Crystal structure of RNA 3′-terminal phosphate cyclase, an ubiquitous enzyme with unusual topology

Gottfried J Palm[1†], Eric Billy[2], Witold Filipowicz[2] and Alexander Wlodawer[1,*]

Background: RNA cyclases are a family of RNA-modifying enzymes that are conserved in eucarya, bacteria and archaea. They catalyze the ATP-dependent conversion of the 3′-phosphate to the 2′,3′-cyclic phosphodiester at the end of RNA, in a reaction involving formation of the covalent AMP–cyclase intermediate. These enzymes might be responsible for production of the cyclic phosphate RNA ends that are known to be required by many RNA ligases in both prokaryotes and eukaryotes.

Results: The high-resolution structure of the *Escherichia coli* RNA 3′-terminal phosphate cyclase was determined using multiwavelength anomalous diffraction [AU: OK?]. Two orthorhombic crystal forms of *E. coli* cyclase (space group $P2_12_12_1$ and $P2_12_12$) were used to solve and refine the structure to R factor 20.4%; R_{free} 27.6% at 2.1 Å resolution [Au: OK?]. Each molecule of RNA cyclase consists of two domains. The larger domain contains three repeats of a folding unit comprising two parallel α helices and a four-stranded β sheet; this fold was previously [Au: OK?] identified in the translation initiation factor 3 (IF3). The large domain is similar to one of the two domains of 5-enolpyruvylshikimate-3-phosphate synthase and UDP–N-acetylglucosamine enolpyruvyl transferase. The smaller domain uses a similar secondary structure element with different topology, observed in many other proteins, such as thioredoxin.

Conclusions: The fold of RNA cyclase consists of known elements connected in a new and unique manner. Although the active site of this enzyme could not be unambiguously assigned, it can be mapped to a region surrounding His309, an adenylate acceptor, in which a number of amino acids are highly conserved in the enzyme from different sources. The structure of *E. coli* cyclase will be useful for interpretation of structural and mechanistic features of this enzyme

Addresses: [1]Macromolecular Crystallography Laboratory, Program in Structural Biology, National Cancer Institute-FCRDC, Frederick, Maryland 21702, USA and [2]Friedrich Miescher-Institut, CH-4002 Basel, Switzerland.

Present address: [†]Institute of Molecular Biotechnology, D-07708 Jena, Germany.

*Corresponding author.
E-mail: wlodawer@ncifcrf.gov

Key words: adenylation, cyclic phosphate, RNA cyclase, selenomethionine

Received: 6 August 1999
Revisions requested: 17 September 1999
Revisions received: 5 October 1999
Accepted: 14 October 1999

Published: Xx Xxxxxx

Structure January 2000, 8:00–000

0969-2126/00/$ – see front matter

Introduction

Terminal 2′,3′-cyclic phosphates are produced during RNA cleavage by many endoribonucleases, as either intermediates or final products of the reaction (reviewed by [1–3]). Sekiguchi and Shuman [4] have shown recently that the type I topoisomerase has endoribonuclease activity producing [AU: OK?] 2′,3′-cyclic phosphate ends. Moreover, the RNA-based enzymes such as the hammerhead, hairpin or hepatitis delta ribozymes, generate cyclic phosphate termini (reviewed by [5]). Interest in the 2′,3′-cyclic phosphate has heightened when it was found that two eukaryotic RNA ligases involved in tRNA splicing require cyclic ends for RNA ligation [6–13]; reviewed by [3,12,14,15]. One of these RNA ligases, an enzyme generating the 3′,5′-phosphodiester, 2′-phosphomonoester linkage, also functions in the splicing of the unusual intron present in *HAC1* pre-mRNA in yeast [16] and might be involved in the circularization of virusoid and viroid RNAs in plants (reviewed by [17,18]). Furthermore, the only known cellular eubacterial RNA ligase, an enzyme that joins RNA ends via the atypical 2′,5′-phosphodiester, also requires 2′,3′-cyclic termini. Physiological substrates of this ligase are not known [18,19].

Discovery of the RNA 3′-terminal phosphate cyclase indicated that endonucleolytic cleavage is not the only mechanism to generate 2′,3′-cyclic-phosphate-terminated RNAs. The cyclase, originally identified in extracts of HeLa cells and *Xenopus* oocytes, catalyzes the conversion of a 3′-phosphomonoester at the end of RNA to the 2′,3′-cyclic diester in a reaction dependent on ATP [7,8]. The cyclase has been purified from HeLa cell extracts and its cDNA has been cloned recently [20,21]. Ubiquitously expressed in all mammalian tissues and cell lines investigated, this enzyme is localized to the nucleoplasm [Au: OK?], consistent with its postulated role in RNA processing. Although the cyclase [AU: OK?] has no apparent sequence motifs in common with proteins of known function, database

searches indicated that genes encoding proteins with significant similarity to the cloned human enzyme are conserved among eucarya, bacteria and archaea [21]. Two different genes encoding cyclase-like proteins are expressed in mammals and *Drosophila* (EB and WF; unpublished observations). The protein encoded in the *Escherichia coli* genome, which has 34% identity and 43% similarity to the human cyclase, was found to have the RNA 3′-phosphate cyclase activity. The *E. coli* cyclase gene forms part of a previously uncharacterized operon, expression of which is controlled by an alternative sigma factor, σ^{54} [17,21].

The properties and substrate specificities of the human and bacterial cyclases are nearly identical. The enzymes cyclize the 3′-phosphate in RNAs with different sequences and base composition, and trinucleotides appear to be the shortest molecules able to act as substrates. With both enzymes, the cyclization of the 3′-phosphate occurs in three steps: first, Enzyme + ATP → Enzyme–AMP + PP_i; second, RNA-N$^{3'}$p + Enzyme–AMP → RNA-N$^{3'}$pp$^{5'}$A + Enzyme; third, RNA-N$^{3'}$pp$^{5'}$A → RNA-N>p + AMP. Evidence for the first step comes from identification of the covalent cyclase–AMP intermediates, whereas the second step is supported by accumulation of the RNA-N$^{3'}$pp$^{5'}$A molecules when the ribose at the RNA 3′ terminus is replaced with the 2′-deoxy- or 2′-*O*-methylribose. The third step probably takes place non-enzymatically [17,20–23]; reviewed by [17,21,24]. **[AU: OK? It is not our house style to have numbered points]**

Mechanistically, with respect to formation of the covalent protein–NMP intermediate and a transfer of NMP to the terminal phosphate in the nucleic acid, cyclase resembles RNA and DNA ligases, as well as capping enzymes (reviewed by [25]). In these enzymes, the nucleotidyl group is transferred to the 5′-terminal phosphate or pyrophosphate and a covalent lysyl–NMP intermediate is formed, with the active site lysine being present in a conserved sequence motif, *K*xDG **(in the single-letter amino acid code and in which x denotes any amino acid) [AU: OK?]** [25]. At the molecular level, however, the 3′-phosphate cyclization must substantially differ from reactions catalyzed by ligases and capping enzymes. The *K*xDG motif is not present in cyclases, and we have found recently that the adenylyl group transfer in the *E. coli* enzyme is transferred not to a lysine but to a histidine [26].

The precise biological role(s) of cyclases is unknown. Requirements for 2′,3′-cyclic phosphate ends in the substrates of several eukaryotic and prokaryotic RNA ligases suggest that the enzymes are involved in the generation (or regeneration) of cyclic termini in RNA molecules undergoing ligation. The cyclase could also be responsible for producing 2′,3′-cyclic ends identified in the spliceosomal U6 RNA [27] and few other small RNAs [28]; for a discussion of additional possible functions of cyclases, [21]. The conservation of cyclases among Eucarya, Bacteria and Archaea argues that the enzymes must perform an important function in RNA metabolism. This view is supported by the findings that the gene encoding cyclase in the yeast *Saccharomyces cerevisiae* is essential for growth (E.B. and W.F.; **unpublished observations). [AU: We allow 'unpublished observations' or 'data not shown', which would you prefer here?]**

Phylogenetic analysis of the *E. coli* and human cyclases, and cyclase-like proteins encoded in other organisms, indicated that cyclases can be subdivided into two classes (W.F. and E.B.; **unpublished observations) [AU: We allow 'unpublished observations' or 'data not shown', which would you prefer here?]**. Members of class I include all prokaryotic proteins, the *Dictyostelium discoideum* protein, and one of the two proteins expressed in *Drosophila* and humans (Figure 1). The previously characterized *E. coli* and human cyclases belong to this class of enzymes. Class II proteins are encoded by the genomes of budding and fission yeast, as well as *Caenorhabditis elegans*, and are second forms of proteins expressed in *Drosophila* and humans.

To gain further insight into the structure and biology of the RNA 3′-phosphate cyclases, we have determined the crystal structure of the *E. coli* enzyme, a member of the class I family, and refined it with 2.1 Å resolution data. The structure and its implications for the catalytic activity are discussed here.

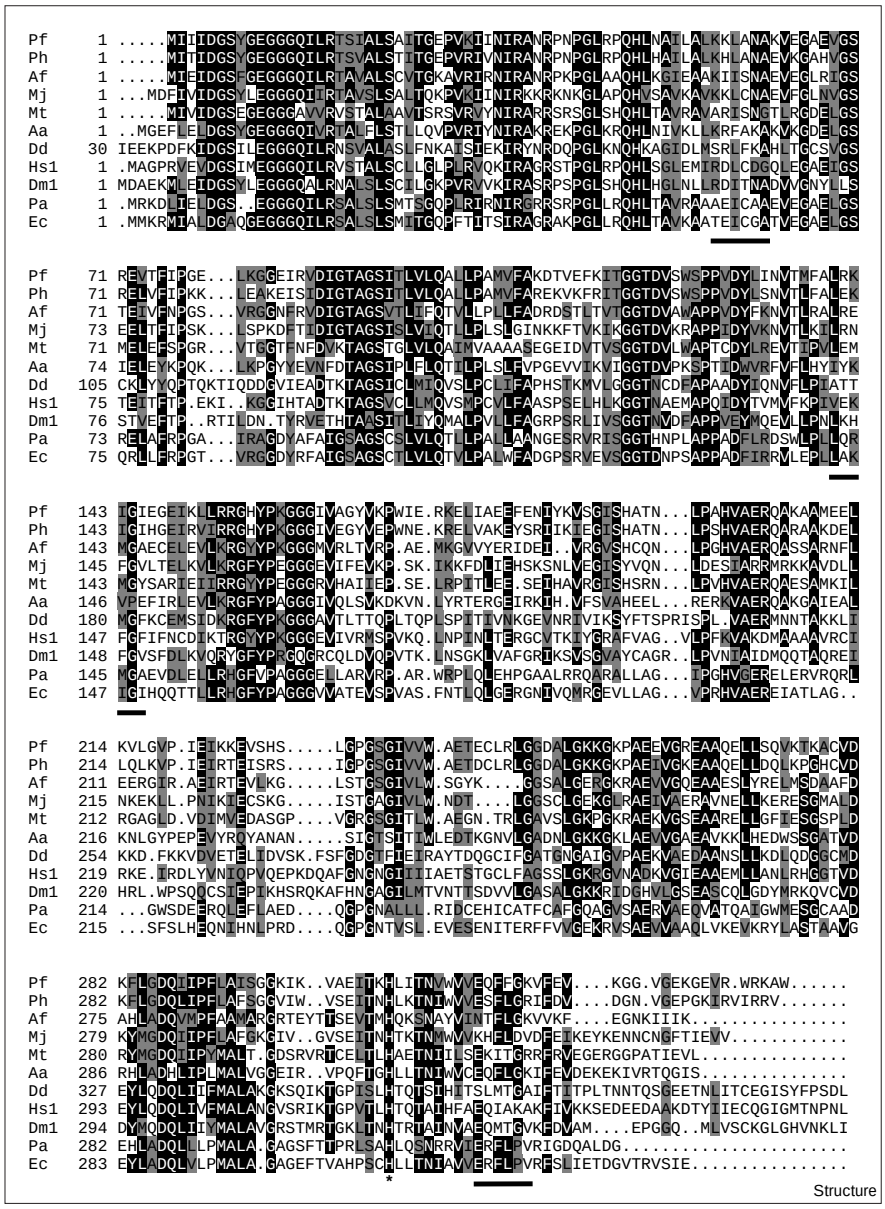
Results

Solution of the crystal structure

RNA cyclase from *E. coli* was crystallized in several forms. Crystals of the selenomethionine (SeMet) protein in form I (see the Materials and methods section) were used to solve the structure using multiwavelength anomalous diffraction (MAD). Crystals of form II of the native protein diffracted better **than those of form I [AU: OK?]** and were used to refine the structure to R = 20.4% (R_{free} = 27.6%) at 2.1 Å **resolution [AU: OK?]**. All detailed structural descriptions are on the basis of this crystal form. A molecular replacement solution was obtained for crystal form III, proving that the quaternary structure of dimers observed in the first two crystal forms is conserved also in this crystal form.

The two monomers found in the asymmetric unit of form II crystals produce a disulfide-linked dimer, through Cys308 of each molecule. The electron density corresponding to this link is very clear and its presence in the crystals is unambiguous (Figure 2a). Only residues 5–338 were included in the model for both monomers, **without the four residues at the N terminus or the C-terminal His**

Figure 1



Alignment of proteins in the cyclase class I subfamily: Pf, *Pyrococcus furiosus* (The Institute for Genomic Research, Gaithersburg, MD); Ph, *Pyrococcus horikoshii* (determined by the Entrez database as [Au: OK? or please suggest and alternative] BAA30639); Af, *Archaeoglobus fulgidus* (Entrez AAB89810); Mj, *Methanococcus jannaschii* (SP Q60335); Mt, *Methanobacterium thermoautotrophicum* (Entrez AAB86375); Aa, *Aquifex aeolicus* (Entrez AAC06852); Dd, *Dictyostelium discoideum* (Entrez AAB70847; this protein contains 29 additional N-terminal amino acids, which are not shown); Hs1, *Homo sapiens* (SP O00442); Dm1, *Drosophila melanogaster* (SPTREMBL O77264); Pa, *Pseudomonas aeruginosa* (The Pseudomonas Genome Project; http://www.pseudomonas.com); Ec, *E. coli* (SP P46849). Multiple sequence alignment was performed with the Clustal W1.5 program [62] using the complete multiple alignment protocol with default parameters. The alignment was improved manually. Identical amino acids and amino acids conserved in at least 50% of sequences are indicated by black and gray boxes, respectively. The histidine that undergoes adenylation (His309 in the *E. coli* protein) is marked with an asterisk, and hexameric sequences that correspond to the motifs Lxx(L)G(A), previously identified in MurA and EPSP synthase (see text), are underlined.

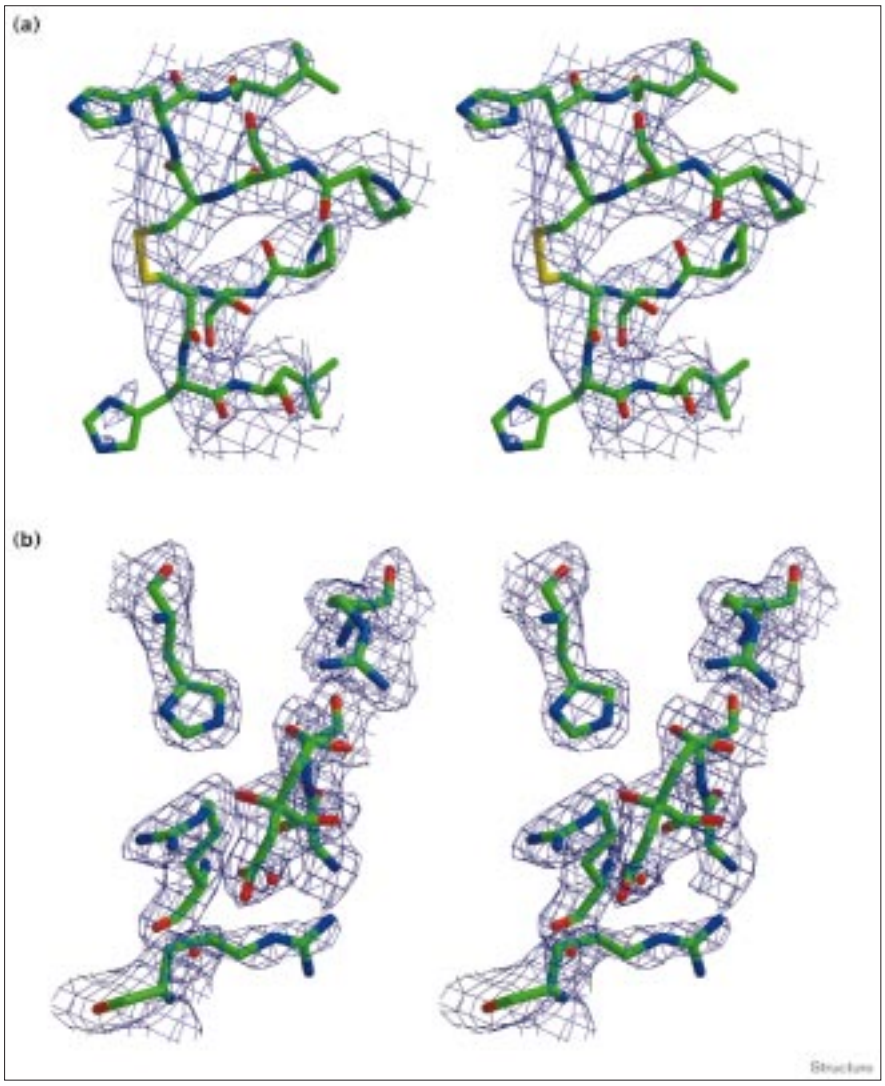
tag [AU: OK?]. Two citrate molecules, one of which is shown in Figure 2b, and a DTT moiety were modeled in addition to the solvent molecules. The Ramachandran plot for the model is of high quality, with the exception of Ser95, which is part of one of the flexible loops described below.

Overall structure of RNA cyclase

A molecule of cyclase consists of two structurally distinct domains, connected by two polypeptide chains (Figure 3a). The larger domain (residues 1–184 and 280–339) consists of three repeats of a folding unit comprising two α helices and a four-stranded β sheet in +1X, +2X,–1 topology [29], with one of the parallel helices in each of the crossover connections. This type of folding motif is usually associated with the bacterial translation initiation factor 3 (IF3), in which it comprises its C-terminal domain [30]. The superposition of cyclase residues 5–87 on IF3 residues 84–170 yields a root mean square (rms) deviation of 3.0 Å for 69 common Cα pairs.

The smaller domain of cyclase, residues 185–279, comprises the same secondary-structure elements — a four-stranded β sheet covered by two parallel helices — but their connection topology (+1X, –2, –1) is different, corresponding to the fold previously observed in human thiore-

Figure 2

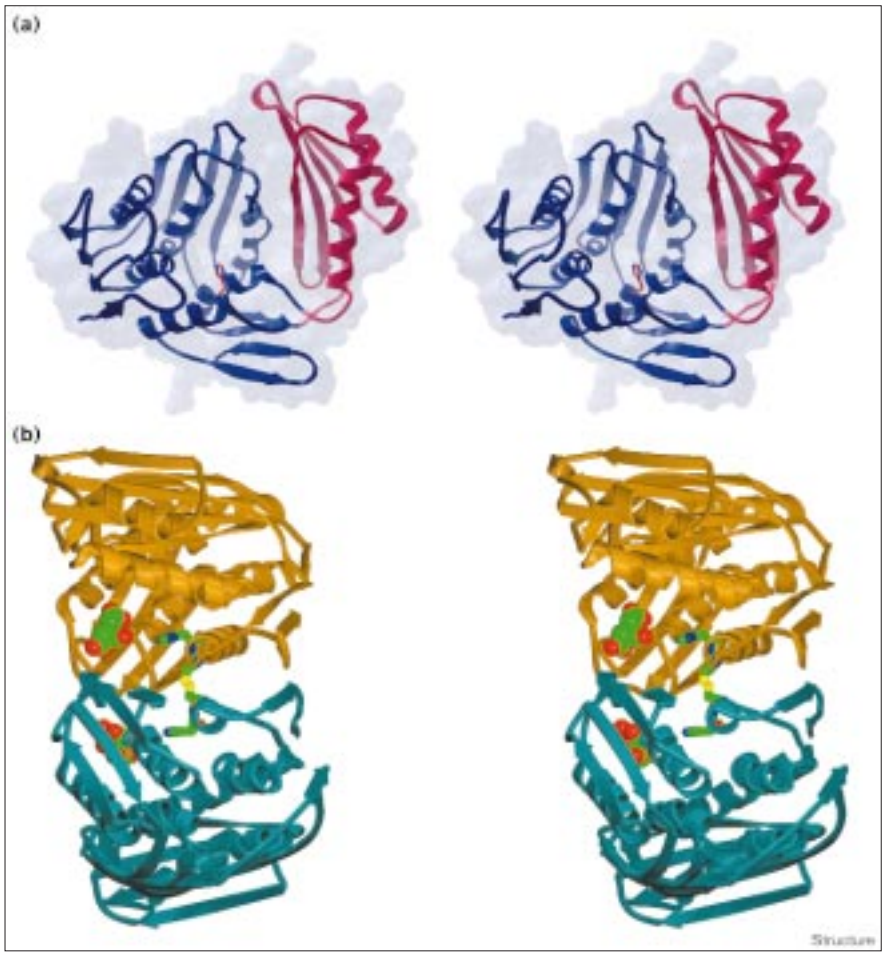


Representative electron-density maps. (a) The original MAD map in the vicinity of Cys308, contoured at the 1.0σ level, clearly shows the disulfide bridge. Noncrystallographic averaging was not used in the calculation of the map shown here. (b) The final, refined 2.1 Å map ($2F_o–F_c$) in the vicinity of the citrate molecule, contoured at the1.0σ level. Note the well-coordinated C-terminal groups of the citrate molecule, which might mimic the backbone of substrate RNA. Atoms are shown in standard colours [Au: OK? or please explain what each colour denotes]

doxin [31]. The superposition of cyclase residues 185–279 on thioredoxin residues 19–105 yields a root mean squared deviation of 2.8 ¯ for 69 Cα pairs. Other proteins utilizing this motif are porphobilinogen deaminase [32], glutathione S-transferase [33] and riboflavin synthase [34]. The small domain is inserted between an α helix and a β strand from the third IF3-like repeat in the large domain (Figure 4).

The finding that in all crystal forms analyzed so far, *E. coli* RNA cyclase is present as a covalent, disulfide-linked dimer, was unexpected. Previous studies performed with the purified human enzyme, which shows 32.5% sequence homology [AU: OK?] to *E. coli* enzyme and has very similar requirements and substrate specificity [17,21], demonstrated that it is present in solution in the form of a monomer [20]. Nevertheless, in the experimentally determined map, the electron density for the disulfide bridge consisting of Cys308 from two molecules is absolutely clear (Figure 2a). A dimer was also indicated by dynamic light scattering (A.W.; unpublished observations). The disulfide bridge must be unique to *E. coli* cyclase, even if the quaternary structure is conserved in cyclase from other organisms because Cys308 is not conserved in the sequences of other class I cyclases (Figure 1). Although we cannot rule out the possibility that the disulfide bond is an artifact of protein purification and crystallization, the identical quaternary structure was observed in four crystallographically independent dimers occurring in three crystal forms under various conditions. Given that crystal forms I and II are closely related, the arrangement of the dimers is therefore similar in both forms, but nevertheless more divergent than would be expected for stable higher oligomers.

Figure 3



Stereo diagrams of the backbone of RNA cyclase. (a) Ribbon representation of the crystal structure of a single cyclase molecule and the atomic surface of the protein. The chain corresponding to the large domain is shown in blue and the small domain is shown in pink. The sidechain of His309 is shown in red. (b) A noncrystallographic dimer of RNA cyclase, with two molecules shown in green and gold. Sidechains of His309 are rendered in atomic colors and the citrate molecules are shown in space-filling representation.

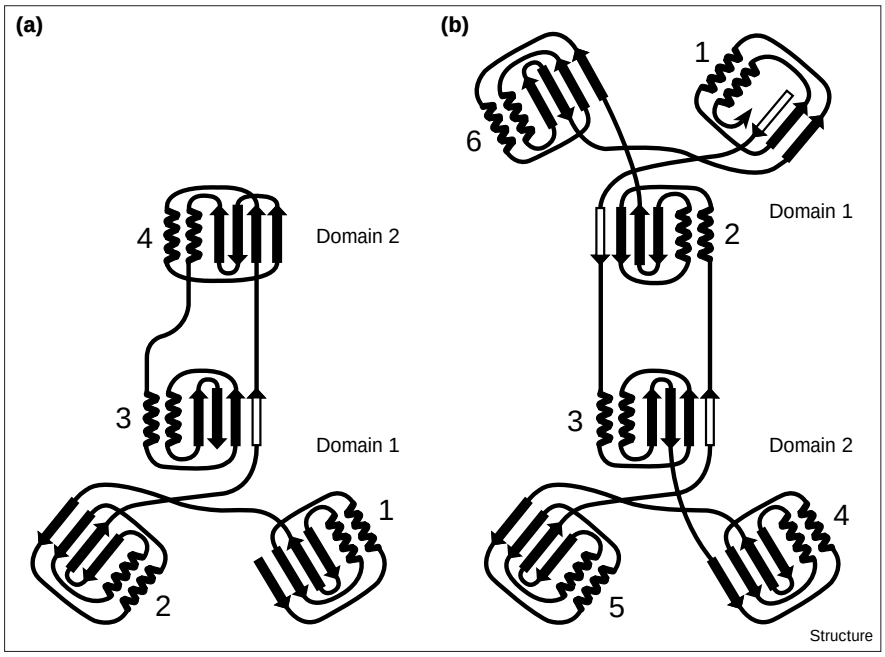
The two monomers of the dimer are related by a dyad that can be noncrystallographic (as in crystal forms I and II) or crystallographic (as in crystal form III). In crystal form II, the two monomers (Figure 3b) have an rms deviation of 0.40 Å between their Cα atoms. Similar comparisons would be meaningless for crystal form I because of strong non-crystallographic symmetry (NCS) restraints utilized during refinement. The rms deviations between different monomers found in crystal forms I and II are, on average, 0.56 Å. The rms deviation between the dimers, 0.65–1.20 Å, shows that the dimer interface is fairly rigid and specific, despite its rather small area (~680 Å^2). The interface is formed by β strands (residues 324–331) [AU: OK?] creating an eight-stranded β sheet and by three loops (residues 12–14, 42–44, and 305–311), and is mainly hydrophobic. Its size is more comparable to the usual crystal contacts rather than to tight domain interfaces.

Location of the active site

The identity of the residues forming the active site of cyclase is largely unknown at this time. Sequencing of proteolytic fragments of cyclase adenylylated with [^{32}α-P] ATP identified His309 as the labeled residue [26], leading to its assignment to the active site. Covalent nucleotidation of lysines and other amino acids has been described either as part of the catalytic mechanism of nucleotidyl transferases such as RNA-capping enzymes and DNA and RNA ligases (reviewed by [25]), or as a mechanism to regulate the activity of proteins (reviewed by [35,36]), and utilization of a histidine for these purposes has also been reported [37,38]. In class I cyclases, His309 is completely conserved, although it is not conserved in the structurally related type II enzymes [26].

When viewed in the three-dimensional structure, His309 is located on the bottom of a large cleft and is surrounded by a number of other strictly conserved residues (Figure 5). Facing His309 are the five loops with the highest B factors in the entire structure (mainchain atom B factors > 50 Å^2: 9–16, 40–47, 93–97, 255–257, 328–333). At the same time, many residues of these loops are highly conserved (28% totally conserved residues compared with

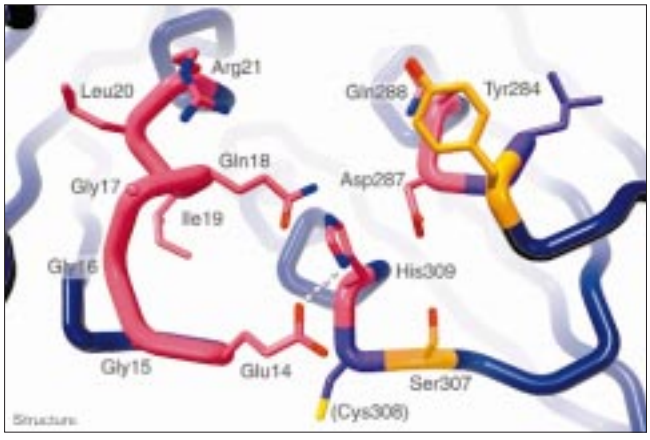
Figure 4



Topology diagram of RNA cyclase and a comparison with EPSP synthase [42]. Secondary structure elements forming compact folding units are solid and those not adjacent in the primary structure are marked as open arrows. (a) Two domains of RNA cyclase, with the larger domain 1 including both the N and C termini. (b) EPSP synthase, adapted from [42]. Although the domain designation of the two proteins is reversed, the similarity of domain 1 of cyclase to domain 2 of EPSP synthase, including the location of the non-contiguous secondary structure element, is clear.[Au: do we need permission to use part (b)?]

14% overall in class I cyclases). Other loops facing the solvent are much better defined. One of the flexible loops includes the stretch 14–21 (EGGGQILR), which is almost entirely conserved in the 11 known sequences of class I cyclases. Glu14 in this stretch appears to form a strong hydrogen bond with the Nδ1 of His309 (distance ~2.7 Å), and Gln18, a very weak one with Nε2. Other neighboring residues include the mostly conserved Tyr284 (replaced only by other aromatic residues, Phe or His), strictly conserved Asp287 (which is on the opposite side of the ring of His309), Ser307 (its only replacement is a threonine) and slightly less conserved Asn313. The citrate bound in crystal form II is liganded by Gly17, Arg21, Arg40, Arg43, Gln51 and His52.

Figure 5



Stereo diagram of the putative active site of RNA cyclase. Conserved residues found within 10 Å of His309 are shown in red and the mostly conserved Tyr284 and Leu285 residues are in gold. The sidechain of the non-conserved Cys308, which makes a disulfide bridge, is also marked. [AU: The figure you sent electronically is different from the hard copy. Is that OK?]

Comparison with other enzymes

A search of the database of all unique structures performed with the program DALI [39] did not identify any other proteins with the complete fold corresponding to that of cyclase. Each individual domain of the cyclase, however, as defined above, has shown similarity to at least some domains found in other proteins. Most striking is the similarity of the larger domain of cyclase to a single domain found in two structurally related proteins, UDP–N-acetylglucosamine enolpyruvyl transferase (MurA) [40,41], and 5-enol-pyruvylshikimate-3-phosphate synthase (EPSP synthase) [42]; see Figure 4). Each of these proteins consists of two domains that are topologically equivalent and structurally similar to each other [AU: OK?]. Thus, each of the domains of MurA and EPSP synthase is not only completely topologically equivalent to the large domain of cyclase but also similar in direct comparison of the coordinates. The rms deviation between the Cα coordinates of residues 22–216 in MurA [41] and residues 5–184 and 280–328 in cyclase is 3.0 Å for 185 atom pairs, and a similar comparison with the recently refined coordinates of EPSP synthase (K Brown, personal communication) [AU: do you have permission to cite this?] yielded an rms deviation of 2.9 Å for 186 atom pairs.

A common feature observed, without exception, in all six IF3 motifs in both MurA and EPSP synthase is a hexapeptide with the sequence Lxx(L)G(A), connecting these motifs into trimers [40–42]. Although leucine and glycine (the latter with positive ϕ/ψ angles) are always found in positions 1 and 5, respectively, positions 4 and 6 are seldom mutated, and positions 2 and 3 are usually occupied by hydrophobic residues. A comparison of the structurally aligned cyclases shows that a similar structure is found there in at least two out of three IF3-like motifs. In motif 1, the corresponding sequence is TEICGA (residues 60–65) and it superimposes very well on its counterparts in MurA and EPSP synthase. This sequence does not appear to be well conserved in other class I cyclases; in particular, glycine is substituted for the most part by asparagine (Figure 1). The alignment of the sequence found in the motif 2 is shifted by one residue but the sequence itself still has some visible similarity to that of motif 1 [AU: OK?]. The cyclase residues 144–149 are LAKIGI and the glycine torsion angles are similar to those found in the equivalent peptides in the other two proteins. Conservation of this sequence in other class I cyclases is quite high, with glycine present in all of the cyclases except for one [AU: OK?]. Finally, the ERFLPV sequence present in a similar position in motif 3 (residues 318–323) is quite different. It does, however, have some similarity, given that the replacement of glycine by proline is quite common, and the nature of most other residues in this sequence bears at least some relation to their counterparts in the other motifs. In almost all of the other class I cyclases, the proline residue is indeed aligned with glycine.

Discussion

Comparison with other enzymes

The primary structures of the human and *E. coli* cyclases have been determined only recently [17,21]. Although cyclase homologs have been found in organisms belonging to all kingdoms, no other proteins seemed to be obviously related to cyclase [AU: OK?]. Thus, no structure could be predicted, making cyclase an interesting target for structure determination by experimental means. It was also hoped that the availability of the structure would provide some clues to the function of the protein, as has been shown in other cases [43].

As expected, the complete fold of cyclase does not correspond to any known proteins. Because the collection of the protein domains is being completed faster than that of the folds, however, it is not surprising that both domains of cyclase have been identified as parts of other proteins. Most prominently, the large domain with approximately 240 residues appears twice in each of two related enzymes, MurA and EPSP synthase. The completely conserved topology of this rather large domain clearly indicates a common origin/ancestor, but domain arrangement and substrates have since diverged. The monomers in cyclase and in the two domains in the other two enzymes create different clefts containing the putative or actual active sites. The substrates, too, are very different: ATP and RNA for cyclase; phosphoenolpyruvate (PEP) and UDP–GlcNAc for MurA; and PEP and 3-phosphoshikimate for EPSP synthase. The strong divergence is obvious at the sequence level. Only careful searches performed with Ψ–BLAST [44] reveal a weak similarity between cyclase and EPSP synthase. In a recent structure prediction exercise with knowledge of only the sequence, the best models built had an rms deviation of 4 Å for only 100 amino acids [45], but the choice of a proper lead model was by no means obvious. The small domain is structurally similar to domains in riboflavin synthase, glutathione S-transferase and porphobilinogen deaminase. Similarly to what was observed for the large domain, none of these structurally related enzymes is functionally related to cyclase.

As outlined above, both domains of cyclase in their present form have diverged so far from proteins with similar domains that functional relationships have been lost. The larger domain is built up by a three-fold repeat of an ancestral structural motif that appears in several nucleic acid binding proteins: C-terminal domain of IF3 [30], DNase I [46], DNA methyltransferase [47], the N-terminal domain of ribosomal protein S8 [48] and the single-stranded nucleic acid-binding motif R3H that has a characteristic Arg/His pair separated by three residues [49]. The course of evolution could thus have proceeded from an RNA binding protein with one repeat to cyclase with three repeats and finally to MurA/EPSP synthase with six repeats and with new substrates.

Active site of cyclase

The location of the active site of cyclase cannot yet be specified precisely, because the structure presented here does not include any bound substrate. We predict that the active site will cover a fairly large area, for at least two reasons. First, the RNA substrate is itself rather large (at least three nucleotides are required), and second, possibly different residues might be involved in catalyzing at least two separate reactions, the first and second steps, outlined in the Introduction. Analysis of the adenylated protein showed that the reaction product of the first step, adenylated His309, is part of the active site. This residue is located at one end of a long cleft located between the large and the small domains of the enzyme. All of the surrounding conserved residues belong to the same molecule of cyclase (Figure 5), although the putative active site is in the vicinity of dimer interface created through the formation of an intermolecular disulfide bond by Cys308 of two molecules. This bond might be the native redox state of the cysteines, even in the reducing environment of the cytosol, because it is observed despite purification and crystallization in the presence of DTT. Several protein

loops located in the vicinity of His309 contain many conserved residues, but also display above-average B factors. These flexible parts of the enzyme might become well ordered only upon binding of the substrate(s) and might be necessary to move the products to new locations, to become substrates in subsequent steps of the overall reaction.

The citrate molecule bound to cyclase in crystal form II (Figures 2a,3b) is coordinated by three arginines, a histidine or glutamine, and a mainchain nitrogen, fitting this site very well. The distance between the two terminal carboxylates (5.2 Å) might mimic the spacing of the phosphate groups in an RNA backbone (5.9 Å), but we cannot postulate which part of the substrate might be mimicked by the citrate. The distance between His309 and the closest carboxylate of the citrate is approximately 6 Å. In addition, the second step of reaction involves not a histidine, but rather its adenylated form, making more precise assignments of the location of the groups participating in the reaction impossible. Only by determining the structure of an RNA complex of cyclase will it be possible to delineate the precise location of the substrate.

Indirect evidence supporting the location of the active site in the large, open channel that includes His309 is provided by the comparison of cyclase with the structurally related enzymes MurA and EPSP synthase. The large domain of cyclase can be superimposed on the catalytic domain of MurA such that the two ligands of the latter enzyme, the inhibitor fosfomycin and UDP–GlcNAc [41], bind to the enzyme's surface corresponding to the general area of His309 (although the pseudo three-fold symmetry axis within the domain allows for three different equivalently valid superpositions). It is commonly accepted that within superfolds, binding sites are similar even in the absence of homology [50], and even more so when weak, but discernible homology is present, as is the case here. Although no ligand-bound structures of EPSP synthase are available at this time, it is also probable that its catalytic site is located in the equivalent area (K Brown; personal communication) [AU: Do you have permission to cite this?]. It should also be pointed out that both MurA and EPSP synthase, in common with cyclase, bind multiple phosphate groups during their catalytic cycles. These indirect indications support the proposed location of the active site of cyclase, even though they do not provide complete proof of such an assignment.

E. coli cyclase is the only protein that has been unambiguously shown to be adenylated on a histidine in a stable manner, and to use ATP as a cofactor [26]. Enzymes from the histidine triad (HIT) family have been shown to form covalent adenylated intermediates; the involvement of a histidine in the linkage has not been demonstrated directly (although structures of transition-state analogs involving adenosine–tungstate covalently attached to histidine have been described [38]). Three other proteins known to undergo modification on a histidine residue with nucleotidyl groups other then adenylyl are galactose-1-phosphate uridylyltransferase [37,51], the Gag protein of the *S. cerevisiae* double-stranded RNA L–A virus [52] and the brine shrimp GTP–GTP guanylyltransferase [53]. The tertiary structure has been established only in the case of the galactose-1-phosphate uridylyltransferase. This enzyme, which is involved in the Leloir pathway for galactose metabolism, is transiently uridylated on Nε2 of the imidazole ring of a histidine; UDP–glucose is a uridyl group donor in this reaction [37,51].

An analysis of the residues interacting with the nucleotide in the modified galactose-1-phosphate uridylyltransferase points to some potentially important similarities with cyclase. In particular, one of the phosphate oxygens interacts with a sidechain of a glutamine. We speculate that Gln18, a residue completely conserved in class I cyclases, might play a similar role because it is located in the vicinity of the adenylation site but does not show strong interactions with His309 in the apoenzyme. It can be shown by modeling that if the binding of an adenylate group is similar to that of uridylate in galactose-1-phosphate uridylyltransferase, Gln18 could reorient and make equivalent interactions. The residue interacting directly with the histidine sidechain is serine in galactose-1-phosphate uridylyltransferase, which might be substituted in cyclase by the similarly conserved Glu14 (Figure 5).

A question that remains unanswered is the structural relationship between class I and class II cyclases. Although the sequence comparison of these two classes shows that they share similar overall structure, the most striking difference between them is the lack of conservation of His309. Preliminary modeling of class II cyclases based on the structure of class I *E. coli* cyclase described here has shown the feasibility of creating reasonable models, with the most difficulty encountered in modeling the third folding unit of domain 1 (Figure 4), which includes His309. The remaining questions about the enzymatic mechanism of cyclases can, therefore, be answered only by an experimentally determined structure of a class II cyclase, as well as structures of substrate complexes of both class I and class II proteins.

Biological implications

Anabolic function of the 2′,3′-cyclic phosphate in RNA first emerged when it was found that eukaryotic and prokaryotic RNA ligases require 2′,3′-cyclic ends for RNA ligation. Such ends can be produced by the cleavage of longer RNA molecules by endonucleases, or by ATP-dependent conversion of the 3′-monophosphate at the end of RNA into the cyclic diester. The latter reaction is catalyzed by the RNA 3′-terminal phosphate

cyclase, an ubiquitous enzyme conserved among eucarya, bacteria and archaea. The crystal structure of *Escherichia coli* cyclase reported here, indicates that cyclases are not structurally related to RNA processing enzymes such as RNA and DNA ligases and capping enzymes. Despite this, mechanistically, reactions catalyzed by these enzymes and the cyclase are similar. In contrast, and rather unexpectedly, the cyclase has common origin with two enzymes that are not involved in RNA metabolism: MurA and EPSP synthase. These two enzymes, however, are similar to cyclase in that all of them bind multiple phosphate groups during their catalytic cycles.

Identification of the probable active site of RNA cyclase, which includes the conserved amino acid residues His309, Glu14, Gln18 and Asp287, will greatly facilitate further investigation of the function of the protein, by construction of negative dominant mutants and other approaches. A long cleft adjacent to these conserved residues forms the putative binding site for the large RNA substrate. The established structure of the *E. coli* protein, a class I enzyme, will also permit modeling and rational mutagenesis of class II cyclases that have so far resisted crystallization (AW; unpublished observations). This should allow us to directly test whether or not the observed essential presence of the class II proteins for growth in the yeast *Saccharomyces cerevisiae* (EB and WF; unpublished observations) is indeed dependent upon enzymatic competence of the protein.

Materials and methods

Protein expression, purification and crystallization

E. coli cyclase was bacterially overexpressed as a fusion protein with a GSHHHHHH peptide at the C terminus, using a pET-11d vector-based expression plasmid [17,21]. The overexpressed protein was purified using metal affinity chromatography, using a Ni^{2+}-NTA column (Qiagen), and desalted on a PD10 column into 50 mM Hepes, pH 7.6, 100 mM NaCl, 0.1 mM EDTA, 0.5 mM DTT, 5% glycerol. The protein solution was stored and transported at –70°C. SeMet-labeled protein was expressed from the same vector, but using B834(DE3) cells (methionine auxotroph). The change to the SeMet expression system lowered the protein yield at least 50-fold to 0.2 mg of purified protein per liter of cell culture.

Crystallizations were carried out using the hanging-drop vapor diffusion technique at 20°C. The stock protein solution was concentrated to approximately 15 mg/ml. The well solution contained 13–15% MPEG 2000, 200 mM Na citrate, pH 4.0, 200 mM Tris-HCl, pH 8.0, 2 mM DTT (resulting pH ~5.5). Drops of 2 µl protein solution plus 2 µl of well solution were equilibrated against the well solution. Two closely related, orthorhombic crystal forms of similar morphology were observed growing together. One of them (form I) was later identified as belonging to space group $P2_12_12_1$, a = 101.8 Å, b = 126.6 Å, c = 128.8 Å, with four molecules in the asymmetric unit, whereas the other (form II) was in space group $P2_12_12$, a = 125.8 Å, b = 133.5 Å, c = 51.0 Å, with two molecules in the asymmetric unit. Because it was not possible to distinguish between these two crystal forms on the basis of the appearance only, each crystal used for data collection had to be characterized separately. The buffers in the well solution could be replaced by high salt (600–900 mM NaCl, 300 mM ammonium sulfate, 800 mM KCl, or 400 mM LiCl), giving lower quality crystals in yet other space groups (in particular, form III, grown with NaCl, belongs to space group $C222_1$ with the unit cell parameters a = 47.9 Å, b = 92.2 Å, c = 144.0 Å). The addition of Mg^{2+}, known to be required for the catalytic activity of RNA cyclase, did not have any visible effect on crystal growth. Attempts to cocrystallize cyclase with ATP analogs and/or with oligoribonucleotide substrates have so far not been successful, nor could such complexes be made by soaking the crystals.

Table 1

Results of crystallographic data collection and processing, and phasing statistics.

	Selenomethionine derivative			Native
	Edge	Peak	Remote	
Wavelength (Å)	0.97981	0.97942	1.02012	0.96456
No. of reflections‡	161,894 (15,228)	162,741 (15,313)	154,068 (11,162)	159,690 (11,690)
Unique Reflections*†	41,451 (4,084)	41,596 (4,074)	40,955 (3,233)	50,057 (4,214)
Resolution (Å)	20.0–2.8	20.0–2.8	20.0–2.8	20–2.1
R_{merge}*,‡	0.043 (0.209)	0.045 (0.198)	0.042 (0.224)	0.043 (0.268)
Completeness (%)*,†,§	99.9 (99.8)	99.9 (99.7)	97.8 (78.7)	97.2 (83.4)
power#/R_{Cullis}¶	96.8 (91.4) 1.44/0.628	96.9 (91.6) 1.77/0581	94.7 (72.2)	93.6 (76.5)Ph

Semet derivitive crystals were in space group $P2_12_12$ with unit-cell dimensions a=101.8, b=126.6, c=128.8 and for molecules per asymmetric unit. The native crystals were in space group $P2_12_12$ with unit-cell dimensions a=125.8, b=133.5, c=51.0 and two molecules per asymmetric unit. Mean overall figure of merit 0.572 (0.817 after solvent flattening). †Reflections with I/σ(I) > –3.0. §Reflections with I/σ(I) > 0. *Last shell (2.9–2.8 Å for crystal form I, 2.17–2.1 Å for form II). ‡$R_{merge} = \Sigma_h\Sigma_i|I(h,i) - \langle I(h)\rangle|/\Sigma_h\Sigma_i I(h,i)$, where I(*h*,*i*) is the intensity of the i^{th} measurement of *h* and <I(*h*)> is the corresponding value of I(*h*) for all *i* measurements. #Phasing power = $\langle F_H\rangle/E$, where $\langle F_H\rangle$ is the rms heavy-atom structure factor and E is the residual lack-of-closure error (the remote data set is considered to be native). ¶$R_{Cullis} = \Sigma||F_{PH} \pm F_P| - |F_{PH(calc)}||/\Sigma|F_{PH} - F_P|$, where F_{PH} and $F_{PH(calc)}$ are the observed and calculated structure factors of a heavy-atom derivative. [Au: are the changes to the table and foonote OK?]

Data collection

Diffraction data were collected on beamline X9B at the NSLS synchrotron source at Brookhaven National Laboratory. Reflection intensities were measured with a MAR345 image plate detector and processed with the HKL2000 suite of programs [54]. Prior to data collection, the crystals were rapidly frozen in a stream of cold nitrogen (100 K) provided by an Oxford Cryosystems device, after being briefly drawn through the mother liquors with half and full concentrations of cryoprotectant (13% MPEG2000 replaced by 26% PEG200). MAD data were collected to a maximum resolution of 2.8 Å from a single, randomly oriented form I crystal at four wavelengths in 97.5° wedges each: 0.97981 Å (edge), 0.97842 Å (peak), 1.02021 Å (remote), and 0.96 Å (second remote). Only a single data set was collected to the maximum resolution of 2.1 Å from a form II crystal at the wavelength 0.96456. The details of data collection and processing are summarized in Table 1. A further data set, not used for subsequent refinement, was collected for a form III crystal to 2.6 Å resolution (Rigaku RU-200 generator, Bruker HiStar multiwire detector, room temperature, R_{sym} 12.8%, 80.4% complete). Data for a native form I crystal were also collected (MAR345 at Brookhaven, beamline X9B, 100 K, 3.0 Å resolution, 93.2% complete, R_{sym} 6.9%), whereas each of the seven useful SeMet cyclase crystals characterized during the course of this work belonged to form I (no form II crystals were ever found for the modified protein).

Structure solution

The structure was solved using multiwavelength anomalous diffraction of the SeMet derivative. Comparison of the different crystal forms clearly suggested a tetramer in the asymmetric unit of form I (V_M = 2.74 Å^3/Da). Excluding the N-terminal methionine of each chain, 16 selenium atoms were thus expected in the SeMet derivative (in place of residues Met5, Met28, Met191 and Met292). SHELXS [55] was used for a direct methods search of the Se sites in the anomalous and dispersive signal. The dispersive signal between either $\lambda_{edge}/\lambda_{low\ energy\ remote}$ or $\lambda_{peak}/\lambda_{low\ energy\ remote}$ yielded 12 sites. These three wavelengths were used for all further calculations in PHASES [56]. Difference Fourier maps based on the first 12 sites revealed four more sites. The 16 sites could be grouped into four repeating units with four selenium atoms each, confirming the four molecules per asymmetric unit. They also obeyed the non-crystallographic translation (0.500, 0.000, 0.042) predicted by a strong non-origin peak (35% of the origin peak) in the native Patterson map. The deviation from higher symmetry (as expected for perfect (1/2, 0, 0) translation) is sufficient to obscure any trends in the intensity distribution. Solvent flattening and fourfold noncrystallographic averaging resulted in interpretable maps (Figure 2a). All but the first four and the last nine residues (6× His-tag and a connecting linker) could be traced.

Table 2

Crystallographic refinement statistics.

Space group	$P2_12_12_1$	$P2_12_12$
Resolution (Å)	10.0–2.8	20.0–2.1
Refinement program	X-PLOR 3.1	REFMAC
$R_{cryst}(R_{free})$	0.230 (0.331)	0.204 (0.276)
Protein atoms	9976	4973
Water molecules	–	424
Ligand atoms	–	35
Rmsd bonds (Å)	0.018	0.022
Rmsd angles (Å)	2.10	3.73

Refinement

The protein model was refined with X-PLOR 3.1 [57] from R = 48.8% (R_{free} = 48.1%) to R = 23.7% (R_{free} = 33.2%) using non-crystallographic restraints (weight 100) (Table 2). At this point, no significant further improvement was expected and the coordinates of one monomer were used as a molecular replacement model for solving the structure of crystal form II, to enable refinement with higher-resolution data. A straightforward solution was obtained in AMoRe [58] (data used 15–3.5 Å, correlation coefficient 64.5%, R factor 34.5%; wrong solutions were eliminated in the rotation function search, because their rotation function values were less than 50% of the correct solution). The resulting model for form II with the two crystallographically independent monomers has the same quaternary structure as the dimers in form I. Refinement of this model was performed with the program REFMAC [59]. Water molecules were added according to the F_o–F_c map. NCS restraints were omitted in the later stages of the refinement. A citrate ion was identified in each monomer close to Arg21, Arg40, Arg43, Gln51 or His52, and the mainchain nitrogen of Gly17. Another contiguous stretch of density could tentatively be described as oxidized DTT, but was seen only in one monomer. This led to a model with R = 20.4% (R_{free} = 27.6%) with good geometry (over 93% of all residues in the most favorable region in the Ramchandran plot, with only Ser95 in the generously allowed/disallowed region as assigned by PROCHECK [60]) (Table 2). The parts of the structure found in the lowest-quality regions were the loops close to the internal dyad around residues 14, 42 and 330 and two loops facing the cleft (around residues 95 and 256). Omit maps for the region 94–97 resulted in well defined contiguous density; nevertheless, no definite conformation could be modeled from this density, despite numerous trials. The buried surface of the interface was calculated by subtracting the accessible surface of the dimer from the sum of monomer surfaces with GRASP [61], omitting all solvent molecules in the calculations.

The structure of crystal form III was solved with AmoRe [58] (data used 15–3.5 Å, correlation coefficient 64.7%, R factor 36.3%; best wrong solution: correlation coefficient 44.2%, R factor 44.4%), but no further refinement was performed.

Accession numbers

The final coordinates and structure factors have been deposited with the PDB (accession code 1qmi for crystal form I and 1qmh for crystal form II).

Acknowledgements

We thank Z. Dauter and the staff of NSLS for help in data collection on the NIH beamline X9B; K. Brown for the unpublished, refined coordinates of EPSP synthase; E. Carpenter, J. Hofsteenge and D. Hess for stimulating discussions; G. Morris for preparation of Figure 3; J. Collins for preparation of Figures 3b and 5; and A. Arthur for editorial comments. A.W. would also like to thank the Master and Fellows of the Sidney Sussex College and the Department of Biochemistry, University of Cambridge, for a Visiting Fellowship during which tenure this paper was written. Research of A.W. and G.P. was sponsored in part by the National Cancer Institute, DHHS, under contract with ABL. The contents of this publication do not necessarily reflect the views or policies of the Department of Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government.

References

1. Blackburn, P. & Moore, S. (1982). Pancreatic Ribonuclease. In *The Enzymes*. (Boyer, P., ed.) 15 pp. 317-433, Academic Press, New York.
2. Takahashi, K. & Moore, S. (1982). Ribonuclease T1. In *The Enzymes.* (Boyer, P., ed.) 15 pp. 435-468, Academic Press, New York.
3. Westaway, S.K. & Abelson, J. (1995). Splicing of tRNA precursors. In *tRNA: Structure, Biosynthesis and Function.* (Soell, D. & RajBhandary, [AU: OK? Or should there be a space between j and B?] U.L., eds.) pp. 79-92. American Society for Microbiology, Washington, DC.
4. Sekiguchi, J. & Shuman, S. (1997). Site-specific ribonuclease activity

of eukaryotic DNA topoisomerase I. *Mol. Cell* **1**, 89-97.
5. Symons, R.H. (1992). Small catalytic RNAs. *Annu. Rev. Biochem.* **61**, 641-671.
6. Konarska, M., Filipowicz, W., Domdey, H., & Gross, H.J. (1981). Formation of a 2′-phosphomonoester, 3′,5′-phosphodiester linkage by a novel RNA ligase in wheat germ. *Nature* **293**, 112-116.
7. Filipowicz, W., Konarska, M., Gross, H.J., & Shatkin, A.J. (1983). RNA 3′-terminal phosphate cyclase activity and RNA ligation in HeLa cell extract. *Nucleic Acids Res.* **11**, 1405-1418.
8. Filipowicz, W. & Shatkin, A.J. (1983). Origin of splice junction phosphate in tRNAs processed by HeLa cell extract. *Cell* **32**, 547-557.
9. Furneaux, H., Pick, L., & Hurwitz, J. (1983). Isolation and characterization of RNA ligase from wheat germ. *Proc. Natl Acad. Sci. USA* **80**, 3933-3937.
10. Greer, C.L., Javor, B., & Abelson, J. (1983). RNA ligase in bacteria: formation of a 2′,5′ linkage by an *E. coli* extract. *Cell* **33**, 899-906.
11. Schwartz, R.C., Greer, C.L., Gegenheimer, P., & Abelson, J. (1983). Enzymatic mechanism of an RNA ligase from wheat germ. *J. Biol. Chem.* **258**, 8374-8383.
12. Phizicky, E.M. & Greer, C. (1993). Pre-tRNA splicing: variation on a theme or exception to the rule? *Trends Biochem. Sci.* **18**, 31-34.
13. Zillmann, M., Gorovsky, M.A., & Phizicky, E.M. (1991). Conserved mechanism of tRNA splicing in eukaryotes. *Mol. Cell Biol.* **11**, 5410-5416.
14. Filipowicz, W. & Gross, H.J. (1984). RNA ligation in eukaryotes. *Trends Biochem. Sci.* **9**, 68-71.
15. Arn, E.A. & Abelson, J.N. (1998). RNA ligases: function, mechanism and sequence conservation. In *RNA Structure and Function.* (Simons, R.W. & Grunberg-Manado, M., eds.) pp. 695-726, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
16. Sidrauski, C. & Walter, P. (1997). The transmembrane kinase Ire1p is a site-specific endonuclease that initiates mRNA splicing in the unfolded protein response. *Cell* **90**, 1031-1039.
17. Genschik, P., Drabikowski, K., & Filipowicz, W. (1998). Characterization of the *Escherichia coli* RNA 3′-terminal phosphate cyclase and its σ^{54}-regulated operon. *J. Biol. Chem.* **273**, 25516-25526.
18. Arn, E.A. & Abelson, J.N. (1996). The 2′-5′ RNA ligase of *Escherichia coli.* Purification, cloning, and genomic disruption. *J. Biol. Chem.* **271**, 31145-31153.
19. Greer, C.L., Peebles, C.L., Gegenheimer, P., & Abelson, J. (1983). Mechanism of action of a yeast RNA ligase in tRNA splicing. *Cell* **32**, 537-546.
20. Vicente, O. & Filipowicz, W. (1988). Purification of RNA 3′-terminal phosphate cyclase from HeLa cells. Covalent modification of the enzyme with different nucleotides. *Eur. J. Biochem.* **176**, 431-439.
21. Genschik, P., Billy, E., Swianiewicz, M., & Filipowicz, W. (1997). The human RNA 3′-terminal phosphate cyclase is a member of a new family of proteins conserved in Eucarya, Bacteria and Archaea. *EMBO J.* **16**, 2955-2967.
22. Filipowicz, W., Strugala, K., Konarska, M., & Shatkin, A.J. (1985). Cyclization of RNA 3′-terminal phosphate by cyclase from HeLa cells proceeds via formation of N(3′)pp(5′)A activated intermediate. *Proc. Natl Acad. Sci. USA* **82**, 1316-1320.
23. Reinberg, D., Arenas, J., & Hurwitz, J. (1985). The enzymatic conversion of 3′-phosphate terminated RNA chains to 2′,3′-cyclic phosphate derivatives. *J. Biol. Chem.* **260**, 6088-6097.
24. Filipowicz, W. & Vicente, O. (1990). RNA 3′-terminal phosphate cyclase from HeLa cells. *Methods Enzymol.* **181**, 499-510.
25. Shuman, S. & Schwer, B. (1995). RNA capping enzyme and DNA ligase: a superfamily of covalent nucleotidyl transferases. *Mol. Microbiol.* **17**, 405-410.
26. Billy, E., Hess, D., Hofsteenge, J., & Filipowicz, W. (1999). Characterization of the adenylation site in the RNA 3′-terminal phosphate cyclase from *Escherichia coli*. *J. Biol. Chem.* in press. [AU: Any update?]
27. Lund, E. & Dahlberg, J.E. (1992). Cyclic 2′,3′-phosphates and nontemplated nucleotides at the 3′ end of spliceosomal U6 small nuclear RNAs. *Science* **255**, 327-330.
28. Gu, J., Shumyatsky, G., Makan, N., & Reddy, R. (1997). Formation of 2′,3′-cyclic phosphates at the 3′ end of human U6 small nuclear RNA *in vitro*. Identification of 2′,3′-cyclic phosphates at the 3′ ends of human signal recognition particle and mitochondrial RNA processing RNAs. *J. Biol. Chem.* **272**, 21989-21993.
29. Richardson, J.S. (1981). The anatomy and taxonomy of protein structure. [AU: please provide the names of the editors.]In *Advances in Protein Chemistry*. pp.167-339. Academic Press, New York.
30. Biou, V., Shu, F., & Ramakrishnan, V. (1995). X-ray crystallography shows that translational initiation factor IF3 consists of two compact alpha/beta domains linked by an alpha-helix. *EMBO J.* **14**, 4056-4064.
31. Weichsel, A., Gasdaska, J.R., Powis, G., & Montford, W.R. (1996). Crystal structures of reduced, oxidized, and mutated human thioredoxins: evidence for a regulatory homodimer. *Structure* **4**, 735-751.
32. Louie, G.V., *et al*., & Jordan P.M. (1992). Structure of porphobilinogen deaminase reveals a flexible multidomain polymerase with a single catalytic site. *Nature* **359**, 33-39.
33. Reinemer, P., *et al*., & Bieseler B. (1996). Three-dimensional structure of glutathione S-transferase from *Arabidopsis thaliana* at 2.2 Å resolution: structural characterization of herbicide-conjugating plant glutathione S-transferases and a novel active site architecture. *J. Mol. Biol.* **255**, 289-309.
34. Risert, K., Huber, R., Turk, D., Ladenstein, R., Schmidt-Base, K., & Bacher, A. (1995). Studies on the lumazine synthase/riboflavin synthase complex of *Bacillus subtilis*: structure analysis of reconstituted icosahedral beta-subunit capsids with bound substrate analogue inhibitor at 2.4 Å resolution. *J. Mol. Biol.* **253**, 151-167.
35. Stadtman, E.R. (1990). Discovery of glutamine synthetase cascade. *Methods Enzymol.* **182**, 793-809.
36. Blaho, J.A., Mitchell, C., & Roizman,B. (1994). An amino acid sequence shared by the herpes simplex virus 1 alpha regulatory proteins 0, 4, 22, and 27 predicts the nucleotidylylation of the UL21, UL31, UL47 and UL49 gene products. *J. Biol. Chem.* **269**, 17401-17410.
37. Wedekind, J.E., Frey, P.A., & Rayment, I. (1996). The structure of nucleotidylated histidine-166 of galactose-1-phosphate uridylyltransferase provides insight into phosphoryl group transfer. *Biochemistry* **35**, 11560-11569.
38. Lima, C.D., Klein, M.G., & Hendrickson, W.A. (1997). Structure-based analysis of catalysis and substrate definition in the HIT protein family. *Science* **278**, 286-290.
39. Holm, L. & Sander, C. (1993). Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.* **233**, 123-138.
40. Schönbrunn, E., *et al*., & Mandelkow, E. (1996). Crystal structure of UDP-*N*-acetylglucosamine enolpyruvyltransferase, the target of the antibiotic fosfomycin. *Structure* **4**, 1065-1075.
41. Skarzynski, T., Mistry, A., Wonacott, A., Hutchinson, S.E., Kelly, V.A., & Duncan, K. (1996). Structure of UDP-*N*-acetylglucosamine enolpyruvyl transferase, an enzyme essential for the synthesis of bacterial peptidoglycan, complexed with substrate UDP-*N*-acetylglucosamine and the drug fosfomycin. *Structure* **4**, 1465-1474.
42. Stallings, W.C., *et al.,* & Kishore G.M. (1991). Structure and topological symmetry of the glyphosate target 5-enol-pyruvylshikimate-3-phosphate synthase: a distinctive protein fold. *Proc. Natl Acad. Sci. USA* **88**, 5046-5050.
43. Zarembinski, T.I., *et al.,* & 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.
44. Altschul, S.F., *et al*., & Lipman D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389-3402.
45. Koehl, P. & Levitt, M. (1999). A brighter future for protein structure prediction. *Nature* **6**, 108-111.
46. Oefner, C. & Suck, D. (1986). Crystallographic refinement and structure of DNase I at 2 Å resolution. *J. Mol. Biol.* **192**, 605-632.
47. Labahn, J., *et al.,* & Saenger, W. (1994). Three-dimensional structure of the adenine-specific DNA methyltransferase M.Taq I in complex with the cofactor S-adenosylmethionine. *Proc. Natl Acad. Sci. USA* **91**, 10957-10961.
48. Davies, C., Ramakrishnan, V., & White, S.W. (1996). Structural evidence for specific S8–RNA and S8–protein interactions within the 30S ribosomal subunit: ribosomal protein S8 from *Bacillus stearothermophilus* at 1.9 Å resolution. *Structure* **4**, 1093-1104.
49. Grishin, N.V. (1998). The R3H motif: a domain that binds single-stranded nucleic acids. *Trends Biochem. Sci.* **23**, 329-330.
50. Russell, R.B., Sasieni, P.D., & Sternberg, M.J.E. (1998). Supersites within superfolds. Binding site similarity in the absence of homology. *J. Mol. Biol.* **282**, 903-918.
51. Frey, P.A., Wong, L.J., Sheu, K.F., & Yang, S.L. (1982). Galactose-1-phosphate uridylyltransferase: detection, isolation, and

characterization of the uridylyl enzyme. *Methods Enzymol.* 87, 20-36.
52. Blanc, A., Ribas, J.C., Wickner, R.B., & Sonenberg, N. (1994). His-154 is involved in the linkage of the *Saccharomyces cerevisiae* L-A double-stranded RNA virus Gag protein to the cap structure of mRNAs and is essential for M1 satellite virus expression. *Mol. Cell Biol.* 14, 2664-2674.
53. Cartwright, J.L. & McLennan, A.G. (1999). Formation of a covalent Nε2-guanylylhistidyl reaction intermediate by the GTP:GTP guanylyltransferase from the brine shrimp *Artemia. Arch. Biochem. Biophys.* 361, 101-105.
54. Otwinowski, Z. & Minor, W. (1997). Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzymol.* 276, 307-326.
55. Sheldrick, G.M. (1997). Patterson superposition and *ab initio* phasing. *Methods Enzymol.* 276, 628-641.
56. Furey, W. & Swaminathan, S. (1995). Phases-95: program package for the processing and analysis of diffraction data from macromolecules. *Methods Enzymol.* 276, 546-552.
57. Brünger, A.T. (1992). *X-PLOR Version 3.1: System for X-ray Crystallography and NMR.* New Haven, CT, Yale University Press.
58. Navaza, J. (1994). An automated package for molecular replacement. *Acta Crystallogr. A* 50, 157-163.
59. Murshudov, G.N., Vagin, A.A., & Dodson, E.J. (1997). Refinement of macromolecular structures by the maximum-likelihood method. *Acta Crystallogr. D* 53, 240-255.
60. Laskowski, R.A., MacArthur, M.W., Moss, D.S., & Thornton, J.M. (1993). PROCHECK: program to check the stereochemical quality of protein structures. *J. Appl. Crystallogr.* 26, 283-291.
61. Nicholls, A., Sharp, K.A., & Honig, B. (1991). Protein folding and association: insights from the interfacial and thermodynamic properties of hydrocarbons. *Proteins: Struct. Funct. Genet.* 11, 281-296.
62. Thompson, J.D., Higgins, D.G., & Gibson, T.J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequencing weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* 22, 4673-4680.

Because *Structure with Folding & Design* operates a 'Continuous Publication System' for Research Papers, this paper has been published on the internet before being printed (accessed from http://biomednet.com/cbiology/str). For further information, see the explanation on the contents page.