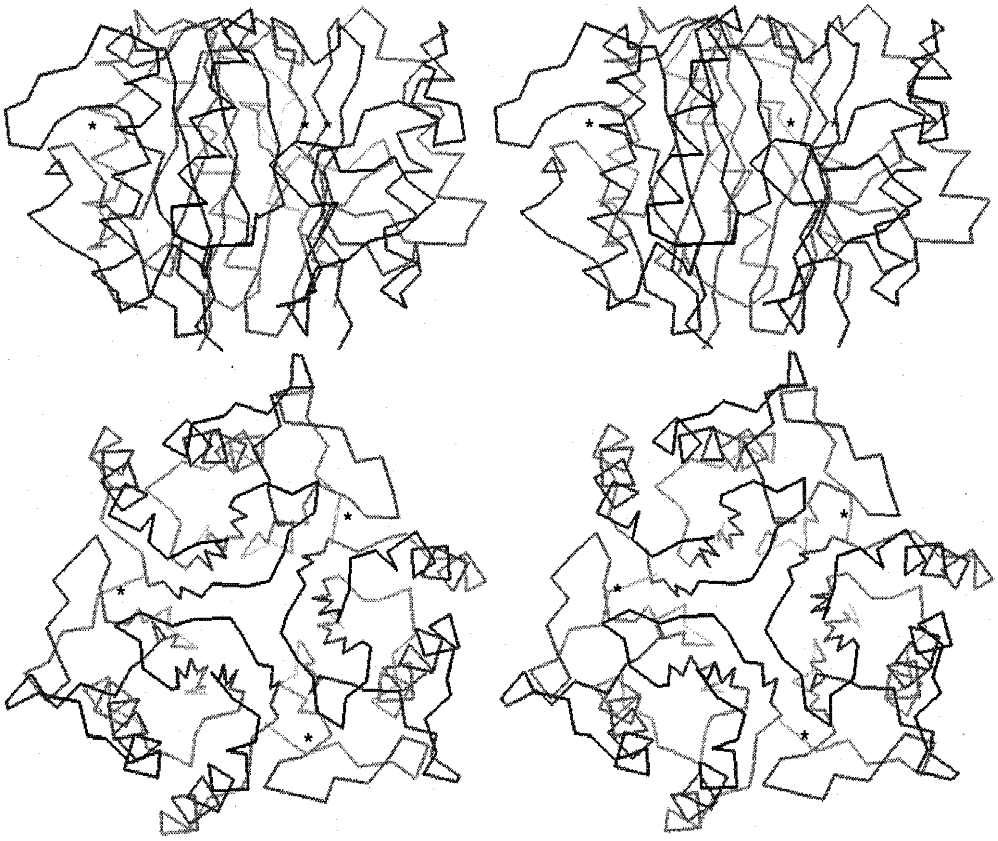

Fig. 3. Stereo diagrams of the YjgF trimer. The asterisks denote the locations of the active sites.

were with the C-terminal domain of FtsZ (PDB code 1FSZ) and a domain of tubulin (PDB code 1TUB). The structural similarities between the C-terminal domain of FtsZ, the *B. subtilis* chorismate monomers, and tubulin have been described elsewhere (Löwe & Amos, 1998; Nogales et al., 1998). The relationship between YjgF and chorismate mutase is more significant than those between YjgF and FtsZ or tubulin because of the conservation of quaternary structure as well as topology.

Identification of active sites in YjgF

Since the YjgF protein has no known function, there is as yet no functional basis for discussing active sites. But in general there are at least three structural characteristics of active sites that can be used to reasonably infer their existence in the absence of functional information: active sites are located in cavities or recessed regions of the molecule; in multimeric proteins, they are usually located on the interfaces between monomers; and finally, they have the most highly conserved amino acids of the protein's family within close proximity.

Three sites on the YjgF trimer exhibit all the above structural attributes of active sites. Near the equatorial surface of the trimer, the three interfaces between the adjoining subunits contain deeply recessed cavities. These cavities are in the same general locations as the three active sites in the *B. subtilis* chorismate mutase trimer. Each cavity is formed by a floor and four "walls (Fig. 4). The four walls comprise (1) the 11-residue loop following the β1 strand on the lower equatorial region (residues 8–18); (2) helix α2 of the adjacent monomer on the left (residues 81–88); (3) the type II turn on the upper region (residues 113–116); and (4) the 9-residue loop following the β3 strand on the right (residues 34–42). As for the last criterion for an active site, the four invariant residues of the family—Gly31, Asn88, Arg105, and Glu120—are all lining the cavity. Based on these observations, it is reasonable to simply refer to the cavities of the YjgF trimer as active sites.

Each active site of YjgF is a narrow and deep opening of about $4 \times 8 \times 8$ Å, containing ~10 ordered solvent molecules. The invariant arginine, Arg105, is prominent and immediately accessible from within the cavity, on the left wall as seen in Figure 4. It is in a strong hydrogen bonding interaction with the invariant Asn88. The role of Asn88 seemingly is to stabilize the position of Arg105. The invariant Glu120 is closing off the inner top of the active site as shown in Figure 4. The side chain of Glu120 is committed to hydrogen bonding interactions with the backbone of β5, but could conceivably rearrange and interact with other groups in the active site. The last invariant residue in the active site, Gly31, is most likely conserved for structural reasons (any Cβ substituent would have steric clash with the invariant Glu120 on β6). The same could be said for the 96% conserved Pro103. Two other highly conserved residues also deserve comment: Tyr17 (92% conserved) lines the bottom of the active site, with its hydroxyl group accessible, and Gln32 (92% conserved), on the far right side, also appears poised for interaction with substituents in the active site.

The most unusual aspect of the active site of YjgF is a covalent modification on the Sγ of Cys107. The Sγ atom occupies two low energy rotameric positions *t*, relatively buried, and $g+$, extending into the active site. In the latter position, the Sγ is 2.1 Å from a tetrahedral arrangement of electron density, which reaches a level of 7σ for the central peak (Fig. 5). The only chemical structures that reasonably fit this density are thiosulfate or thiophosphate.

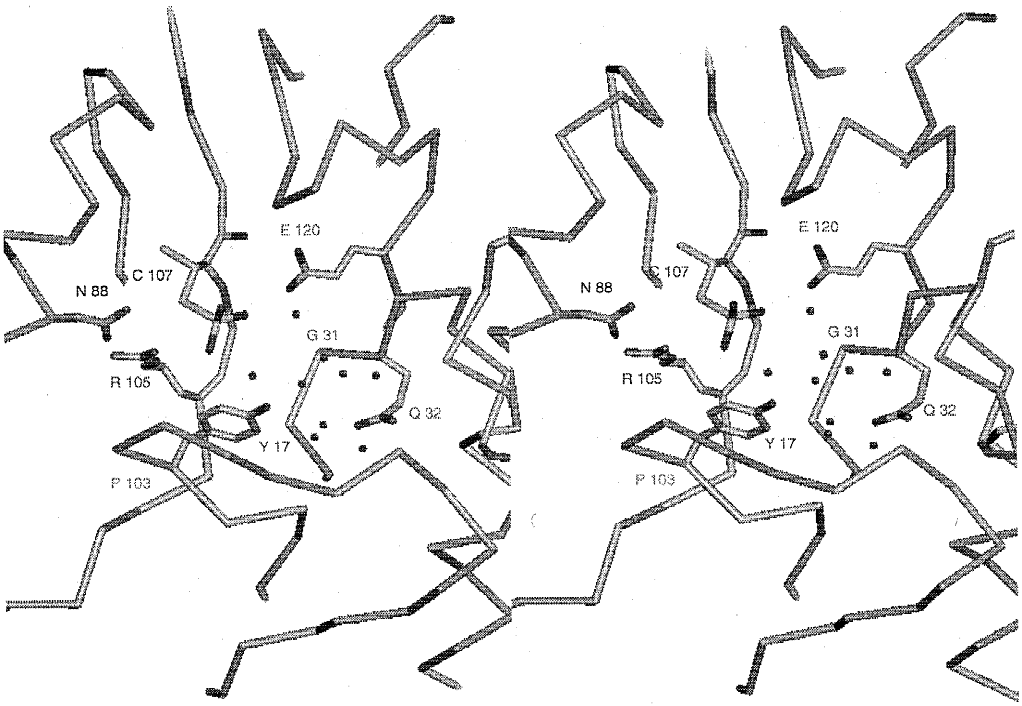

Fig. 4. Stereo diagram of the active site region of the YjgF protein. The four invariant residues of the YjgF family—Gly31, Asn88, Arg105, and Glu120—are labeled, as well as the three most highly conserved residues Tyr17, Gln32, and Pro103. Cys107 is also shown. The active site solvent molecules are shown as spheres.

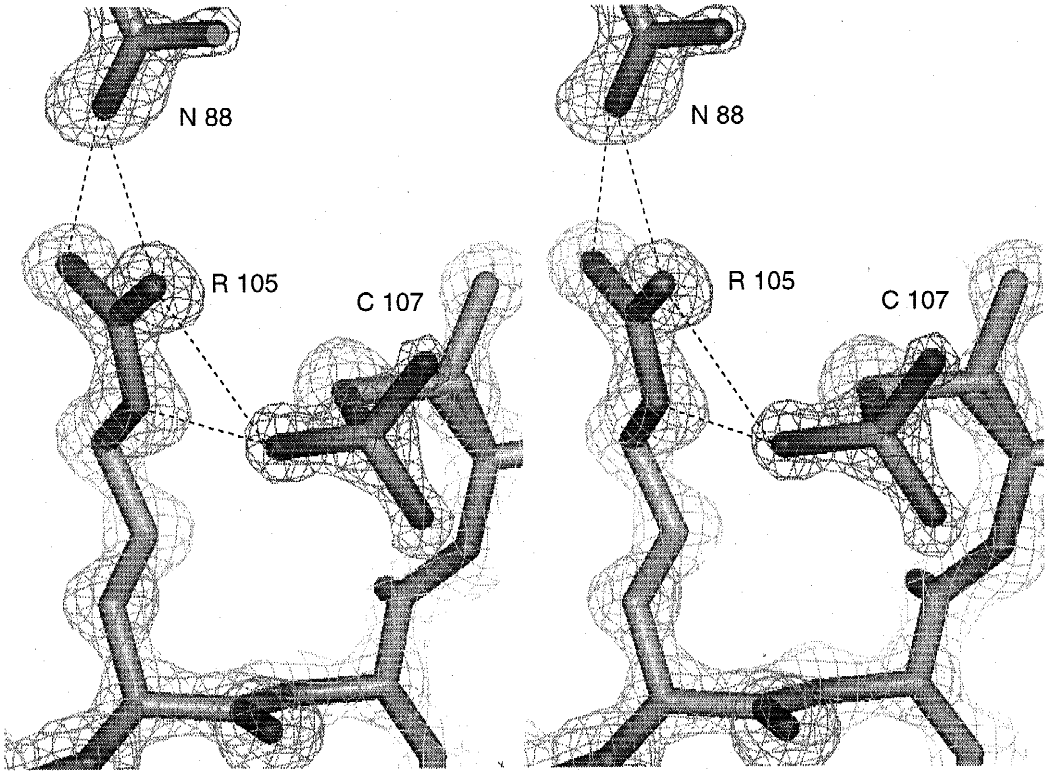

Fig. 5. Stereo diagram of electron density from final $|2F_o - F_c|\alpha_c$ map for the active site of YjgF. The electron density is contoured at 1.5σ.

When modeled as a thiophosphate with an occupancy of 0.67, the group refined to an average temperature factor of 14.9 Å^2 (the mean temperature factor for all protein atoms was 7.6 Å^2). At this very high resolution of 1.2 Å, the electron density for all atoms is clearly resolved.

Presence of the cysteine-bound moiety has been verified through mass spectrometry (data not shown). However, the masses of thiosulfate and thiophosphate are the same, so mass spectrometry could not discriminate between the two. Both thiosulfates and thiophosphates are rare and relatively unstable. One reason for the stability of this modified group on YjgF is the bidentate hydrogen bond between one of its own oxygen atoms and the Nϵ and NH_1 atoms of the guanidinium group of Arg105 (Fig. 5).

Functional implications for YjgF and the YjgF family

The major structural conclusions for YjgF are twofold. First, there is a clear topological relationship between *E. coli* YjgF and *B. subtilis* chorismate mutase. Similarity in their topologies is not likely to be a result of convergent evolution, but common ancestry may be impossible to prove. Second, although the active sites of YjgF and chorismate mutase are in the same relative locations in the molecules, they are completely unrelated in their constitution and configuration of amino acids. Therefore, it is reasonable to conclude that they have different functions.

The function of YjgF must relate to the conserved amino acids and any unique features present in its active site. As for the latter, the covalent modification of Cys107 is a clear indicator of unusual chemistry. Unfortunately, the results do not indicate whether the modified Cys107 is a thiosulfate or a thiophosphate. Cysteinyl sulfates have never been observed in structures of naturally occurring proteins, so if the group on Cys107 is a thiosulfate it is either an experimental artifact or a new and unprecedented structural result. Alternatively, one may assume that the Cys107 modification is a thiophosphate. There are two types of proteins that utilize thiophosphate chemistry: the protein-tyrosine phosphatases (PTPs) (Walton & Dixon, 1993; Fauman & Saper, 1996; Barford et al., 1998), and the structurally similar bacterial IIB$^{\text{cellobiose}}$ protein (van Monfort et al., 1997). The PTPs are cysteine-dependent phosphotransferases that proceed through a covalent cysteinyl-phosphate intermediate in the last step of a Mg^{2+} independent phosphoryl-group transfer reaction (Guan & Dixon, 1991; Zhang & Van Etten, 1991; Pannifer et al., 1998). All PTPs contain a highly conserved and very characteristic active site structure (Fig. 6A), composed of a phosphoryl-binding P-loop, a nucleophilic cysteine, a phosphoryl-binding arginine, an arginine-stabilizing glutamate, and a catalytic acidic group that assists a water molecule in nucleophilic attack on the phosphorous atom.

YjgF has no primary or tertiary structural relationship to the PTP superfamily, but some features of YjgF's active site resemble those of the PTPs. Although YjgF has no P-loop, the cysteinyl moiety is stabilized by the guanidinium group of invariant Arg105. In turn, the position of Arg105 is stabilized by interaction with the invariant Asn88. There is, however, no nucleophile in position for hydrolysis, so the cysteinyl modification is stable. The relative arrangement of these conserved active site groups is similar to that found in the structure of the phosphorylated PTP1B mutant Q262A (Pannifer et al., 1998) (Fig. 6B vs. 6A).

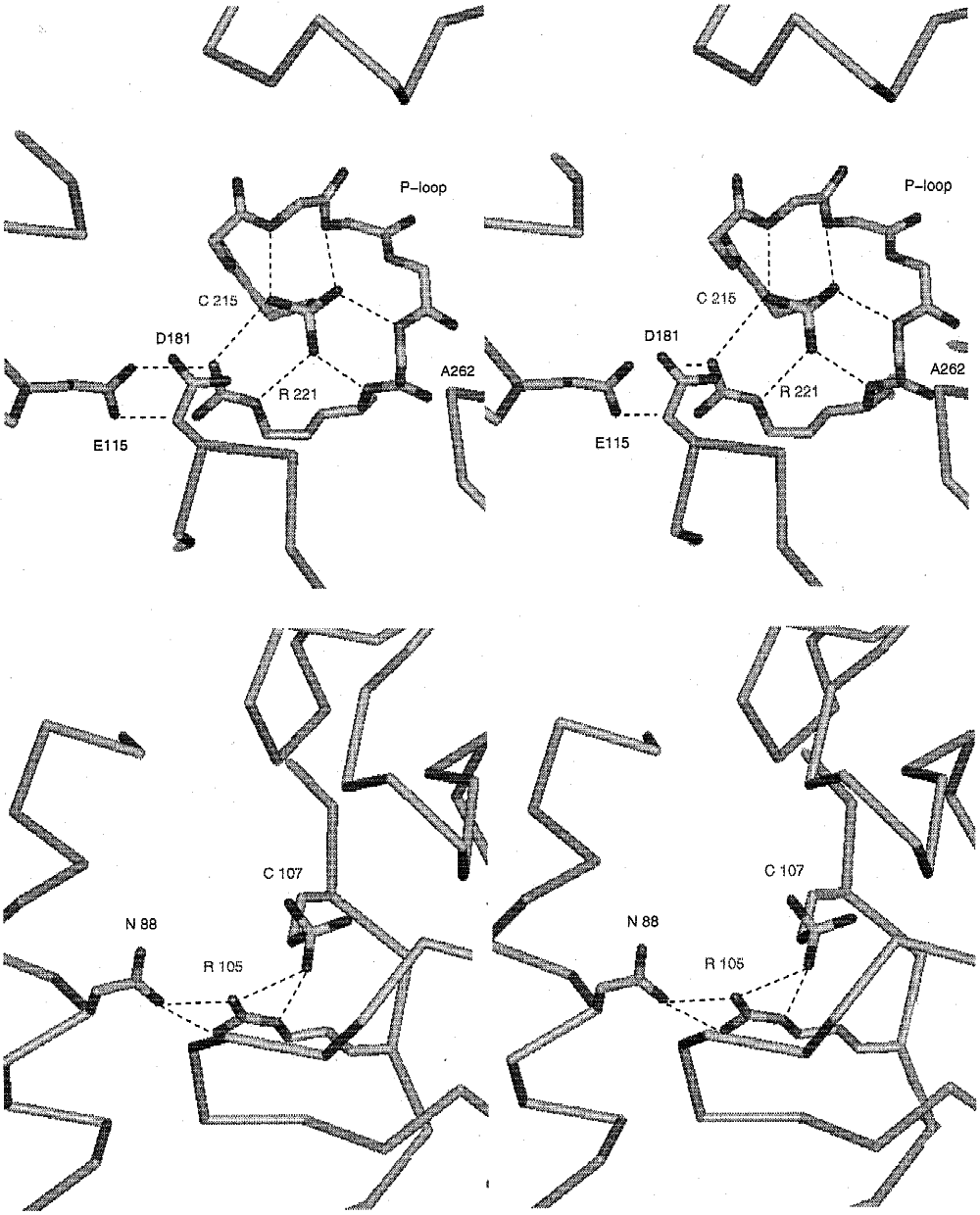

Fig. 6. Stereo diagram illustrating structural similarities between the active sites of the phosphorylated human protein tyrosine phosphatase 1B (mutant Q262A) (Pannifer et al., 1998) and the covalently modified *E. coli* YjgF protein.

There are two main problems with these assumptions and functional interpretations. In the first place, if YjgF is a phosphotransferase, then it is a most inefficient one, suffering from permanent product inhibition (YjgF shows no detectable phosphatase activity in a standard pNPP assay; S. Brakenridge & K. Volz, unpubl. data). The stability of the cysteinyl modification is not consistent with it being an intermediate in an enzyme reaction. Second, if one assumes that Cys107 is essential to the function of YjgF, then it would likely be invariant within the family. However, position 107 is poorly conserved: it is only 42% cysteine, with substitutions to alanine (42%), threonine (12%), and isoleucine (4%). Threonine could conceivably participate in phosphotransfer, but not alanine or isoleucine. Thus, this assignment of functional significance to Cys107 leads to the conclusion that the Yjgf family is divided in two, with different functions. This is not without precedent: other protein families are known to have members of the same fold but different functions (e.g., Hunt & Dayhoff, 1980; Babbitt et al., 1995). Interestingly, the PTP family itself contains members that are not phosphatases due to critical active site mutations (Wishart et al., 1995; Wishart & Dixon, 1998). The extent of this lack of functional correlation within other families is unknown, but could complicate structure-based assignment of function. Further difficulties are that the same protein can have multiple functions (Jeffery, 1999), and the same sequence can have different folds

(Goldstein, 1998). These functional assignment ambiguities will cause general problems in structural genomics, especially in the early stages when databases will be largely incomplete.

Although the sequence-to-structure-to-function approach was not successful in this test case, the results suggest a variety of experiments that should complete the goal. The first priority is to identify the Cys107 substituent and determine if it occurs in vivo. If so, Cys107's relationship to YjgF's function must be ascertained. It could be centrally important, or it could be a post-transcriptional modification not directly related to function. More general experiments include determination of the phenotype of an organism (*E. coli*, *S. cerevesiae*, or *Caenorhabditis elegans*) after deletions of all YjgF paralogs. Preliminary results toward this approach with *S. typhimurium* have already been reported (Enos-Berlage et al., 1998). Similarly, a yeast dual-hybrid system may identify other proteins that interact with YjgF. Experiments to test these possibilities are currently under way.

Materials and methods

Sequence and structural searches

The PSI-BLAST (Altschul et al., 1997) search was done at the National Center for Biotechnology website (http://www.ncbi.nlm.nih.gov/BLAST) using the tblastn option to search the nonredundant nucleotide sequence database. Also, a threading calculation (Marchler-Bauer & Bryant, 1997) was performed blindly by Drs. Bryant and Marchler-Bauer, but it did not retrieve the *B. subtilis* chorismate mutase structure. The best hit (P-value of 0.000554) was an all α-helical protein, the ciliary neurotrophic factor cytokine (PDB code 1CNT). The apparent reason for the failure to identify the match with chorismate mutase was that the superimposable substructure common in the two proteins is too small. The DALI calculation (Holm & Sander, 1993) was performed at the DALI server website at http://www.ebi.ac.uk/dali/.

Purification and crystallization

Purified *E. coli* YjgF protein was a gift from Dr. J. Wild's laboratory. The protein was concentrated by serial ammonium sulfate fractionation and centrifugation to a final concentration of 15 mg mL^{-1}. High ammonium sulfate conditions were explored for crystallization using microdialysis techniques. Phosphate salts were never employed. Microdialysis wells (Cambridge Repetition Engineers) of 30 μL volumes were used, with Spectrapor dialysis membrane having an M_r 3,500 (pore size) cutoff. All crystallization experiments were performed at 4 °C, with solutions buffered by 50 mM Tris-HCl at pH ranges from 7.9 to 8.5. The critical precipitation/crystal growth condition was 2.35 M ammonium sulfate. Single crystals grew within one week. The YjgF crystals assumed a cubic crystal habit, with a variety of truncations, with average dimensions of 0.2 mm on an edge.

Data collection and processing

The YjgF protein crystals were of excellent diffraction quality. The space group was found to be cubic, P23, with the unit cell parameter $a = 73.20$ Å, and one monomer per asymmetric unit. The V_m was 2.3 $Å^3$/Da. A number of data sets were collected for MIR phase determination, all at room temperature: a complete native data set was measured to 1.7 Å resolution on a Siemens multiwire area detector, and data sets of eight different potential heavy atom derivatives were collected to 2.0 Å resolution on either a Siemens multiwire area detector or a Rigaku RAXISII detector, using Cu Kα radiation from a Rigaku RU-H2R rotating anode generator. The final, complete set for high-resolution refinement was measured to 1.2 Å resolution at a wavelength of 0.961 Å under cryogenic conditions at the IMCA beam line ID17, Advanced Photon Source, Argonne National Laboratory. Crystal preparation for the cryoconditions was serial transfer of the crystal through 5% increases of glycerol in the original mother liquor, up to 20% over a total period of a few minutes, followed by flash cooling by immersion in liquid nitrogen. The unit cell parameter for the low-temperature crystal was $a = 72.22$ Å. Statistics for this final data set are given in Tables 1 and 2.

MIR phasing and structural solution

The YjgF structure was solved by multiple isomorphous replacement. All heavy atom soaks were at 10 mM concentrations of reagent for ~4 days. All Patterson maps were solved manually, starting with the $K_2Pt(NO_2)_4$ derivative. Since no protein crystallography program suites supported the cubic space group P23, MIR solution and refinement calculations (MLPHARE in CCP4, 1979; Otwinowski, 1991) were done with a threefold expansion of the data sets to space group P222, and expansion of the real-space asymmetric unit to include the entire trimer. The MIR solution finally yielded after the $UO_2(NO_3)_2$, $NaAuCl_4$, and sodium ethyl mercury thiosalicylate derivatives were solved and combined with the $K_2Pt(NO_2)_4$ derivative.

Model building and crystallographic refinement

The initial electron density maps were phased with the four heavy atom derivatives (MLPHARE, Otwinowski, 1991), and after density modification and threefold averaging (DM, Cowtan, 1994; RAVE, Kleywegt & Jones, 1994), they showed interpretable electron density. The maps were manually interpreted (using QUANTA,

Table 1. *Shell statistics for data and final refinement of E. coli YjgF*[a]

Shell Res limits	No. of reflections predicted	% Complete	% $>2\sigma_F$	*R*-value (%)[b]
$\infty \rightarrow 10.00$	265	100.0	—	—
$10.00 \rightarrow 2.31$	16,901	99.0	98.3	18.0
$2.31 \rightarrow 1.83$	16,927	98.0	96.0	15.8
$1.83 \rightarrow 1.60$	16,492	96.7	92.5	16.1
$1.60 \rightarrow 1.45$	17,027	95.3	87.9	15.3
$1.45 \rightarrow 1.35$	15,989	94.4	84.5	15.6
$1.35 \rightarrow 1.27$	16,538	91.7	79.9	15.0
$1.27 \rightarrow 1.20$	18,335	90.6	76.9	14.1
$10.00 \rightarrow 1.20$	118,474	94.8	87.7	16.4

[a] The number of reflections in the table correspond to space group P222.

[b] $R = \dfrac{\sum |F_{obs} - F_{calc}|}{\sum |F_{obs}|} \times 100.$

Table 2. *Refinement restraints and RMSDs from ideal of E. coli YjgF using the 103,863 reflections greater than 2σ from 10 to 1.20 Å*

Parameter	Target σ	Final value
Distance restraints (Å)		
Bond distance	0.020	0.011
Angle distance	0.040	0.031
Planar distance	0.050	0.048
Plane restraint (Å)	0.020	0.018
Chiral-center restraint (Å^3)	0.150	0.131
Nonbonded contact restraints (Å)		
Single torsion contact	0.500	0.164
Multiple torsion contact	0.500	0.142
Possible hydrogen bond	0.500	0.132
Conformational torsion angle restraint		
Planar (ω, 0, 180)	3.0	3.6
Staggered (±60, 180)	15.0	11.8
Orthonormal (±90)	20.0	15.6
Isotropic thermal factor restraints (Å^2)		
Main-chain bond	1.000	0.727
Main-chain angle	1.500	1.124
Side-chain bond	1.000	1.145
Side-chain angle	1.500	1.794
X-ray	$0.7*\|F_o - F_c\|$	16.4%
No. of reflections/No. of variables	—	8.0

1997, on a Silicon Graphics Indigo^2). During model building and refinement (X-PLOR, Brünger, 1993; PROLSQ/PROFFT, Hendrickson, 1985; Finzel, 1987), electron density maps were phased with phase combination of the partial models with the MIR phases (SFALL and SIGMAA in CCP4, 1979). Midway through the iterative rebuild/refine process, the topology of the last two-thirds of the molecule was recognized to be similar to that of *B. subtilis* chorismate mutase. Rebuilding and refinement of that portion of the molecule proceeded rapidly, primarily because of the extensive amount of regular secondary structure. The most difficult building was in the N-terminal third of the molecule, because it contains a rather high content of proline, glycine, and irregular secondary structure. Part of the loop between β1 and β2 (including Ile14 and Gly15) is a tenuous, extended chain with little hydrogen bonding support. Near the end of refinement, strong and persistent electron density centered 2.1 Å away from the Sγ of Cys107 was modeled as a phosphoryl group covalently bound to the thiol.

The final structure of the YjgF molecule contains 2,847 protein atoms for the trimer, with an *R*-factor of 16.4% for 103,863 reflections to 1.20 Å resolution (Tables 1, 2). For each monomer, the amino terminal methionine was absent, residues Ile14 and Gly15 had weak density, and the side chain of the C-terminal arginine was not able to be modeled. All other residues had clear and interpretable density for their backbones and side chains. There was no backbone disorder. There are 158 solvent molecules per monomer. There are no cis peptides. Cys107 is the only rotameric side chain, with ~67% occupancy of the side chain in the modified state, and ~33% in the reduced thiol form in an alternate conformation.

Mass spectrometry

Samples were analyzed using a Quattro II electrospray mass spectrometer (Micromass, Manchester, United Kingdom). Scans were done in both positive and negative mode. The largest peak in each spectrum corresponded to the molecular mass of the covalently modified form of the molecule (minus the amino terminal methionine) within the expected experimental error of 1.0 amu (<0.01%).

Coordinates

The atomic coordinates and structure factors are in the Protein Data Bank with the PDB code 1QU9.

Acknowledgments

I thank J. Wild and M. Wales at Texas A&M for their gift of the protein, A. Howard and coworkers at the IMCA ID17 beam line, Advanced Photon Source, Argonne National Laboratory, for access and beam time, R. Van Breemen and coworkers of the University of Illinois, Chicago, for collection of the electrospray mass spectrometry data, and S. Bryant and A. Marchler-Bauer for their threading calculations. This work was supported by National Institute of Health Grant GM 47522 to K.V.

References

Abola EE, Sussman JL, Prilusky J, Manning NO. 1997. Protein Data Bank archives of three-dimensional macromolecular structures. *Methods Enzymol 277*:556–571.

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucl Acids Res 25*:3389–3402.

Babbitt PC, Mrachko GT, Hasson MS, Huisman GW, Kolter R, Ringe D, Petsko GA, Kenyon GL, Gerlt JA. 1995. A functionally diverse enzyme superfamily that abstracts the a protons of carboxylic acids. *Science 267*:1159–1161.

Barford D, Das AK, Egloff M-P. 1998. The structure and mechanism of protein phosphatases: Insights into catalysis and regulation. *Annu Rev Biophys Biomol Struct 27*:133–164.

Bartorelli A, Bussolati B, Millesimo M, Gugliotta P, Bussolati G. 1996. Antibody-dependent cytotoxic activity on human cancer cells expressing UK 114 tumor membrane antigen. *Int J Oncol 8*:543–548.

Brünger AT. 1993. *X-PLOR version 3.1: A system for X-ray crystallography and NMR*. New Haven, Connecticut: Yale University Press.

Burland V, Plunkett G III, Sofia HJ, Daniels DL, Blattner FR. 1995. Analysis of the *Escherichia coli* genome VI: DNA sequence of the region from 92.8 through 100 minutes. *Nucl Acids Res 23*:2105–2119.

CCP4 (Collaborative Computing Project Number 4). 1979. *The SERC (UK) Collaborative Project No. 4, A suite of programs for protein crystallography*. Warrington, UK: Daresbury Laboratory.

Ceciliani F, Fatto L, Negri A, Colombo I, Berra B, Bartorelli A, Ronchi S. 1996. The primary structure of UK114 tumor antigen. *FEBS Lett 393*:147–150.

Chook YM, Gray JV, Ke H, Lipscomb WN. 1994. The monofunctional chorismate mutase from *Bacillus subtilis*: Structure determination of chorismate mutase and its complexes with a transition state analog and prephenate, and implications for the mechanism of the enzymatic reaction. *J Mol Biol 240*: 476–500.

Chook YM, Ke H, Lipscomb WN. 1993. Crystal structures of the monofunctional chorismate mutase from *Bacillus subtilis* and its complex with a transition state analog. *Proc Natl Acad Sci USA 90*:8600–8603.

Colovos C, Cascio D, Yeates TO. 1998. The 1.8 Å crystal structure of the ycaC gene product from *Escherichia coli* reveals an octameric hydrolase of unknown specificity. *Structure 6*:1329–1337.

Cowtan K. 1994. An automated procedure for phase improvement by density modification. *Joint CCP4 and ESF-EACBM Newsletter on Protein Crystallography 31*:34–38.

Enos-Berlage JL, Langendorf MJ, Downs DM. 1998. Complex metabolic phenotypes caused by a mutation in yjgF, encoding a member of the highly conserved YER057c/YjgF family of proteins. *J Bacteriol 180*:6519–6528.

Fauman EB, Saper MA. 1996. Structure and function of the protein tyrosine phosphatases. *Trends Biochem Sci 21*:413–417.

Finzel BF. 1987. Incorporation of fast Fourier transforms to speed restrained least-squares refinement of protein structures. *J Appl Crystallogr 20*:53–55.

Gloeckler R, Ohsawa I, Speck D, Ledoux C, Bernard S, Zinsius M, Villeval D, Kisou T, Kamogawa K, Lemoine Y. 1990. Cloning and characterization of the *Bacillus sphaericus* genes controlling the bioconversion of pimelate into dethiobiotin. *Gene 87*:63–70.

Goldstein DJ. 1998. An unacknowledged problem for structural genomics? *Nature Struct Biol 16*:696.
Guan K-L, Dixon JE. 1991. Evidence for protein tyrosine phosphatase proceeding via a cysteine-phosphate intermediate. *J Biol Chem 266*:17026–17030.
Han K-S, Archer JAC, Sinskey AJ. 1990. The molecular structure of the Corynebacterium glutamicum threonine synthetase gene. *Mol Micro 4*:1693–1702.
Hendrickson WA. 1985. Stereochemically restrained refinement of macromolecular structures. *Methods Enzymol 115*:252–270.
Holm L, Sander C. 1993. Protein structure comparison by alignment of distance matrices. *J Mol Biol 233*:123–128.
Hunt LT, Dayhoff MO. 1980. A surprising new protein superfamily containing ovalbumin, antithrombin III and a1-proteinase inhibitor. *Biochem Biophys Res Commun 95*:864–871.
Jeffery CJ. 1999. Moonlighting proteins. *Trends Biochem Sci 24*:8–11.
Joerger RD, Jacobson MR, Bishop PE. 1989. Two nifA-like genes required for expression of alternative nitrogenases by *Azotobacter vinelandii*. *J Bacteriol 171*:3258–3267.
Kim S-H. 1998. Shining a light on structural genomics. *Nature Struct Biol 5*:643–645.
Kleywegt GJ, Jones TA. 1994. Halloween . . . masks and bones. In: Bailey S, Hubbard R, Waller D, eds. *From first map to final model*. Warrington, UK: SERC Daresbury Laboratory. pp 59–66.
Lima CD, Klein MG, Hendrickson WA. 1997. Structure-based analysis of catalysis and substrate definition in the HIT protein family. *Science 278*:286–290.
Löwe J, Amos LA. 1998. Crystal structure of bacterial cell-division protein FtsZ. *Nature 391*:203–206.
Marchler-Bauer A, Bryant SH. 1997. A measure of success in fold recognition. *Trends Biochem Sci 22*:236–240.
Melloni E, Michetti M, Salamino F, Pontremoli S. 1998. Molecular and functional properties of a calpain activator protein specific for m-isoforms. *J Biol Chem 273*:12827–12831.
Montelione GT, Anderson S. 1999. Structural genomics: Keystone for a human proteome project. *Nature Struct Biol 6*:11–12.
Nogales E, Wolf SG, Downing K. 1998. Structure of the ab tubulin dimer by electron crystallography. *Nature 391*:199–203.
Oka T, Tsuji H, Noda C, Sakai K, Hong Y-M, Suzuki I, Muñoz S, Natori Y. 1995. Isolation and characterization of a novel perchloric acid soluble protein inhibiting cell-free protein synthesis. *J Biol Chem 270*:30060–30067.
Otwinowski Z. 1991. Maximum likelihood refinement of heavy atom parameters. In: Wolf W, Evans PR, Leslie AGW, eds. *Isomorphous replacement and anomalous scattering*. Daresbury, UK: Daresbury Laboratory. pp 80–86.
Pannifer ADB, Flint AJ, Tonks NK, Barford D. 1998. Visualization of the cysteinyl-phosphate intermediate of a protein-tyrosine phosphatase by X-ray crystallography. *J Biol Chem 273*:10454–10462.
QUANTA. 1997. San Diego: Molecular Simulations, Inc.
Samuel SJ, Tzung S-P, Cohen SA. 1996. Hrp12, a novel heat-responsive, tissue-specific, phosphorylated protein isolated from mouse liver. *Hepatology 25*:1217–1222.
Schmiedeknecht G, Kerkhoff C, Orsó E, Stöhr J, Aslanidis C, Nagy G, Knuechel R, Schmitz G. 1996. Isolation and characterization of a 14.5-kDa trichloroacetic-acid-soluble translational inhibitor protein from human monocytes that is upregulated during cellular differentiation. *Eur J Biochem 242*:339–351.
van Monfort RLM, Pijning T, Kalk KH, Reizer J, Saier MH, Thunnissen MMGM, Robillard GT, Dikstra BW. 1997. The structure of an energy-coupling protein from bacteria, IIBcellobiose, reveals similarity to eukaryotic protein phosphatases. *Structure 5*:217–225.
Walton KM, Dixon JE. 1993. Protein tyrosine phosphatases. *Annu Rev Biochem 62*:101–120.
Wishart MJ, Denu JM, Williams JA, Dixon JE. 1995. A single mutation converts a novel phosphotyrosine binding domain into a dual-specificity phosphatase. *J Biol Chem 270*:26782–26785.
Wishart MJ, Dixon JE. 1998. Gathering STYX: Phosphatase-like form predicts functions for unique protein-interaction domains. *Trends Biochem Sci 23*:301–306.
Yang F, Gustafson KR, Boyd MR, Wlodawer A. 1998. Crystal structure of *Escherichia coli* HdeA. *Nature Struct Biol 5*:763–764.
Zarembinski TI, Hung LW, Mueller-Dieckmann H-J, Kim K-K, Yokota H, Kim R, Kim S-H. 1998. Structure-based assignment of the biochemical function of a hypothetical protein: A test case of structural genomics. *Proc Natl Acad Sci USA 95*:15189–15193.
Zhang Z-Y, Van Etten RL. 1991. Leaving group dependence and proton inventory studies of the phosphorylation of a cytoplasmic phosphotyrosyl protein phosphatase from bovine heart. *Biochemistry 30*:8954–8959.