## STRUCTURE NOTE

# Crystal Structure of the Hypothetical Protein ST2072 From *Sulfolobus tokodaii*

**Yoshikazu Tanaka,[1,2] Kouhei Tsumoto,[1,3]* Eiki Tanabe,[1] Yoshiaki Yasutake,[2] Naoki Sakai,[2] Min Yao,[2,4] Isao Tanaka,[2,4] and Izumi Kumagai[1]**

[1] *Department of Biomolecular Engineering, Graduate School of Engineering, Tohoku University, Sendai, Japan*
[2] *Division of Biological Sciences, Graduate School of Science, Hokkaido University, Sapporo, Japan*
[3] *Department of Medical Genome Sciences, Graduate School of Frontier Sciences, University of Tokyo, Kashiwa, Japan*
[4] *Division of Bio-Crystallography Technology, RIKEN Harima Institute/SPring-8, Hyogo, Japan*

***Introduction.*** ST2072 is a hypothetical function unknown protein from the thermophilic archaeon *Sulfolobus tokodaii* and is categorized into the protein family named "protein of unknown function UPF0047." The presence of UPF0047 among diverse bacteria and archaea (i.e., 15 of 18 archaeal and 44 of 136 bacterial genome sequences evaluated) suggests that this protein plays an essential role in these microorganisms, but no candidate function has been identified thus far. To obtain insights into the functions of UPF0047 from the structural viewpoint, we determined the crystal structure of ST2072 by using the multiple-wavelength anomalous dispersion (MAD) method at 2.0-Å resolution. Structural analysis revealed that three monomers of ST2072 associate into a trimer and that there is a cleft surrounded by highly conserved residues in the intersubunit interface. In this cleft, one sulfate molecule and a zinc ion, which was identified by inductively coupled plasma (ICP) atomic emission spectrometry, were captured. These observations suggest the critical role of this cleft in the function of ST2072.

***Materials and Methods.*** *Construction of expression vectors for ST2072.* The gene encoding ST2072 was amplified by KOD-Plus DNA polymerase by using *S. tokodaii* genomic DNA as a template and the following designed primers. *Nde*I and *Bam*HI sites were incorporated into the ST2072-reverse (5′-NNNNNNNCATATGAAGATAATTTCAAAAGAATTCACTGTAAAAACG-3′) and ST2072-forward (5′-NNNGGATCCTTATTCTCCCATACTTTTCACTAATACTGTCCTAG-3′) sequences, respectively (restriction enzyme sites are underlined). The polymerase chain reaction (PCR) products were inserted into the *Nde*I and *Bam*HI sites of the pET20b vector (Novagen). The correctness of the DNA sequence was confirmed by using an ABI 310 Genetic Analyzer (Applied Biosystems). *Expression and purification of ST2072.* *Escherichia coli* strain BL21 (DE3) transformed with the expression vector encoding the ST2072 gene was grown at 37°C in Luria–Bertani (LB) medium supplemented with 100 μg mL$^{-1}$ ampicillin until early stationary phase. To induce expression of the desired protein, isopropylthio-β-D-galactoside (IPTG) was added to a final concentration of 1 m*M*, and the culture was grown for 3 h at 37°C. The selenomethionine (Se-Met) derivative of ST2072 was expressed in a methionine auxotroph, *E. coli* strain B834(DE3), grown in M9 medium supplied with 1 m*M* of Se-Met. Cells were harvested by centrifugation at 4°C and 5000 × *g* for 10 min, washed with a buffer comprising 50 m*M* Tris-HCl (pH 8.0) and 50 m*M* NaCl, and then resuspended in the same buffer. The suspension was sonicated, followed by centrifugation at 8000 × *g* for 30 min at 4°C. The soluble fraction was incubated at 70°C for 30 min and then centrifuged at 8000 × *g* for 30 min at 20°C. The supernatant was dialyzed against 50 m*M* Tris-HCl (pH 9.0), and then loaded onto a 5-mL HiTrap QXL column (Amersham Bioscience) previously equilibrated with 50 m*M* Tris-HCl (pH 9.0). After the column had been washed with 15 mL of 50 m*M* Tris-HCl (pH 9.0), the adsorbed protein was eluted with 75 mL of a 0–1.0 *M* gradient of NaCl in 50 m*M* Tris-HCl (pH 9.0). Fractions containing ST2072 were further purified on a HiLoad 26/60 Superdex 75-pg column (Amersham Bioscience) equilibrated with 50 m*M* Tris-HCl (pH 8.0) containing 200 m*M* NaCl. Fractions containing the desired protein were dialyzed against 50 m*M* Tris-HCl (pH 9.0) and then loaded onto a ResourceQ column (Amersham Bioscience), previously equilibrated with 50 m*M* Tris-HCl (pH 9.0). Adsorbed protein was eluted with 30 mL of a 0–0.4 M gradient of NaCl in 50 m*M* Tris-HCl (pH 9.0).

Grant sponsor: Ministry of Education, Culture, Sports, Science and Technology of Japan: National Project on Protein Structural and Functional Analyses; Grants-in-Aid for Younger Scientists (to K. Tsumoto); and Grant-in-Aid for the COE project Giant Molecules and Complex Systems, 2002. Grant sponsor: New Energy and Industrial Technology Development Organization (NEDO) of Japan: Industrial Technology Research Grant Program, 2002 (to K. Tsumoto).

*Correspondence to: Kouhei Tsumoto, University of Tokyo, Medical Genome Sciences, 5-1-5, Kashiwanoha, Kashiwa, 277-8562, Japan. E-mail: tsumoto@k.u-tokyo.ac.jp

Received 9 March 2005; Accepted 18 March 2005

Published online 13 October 2005 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20619

**TABLE I. X-ray Data Collection and Refinement Statistics**

| | Peak | Edge | Remote |
|---|---|---|---|
| Data collection | | | |
| Resolution (Å)[a] | 50.0–2.00 (2.07–2.00) | 50.0–2.00 (2.07–2.00) | 50.0–2.00 (2.07–2.00) |
| Wavelength (Å) | 0.9789 | 0.9794 | 0.9000 |
| $R_{sym}$ (%)[a,b] | 0.069 (0.367) | 0.059 (0.370) | 0.065 (0.389) |
| Completeness (%)[a] | 100.0 (100.0) | 100.0 (100.0) | 100.0 (100.0) |
| Unique reflections | 8557 | 8586 | 8590 |
| Averaged $I/\sigma$ ($I$) | 11.0 | 11.3 | 10.6 |
| Average redundancy[a] | 21.4 (21.6) | 21.4 (21.7) | 21.5 (21.8) |
| Refinement and model quality | | | |
| Resolution range (Å) | | | 10–2.00 |
| No. of reflections in working set | | | 7626 |
| No. of reflections in test set | | | 832 |
| $R$-factor[c] | | | 18.1 |
| $R_{free}$-factor[d] | | | 22.2 |
| Total protein atoms | | | 1049 |
| Total ligand atoms | | | 6 |
| Total water atoms | | | 45 |
| Average B-factor (Å$^2$) | | | 26.18 |
| RMSD bond lengths (Å) | | | 0.007 |
| RMSD bond angles (°) | | | 1.42 |
| Ramachandran plot | | | |
| In most favored regions (%) | | | 90.7 |
| In additional allowed regions (%) | | | 9.3 |

[a] The values in parentheses refer to data in the highest resolution shell.
[b] $R_{sym} = \Sigma_h \Sigma_i |I_{h,i} - \langle I_h \rangle / \Sigma_h \Sigma_i |I_{h,i}|$, where $I_h$ is the mean intensity of a set of equivalent reflections.
[c] $R$-factor $= \Sigma|F_{obs} - F_{calc}|/\Sigma\, F_{obs}$, where $F_{obs}$ and $F_{calc}$ are observed and calculated structure factor amplitudes, respectively.
[d] $R_{free}$-factor was calculated for $R$-factor, with a random 10% subset from all reflections.

*Crystallization of ST2072.* ST2072 was dialyzed against 10 m$M$ Tris-HCl (pH 8.0) and concentrated to 20 mg mL$^{-1}$. Initial crystallization conditions were screened by the sparse matrix method at 20°C by using Crystal Screen and Crystal Screen 2 kits (Hampton Research). Crystals of ST2072 most suitable for further analyses could be grown by the hanging-drop vapor diffusion method from 100 m$M$ Tris-HCl (pH 8.5), 22% polyethylene glycol 4000, 0.17 $M$ lithium sulfate, and 15% glycerol.

*X-ray diffractions.* X-ray diffraction of Se-Met–substituted ST2072 was performed on the beamline BL44B2 at SPring-8 (Harima, Japan) under cryogenic conditions (100 K). For MAD phasing, three wavelengths were chosen on the basis of the fluorescence spectrum of the Se $K$ absorption edge, corresponding to the maximum *f″ (peak, 0.9789 Å), minimum f′* (edge, 0.9794 Å), and the reference point (remote, 0.9000 Å). The three-wavelength diffraction data set was collected to a resolution of 2.0 Å with a Quantum210 detector (Area Detector Systems Corporation). The crystal of Se-Met–substituted ST2072 belongs to space group $P2_13$ with unit cell dimensions of $a = b = c =$ 71.72 Å. One molecule of ST2072 was in the asymmetric unit ($V_M$ = 2.06 Å$^3$/Da). The diffraction data were indexed, integrated, scaled, and merged using the HKL2000 program package.[1]

*Structure solution and refinement.* The structure of Se-Met–substituted ST2072 was solved by the MAD method. The program SOLVE/RESOLVE[2–4] was used to calculate the initial phases, to improve the phases, and to perform automatic model building. The program RESOLVE successfully placed 74 out of 134 residues. The complete atomic model with the total 134 residues, including side-chains, was rebuilt by using the program Lafire.[5] Positional and individual $B$ factor refinement was carried out with the program Crystallography & NMR System[6] (CNS), using reflections ranging from 10 Å to 2.0 Å. A random 10% of all observed reflections were set aside for cross-validation analysis and were used to monitor throughout the refinement by calculating the free $R$ value ($R_{\text{free}}$). The final model consisted of 134 residues, one sulfate molecule, one zinc ion, and 45 water molecules with crystallographic $R$ and $R_{\text{free}}$ values of 18.1% and 22.2%, respectively. The stereochemical quality of the final refined model was analyzed with the program PROCHECK.[7] The crystallographic parameters and refinement statistics are summarized in Table I.

*ICP atomic emission spectrometry.* Identification of the metal ion in ST2072 was carried out with a SPS7800 ICP atomic emission spectrometer (SEIKO Instruments Inc.). Purified protein was dialyzed three times against 100 volumes of ultrapure water and then used for the ICP atomic emission measurement. The concentrations of protein were measured using the bicinchoninate method with bovine serum albumin as the standard protein.

***Results and Discussion.*** Crystal structure of ST2072 was determined at a resolution of 2.0 Å by MAD method. The final model contains all 134 residues of ST2072. The monomeric structure of ST2072 is composed of seven β-strands (β1–β7) and four α-helices (α1–α4). The seven β-strands constitute two antiparallel β-sheets; one comprising β1, β7, β3, and β5, and the other, β4, β6, and β2. These two β-sheets are surrounded by four α-helices [Fig. 1(A)]. There is one molecule of ST2072 per asymmetric unit, which formed a crystallographically related three-fold symmetric trimer in crystals. This arrangement is consistent with the results of size exclusion chromatography (data not shown). The trimer of ST2072 assumes a pyrami-

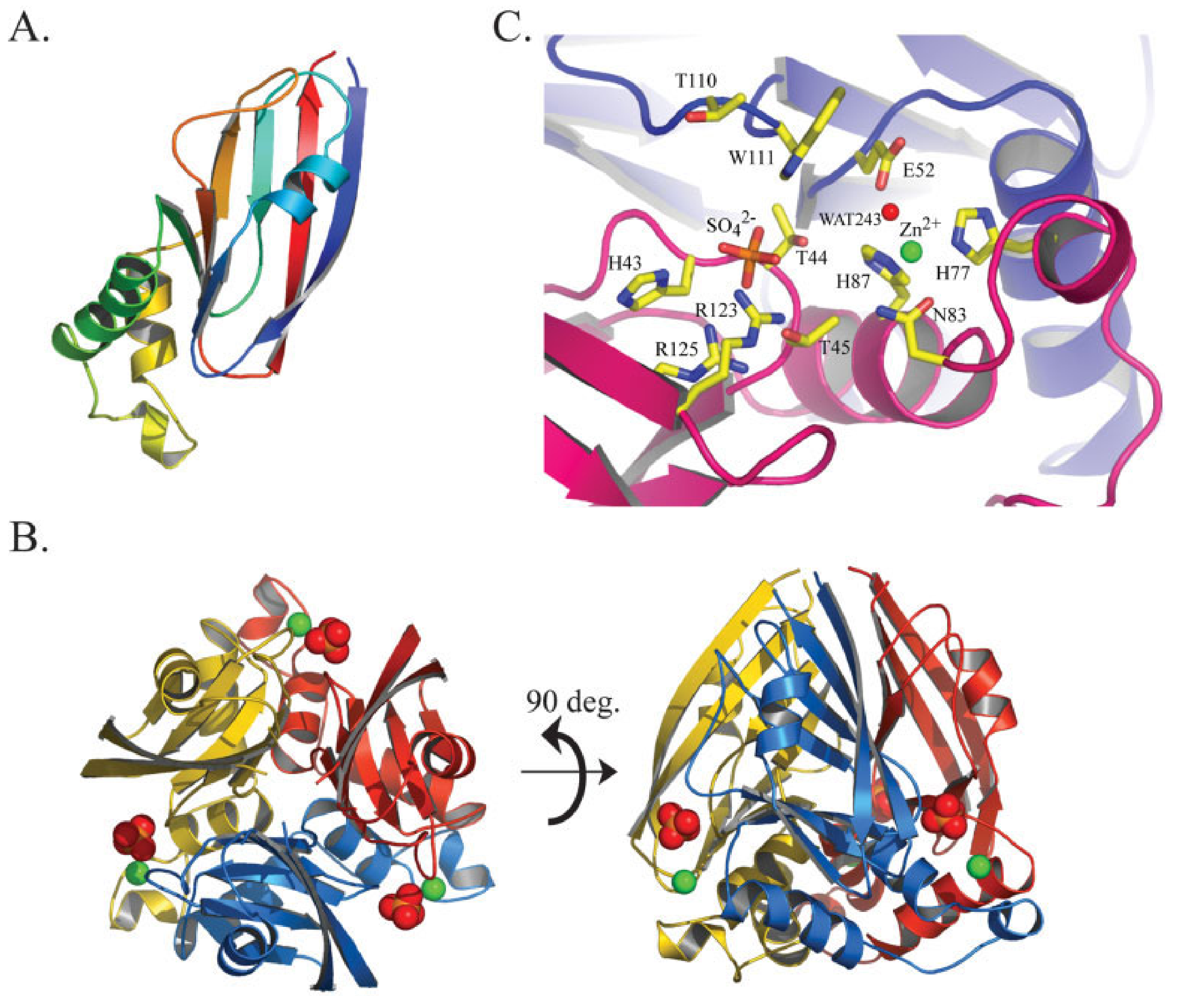

Fig. 1. Crystal structure of ST2072. (**A**) Ribbon diagram of a monomer of ST2072, which is colored according to the sequence by a rainbow color ramp going from blue at the N-terminus to red at the C-terminus. (**B**) Trimeric structure of ST2072. Bound sulfate molecule and zinc ion are also shown as a CPK space-filling model, and the atoms are colored as follows: oxygen (red), phosphate (orange), and zinc (green). (**C**) Close-up view of the cleft where a sulfate molecule and a zinc ion were captured. The sulfate molecule, zinc ion, water molecule, and nearby completely or highly conserved residues are shown. The ribbon diagrams shown are colored according to subunit.

dal shape. All of the β-strands are located in the upper side of the pyramid, which is surrounded by the three α1"s. The bottom of the triangular pyramid consists of only α-helices α2, α3, and α4 [Fig. 1(B)].

The refined structure shows the existence of a cleft in the subunit interface, which is located toward the bottom of one side of the pyramid, between the β-strand-rich and α-helix-rich regions at the bottom. The *Fo-Fc* electron density map indicated that a sulfate molecule, derived from the crystallization buffer, was located at the cleft. The residues participating in binding of the sulfate molecule are His43, Thr44, Thr45, Arg123, Arg125, Gly109, Thr110, and Trp111; the former four residues are included in the same subunit, and the latter three are in the counterpart subunit [Fig. 1(C)]. These residues are highly conserved in homolog proteins (Fig. 2), especially Gly109, Thr110, and Trp111, which are included in a consensus pattern for proteins in this family, Sxx[L, I or V]x[L, I or V]xxGxxxxGTWQx[L, I or V] (corresponding sequence is underlined).

In the cleft that has captured the sulfate molecule, electron density probably derived from some metal ion (on the basis of the coordination of the surrounding residues) was also observed 7.2 Å distant from the sulfur atom in the sulfate molecule. ICP atomic emission spectrometry analysis of purified ST2072 revealed that zinc ion is included in this protein. The molar ratio of zinc ion to protein, as determined by quantitative analysis, is 0.98, which corresponds well to that of zinc ion in crystal structure: three atoms of zinc ion bound to trimeric ST2072. None of the zinc compounds was added in the purification and/or crystallization steps, the zinc ion presumably bound during the culturing of the *E. coli* cells, and ST2072 would have zinc ions throughout all of the experimental proce-

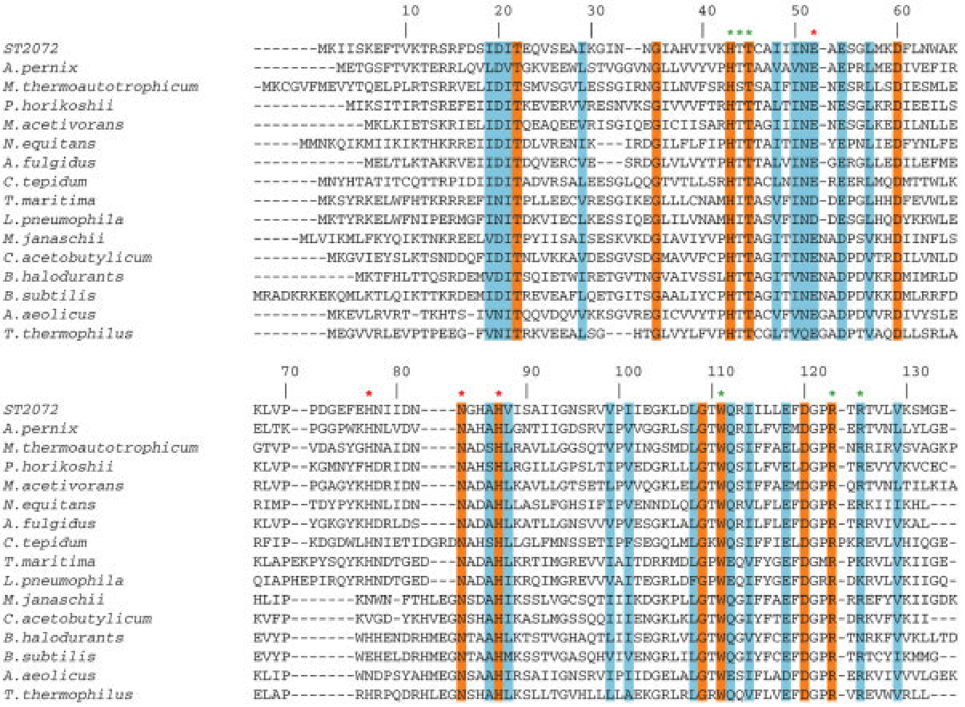

Fig. 2. Multiple sequence alignment of ST2072 and homologous proteins. The residue number refers to ST2072. Completely conserved residues are orange, and highly conserved residues are blue. Residues that participate in binding to the zinc and sulfate ions are indicated by red and green asterisks (*), respectively. *A. pernix*, *Aeropyrum pernix*; *M. thermoautotrophicum*, *Methanobacterium thermoautotrophicum*; *P. horikoshii*, *Pyrococcus horikoshii*; *M. acetivorans*, *Methanosarcina acetivorans*; *N. equitans*, *Nanoarchaeum equitans*; *A. fulgidus*, *Archaeoglobus fulgidus*; *C. tepidum*, *Chlorobium tepidum*; *T. maritima*, *Thermotoga maritima*; *L. pneumophila*, *Legionella pneumophila*; *M. jannaschii*, *Methanococcus jannaschii*; *C. acetobutylicum*, *Clostridium acetobutylicum*; *B. halodurans*, *Bacillus halodurans*; *B. subtilis*, *Bacillus subtilis*; *A. aeolicus*, *Aquifex aeolicus*; *T. thermophilus*, *Thermus thermophilus*.

dures. The zinc ion is bound to His77, Asn83, His87, and one water molecule, resulting in a tetrahedral coordination—a geometry that is observed frequently with zinc ions.[8] In addition, His77, Asn83, and His87 are highly conserved, indicating that the bound zinc ion may play a critical role in the function of this protein.

A large proportion of conserved residues are arranged around the cleft that binds a sulfate molecule and a zinc ion, suggesting a critical functional role for this site (e.g., a catalytic site). Recently, crystal structures of homologous proteins from *Clostridium acetobutylicum*, *Bacillus halodurans*, and *Thermotoga maritima* were determined, and the atomic coordinates were published (1XBF, 1VMF, and 1VMJ, respectively). All of these proteins also contained negatively charged compounds at the site we focus on here. The cleft is rather broad, with a width of approximately 18 Å; thus, the substrate may be a compound that has relatively long scaffolds, or perhaps multiple compounds may bind at this site as substrates. Indeed, the crystal structure of the homologous protein from *B. halodurans* binds 4-(2-hydroxyethyl)-1-piperazine ethanesulfonic acid (HEPES) and 1,2-ethandiol at this site. The electrostatic surface map of ST2072 shows that the right side of the cleft [Fig. 1(C)], where zinc ion bound is negatively charged, and the left side, where sulfate molecule bound is positively charged, and the completely conserved His43 and Arg125 are in the positively charged area. For a large number of enzymes, histidine plays an essential role in catalysis as a nucleophilic or dehydration residue, usually where N$\epsilon$ is important.[9,10] However, N$\epsilon$ of His43 in ST2072 or its homologs does not turn to the inside of the cleft, suggesting that this residue might not participate directly in the function of the protein (e.g., in catalytic activity). Instead, the capture of a sulfate molecule in the crystal structure suggests that perhaps His43 and Arg125 are used for identifying a negatively charged functional group of certain ligands (i.e., a substrate).

In several enzymes, zinc ion bound at the active site participates directly in chemical catalysis, with the zinc coordination involving at least one water molecule.[8,11,12] In the crystal structure of ST2072, one water molecule is bound to the zinc ion, suggesting that it is catalytically active. As with other enzymes in which the zinc ion directly contributes to a chemical reaction, the zinc ion in ST2072 likely activates the bound water molecule for nucleophilic attack, polarization of a scissile bond, or stabilization of the transition state.[8,12,13] It should be noted that there is a completely conserved Trp111 opposite the zinc ion binding side of the negatively charged cleft, suggesting that Trp111 may cooperate in the reaction of the water molecule by cation–π interaction with the positively charged part in the substrate.

We now are trying to identify specific ligands for ST2072 by using a chemical compound library from combinatorial approaches, and docking analyses.

***Data Deposition.*** The atomic coordinates for ST2072 are available in the Protein Data Bank (PDB code: 1ve0).

***Acknowledgments.*** Our thanks to Drs. H. Naitow and T. Matsu for their kind help with the data collection on beamline BL44B2 at SPring-8.

## REFERENCES

1. Otwinowski Z, Minor W. Processing of X-ray diffraction data collected in oscillation mode. Methods Enzymol 1997;276:307–326.
2. Terwilliger TC, Berendzen J. Automated MAD and MIR structure solution. Acta Crystallogr D Biol Crystallogr 1999;55:849–861.
3. Terwilliger TC, Berendzen J. Evaluation of macromolecular electron-density map quality using the correlation of local r.m.s. density. Acta Crystallogr D Biol Crystallogr 1999;55:1872–1877.
4. Terwilliger TC. Maximum-likelihood density modification. Acta Crystallogr D Biol Crystallogr 2000;56:965–972.
5. Yao M, Zhou Y, Tanaka I. Lafire: an automatic refinement program for protein crystallography. To be published.
6. Brunger AT, Adams PD, Clore GM, DeLano WL, Gros P, Grosse-Kunstleve RW, Jiang JS, Kuszewski J, Nilges M, Pannu NS, Read RJ, Rice LM, Simonson T, Warren GL. Crystallography & NMR system: a new software suite for macromolecular structure determination. Acta Crystallogr D Biol Crystallogr 1998;54:905–921.
7. Laskowski RA, Macarthur MW, Moss DS, Thornton JM. PROCHECK—a program to check the stereochemical quality of protein structures. J Appl Crystallogr 1993;26:283–291.
8. McCall KA, Huang C, Fierke CA. Function and mechanism of zinc metalloenzymes. J Nutr 2000;130:1437S–1446S.
9. Chapman E, Best MD, Hanson SR, Wong CH. Sulfotransferases: structure, mechanism, biological activity, inhibition, and synthetic utility. Angew Chem Int Ed Engl 2004;43:3526–3548.
10. Kai Y, Matsumura H, Izui K. Phosphoenolpyruvate carboxylase: three-dimensional structure and molecular mechanisms. Arch Biochem Biophys 2003;414:170–179.
11. Vallee BL, Auld DS. Active-site zinc ligands and activated H2O of zinc enzymes. Proc Natl Acad Sci USA 1990;87:220–224.
12. Hightower KE, Fierke CA. Zinc-catalyzed sulfur alkyation: insights from protein farnesyltransferase. Curr Opin Chem Biol 1999;3:176–181.
13. Christianson DW, Cox JD. Catalysis by metal-activated hydroxide in zinc and manganese metalloenzymes. Annu Rev Biochem 1999;68:33–57.