Structures of a histidine triad family protein from *Entamoeba histolytica* bound to sulfate, AMP and GMP

Donald D. Lorimer,^{a,b,*} Ryan Choi,^{a,c} Ariel Abramov,^{a,d} Stephen Nakazawa Hewitt,^{a,c} Anna S. Gardberg,^{a,b,‡} Wesley C. Van Voorhis,^{a,e} Bart L. Staker,^{a,d} Peter J. Myler^{a,d,f} and Thomas E. Edwards^{a,b}

Received 9 December 2014

Accepted 4 February 2015

Edited by W. N. Hunter, University of Dundee, Scotland

‡ Present address: EMD Serono Research & Development Institute Inc., Billerica, MA 01821, USA

Keywords: hydrolase; structural genomics; Seattle Structural Genomics Center for Infectious Disease; SSGCID; metal-binding protein.

PDB references: histidine triad family protein, complex with sulfate, 3oj7; complex with GMP, 3oxk; complex with AMP, 3omf

Supporting information: this article has supporting information at journals.iucr.org/f

^aSeattle Structural Genomics Center for Infectious Disease, USA, ^bBeryllium, Bainbridge Island, WA 98110, USA, ^cCERID, University of Washington, Seattle, WA 98109, USA, ^dSeattle Biomedical Research Institute, Seattle, WA 98109, USA, ^eDepartment of Medicine, Division of Allergy and Infectious Diseases, University of Washington, Seattle, WA 98195, USA, and ^fDepartment of Global Health and Medical Education and Biomedical Bioinformatics, University of Washington, Seattle, WA 98109, USA. *Correspondence e-mail: dlorimer@be4.com

Three structures of the histidine triad family protein from *Entamoeba histolytica*, the causative agent of amoebic dysentery, were solved at high resolution within the Seattle Structural Genomics Center for Infectious Disease (SSGCID). The structures have sulfate (PDB entry 3oj7), AMP (PDB entry 3omf) or GMP (PDB entry 3oxk) bound in the active site, with sulfate occupying the same space as the α -phosphate of the two nucleotides. The C α backbones of the three structures are nearly superimposable, with pairwise r.m.s.d.s ranging from 0.06 to 0.13 Å.

1. Introduction

The histidine triad proteins are a large and highly conserved family of proteins characterized as having the conserved sequence H φ H φ H/Q φ (where φ is any hydrophobic amino acid). This family of proteins can be classified into three groups based upon sequence, substrate specificity, structure and mechanism: (i) the HINT branch, which consists of adenosine 5'-monophosphoramidate hydrolases, (ii) the Fhit branch, which consists of diadenosine polyphosphate hydrolases, and (iii) the GalT branch, which consists of specific nucleoside monophosphate transferases (Brenner, 2002). At least one human representative is lost in some diseases. For example, apraxin, a HINT family member, is lost in ataxia-oculomotor apraxia syndrome, Fhit is lost in the development of epithelially derived tumors and GalT is lost in galactosemia (Brenner *et al.*, 1997; Brenner, 2002).

The Seattle Structural Genomics Center for Infectious Disease (SSGCID) protein EnhA.01296.a (*EhHIT*) was originally identified in the *Entamoeba histolytica* genome in 2005 (Loftus *et al.*, 2005) as a hypothetical protein called EHI_093910 and given GenBank accession XP_655618. The target was one of 1477 eukaryotic pathogens selected by SSGCID based on sequence similarity to protein targets in the DrugBank database (Wishart *et al.*, 2008; Myler *et al.*, 2009; Stacy *et al.*, 2011; Phan *et al.*, 2014). Genes from several eukaryotic genomes, including *E. histolytica*, were used in a BLAST search against DrugBank and those that showed >50% similarity over >75% of their length were passed to the next filter. Additional filters included a rigorous screen of physical properties to filter out those with unreasonable sizes or amino-acid compositions or with the presence of trans-

Table 1
Macromolecule-production information.

Source organism	<i>E. histolytica</i>
DNA source	<i>E. histolytica</i> strain HM-1:IMSS, genomic DNA
Forward primer	GGGTCCTGGTTCGATGGCTGATTCATGCATTTCTGTA
Reverse primer	CTTGTTTCGTGCTGTTTATTATAATTTATCCCAAG-CAAGTTTTTTCC
Cloning vector	AVA0421
Expression vector	AVA0421, IPTG or lactose inducible
Expression host	<i>E. coli</i> Rosetta Oxford BL21(DE3)R3
Complete amino-acid sequence of the construct produced	GP GSMADSCIFCKIAQKQIPSTIVYEDDEIFAFK-DINPIAPIHILVIPKQHIASLNEITEENEAFI-GKVLKYVSLIGKKECEGYRVVNNIGEDAGQT-VKHIHFHILGGKKLAWDKL

membrane domains or low-complexity sequences, which are associated with difficulty in soluble expression and crystallization. Additionally, targets with close similarity to sequences already in the PDB or already targeted by other structural genomics groups were removed (Myler *et al.*, 2009). For *EhHIT*, *XtalPred-EP* gives a score of three out of five for likelihood of crystallizability, which is average for a protein of this size (Slabinski *et al.*, 2007). Similarly, *XtalPred-RF* gives a score of four out of 11 or slightly better than average crystallizability. No transmembrane domains were predicted by *TMpred* (Hofmann & Stoffel, 1993). Based upon sequence similarity only, this target appears to be in the HINT branch of the HITs. We initially determined the structure of the apo protein, which had a molecule of sulfate in its active site, and we then co-crystallized this target with AMP and GMP. Here, we present the sulfate-bound structure and the two nucleotide-bound structures.

2. Materials and methods

2.1. Macromolecule production

The target gene was PCR-amplified using Phusion DNA polymerase (New England Biolabs) and suitable primers (GGGTCCTGGTTCGATGGCTGATTCATGCATTTCTGTA and CTTGTTTCGTGCTGTTTATTATAATTTATCCCAAGCAAGTTTTTTCC), with *E. histolytica* HM-1:IMSS cDNA as the template (Table 1). The PCR product was gel-purified, treated with T4 DNA polymerase and annealed *via* ligation-independent cloning (LIC) to the AVA0421 vector, a T7 expression vector containing a cleavable six-histidine tag at the N-terminus (Choi *et al.*, 2011). The LIC-annealed construct was propagated in *Escherichia coli* NovaBlue competent cells (EMD Biosciences) and subsequently transformed into the *E. coli* Rosetta Oxford BL21(DE3)R3 expression host strain for expression testing. After transformation into the expression host, a starter culture from a single colony in LB broth was grown for ~18 h at 310 K. Protein was expressed in a LEX bioreactor in the presence of 50 µg ml⁻¹ ampicillin, 50 µg ml⁻¹ carbenicillin and 34 µg ml⁻¹ chloramphenicol in 2 l ZYP-5052 auto-induction medium (Studier, 2005). The culture was grown for 24 h at 298 K and was then incubated for a further 60 h at 288 K. The cells were collected

by centrifugation at 4000g for 20 min at 277 K. The cell paste was stored frozen at 193 K. The frozen *E. coli* cell-paste pellet was lysed in 200 ml lysis buffer (0.3 M NaCl, 20 mM HEPES pH 7.4, 5% glycerol, 30 mM imidazole, 0.5% CHAPS, 2 mM MgCl₂) with lysozyme, five tablets of Roche protease-inhibitor cocktail and 250 µl 14.3 M β-mercaptoethanol. The suspension was sonicated for 45 min with a pulse frequency of 5 s on, 10 s off. The lysate was incubated with 20 µl Benzonase for 40 min at room temperature and clarified by centrifugation for 60 min at 15 200g in a Sorvall SLA-1500 rotor. The lysate was filtered using a 0.45 µm low-protein-binding filter and loaded onto a 5 ml HisTrap FF column (GE Healthcare, Piscataway, New Jersey, USA). The protein was eluted with an 80 ml gradient of 0–100% buffer B (30–500 mM imidazole). Fractions containing the target protein were pooled and concentrated for size-exclusion chromatography (SEC). The protein was loaded onto a HiLoad 26/60 Superdex preparative-grade SEC column (GE Healthcare, Piscataway, New Jersey, USA) and run in a buffer consisting of 0.3 M NaCl, 20 mM HEPES pH 7.0, 5% glycerol, 1 mM TCEP at 1.5 ml min⁻¹; 2.5 ml fractions were collected. The protein eluted at a volume consistent with a monomer. The fractions containing the target protein were pooled, concentrated to 83.24 mg ml⁻¹, flash-frozen in 200 µl aliquots in liquid nitrogen and stored at 193 K. Protein concentrations were determined using a Nanodrop ND-1000 (Thermo Fisher Scientific, Waltham, Massachusetts, USA).

2.2. Crystallization

All crystallization experiments were set up as sitting-drop vapour-diffusion trials using XJR Junior crystallization trays (Rigaku Reagents). The drops were formed by mixing 0.4 µl protein solution with 0.4 µl precipitant solution and were incubated against 80 µl reservoir solution at 289 K. Crystallization conditions were searched for using the commercial screens JCSG+ (Rigaku Reagents) and PACT (Molecular Dimensions). Crystallization conditions for *EhHIT* at 83.2 mg ml⁻¹ were as follows: *EhHIT* + sulfate, JCSG+ condition A1 (50% PEG 400, 0.1 M acetate buffer pH 4.5, 200 mM Li₂SO₄), direct cryocooling; *EhHIT* + AMP, JCSG+ condition B6 [40%(v/v) ethanol, 5%(w/v) PEG 1000, phosphate-citrate buffer pH 4.2, 10 mM AMP], cryopreserved with 20% ethylene glycol; *EhHIT* + GMP, JCSG+ condition B6 [40%(v/v) ethanol, 5%(w/v) PEG 1000, phosphate-citrate buffer pH 4.2, 10 mM AMP], cryopreserved with 20% ethylene glycol. All crystals were incubated for several seconds in the cryosolution and vitrified by plunging them into liquid nitrogen.

2.3. Data collection and processing

The diffraction data set for the *EhHIT* + sulfate crystal was collected on ALS beamline 5.0.3 with an ADSC Quantum 315 CCD detector. The diffraction data sets for the *EhHIT* + AMP and *EhHIT* + GMP crystals were collected in-house with a Rigaku FR-E+ SuperBright generator using a Rigaku 944+ CCD detector. All data sets were collected at 100 K. The

Table 2
Data collection and processing.

Values in parentheses are for the outer shell.

	<i>Eh</i> HIT + sulfate (PDB entry 3oj7)	<i>Eh</i> HIT + AMP (PDB entry 3omf)	<i>Eh</i> HIT + GMP (PDB entry 3oxk)
Wavelength (Å)	0.9765	1.5418	1.5418
Space group	C222 ₁	C222 ₁	C222 ₁
<i>a</i> , <i>b</i> , <i>c</i> (Å)	52.89, 61.03, 67.31	53.58, 60.63, 67.88	53.66, 60.86, 67.93
α , β , γ (°)	90, 90, 90	90, 90, 90	90, 90, 90
Resolution range (Å)	34.37–1.40 (1.44–1.40)	34.56–1.80 (1.85–1.80)†	27.77–1.55 (1.59–1.55)
No. of unique reflections	21814 (1582)	10570 (596)	16495 (1185)
Completeness (%)	99.8 (100.0)	91.6 (80.3)	90.3 (82.4)
Multiplicity	7.03 (7.12)	5.8 (3.7)	6.6 (5.2)
$\langle I/\sigma(I) \rangle$	25.27 (3.80)	52.96 (21.40)	25.18 (3.80)
$R_{\text{r.i.m.}}$ ‡	0.039 (0.534)	0.025 (0.053)	0.054 (0.434)
Overall <i>B</i> factor from Wilson plot (Å ²)	24.26	16.26	18.01

† The resolution cutoff was restricted by the detector size as well as the completeness in the outer shell. ‡ Estimated $R_{\text{r.i.m.}} = R_{\text{merge}}[N/(N - 1)]^{1/2}$, where *N* is the data multiplicity.

Table 3
Structure solution and refinement.

Values in parentheses are for the outer shell.

	<i>Eh</i> HIT + sulfate (PDB entry 3oj7)	<i>Eh</i> HIT + AMP (PDB entry 3omf)	<i>Eh</i> HIT + GMP (PDB entry 3oxk)
Resolution range (Å)	34.37–1.40 (1.44–1.40)	34.56–1.80 (1.847–1.800)	27.77–1.55 (1.59–1.55)
Completeness (%)	99.8	91.4	90.0
σ Cutoff	$F > 0.000\sigma(F)$	$F > 0.00\sigma(F)$	$F > 0.00\sigma(F)$
No. of reflections			
Working set	21706 (1479)	9660 (596)	14847 (932)
Test set	1111 (95)	467 (27)	712 (40)
Final R_{cryst}	0.162 (0.247)	0.146 (0.146)	0.157 (0.187)
Final R_{free}	0.184 (0.251)	0.191 (0.206)	0.208 (0.210)
No. of non-H atoms			
Protein	856	879	897
Ion	1 Zn ²⁺ , 5 SO ₄ ^{2−}	1 Zn ²⁺	1 Zn ²⁺
Ligand	0	1 AMP	1 GMP
Water	98	165	188
Total	960	1068	1113
R.m.s. deviations			
Bonds (Å)	0.015	0.012	0.013
Angles (°)	1.697	1.516	1.628
Average <i>B</i> factors (Å ²)			
Protein	19.9	7.9	10.2
Ion	32.4	9.3	9.1
Ligand	0.0	18.6	14.9
Water	26.4	19.8	19.8
Ramachandran plot (%)			
Most favored	99.1	100	99.1
Outliers	0	0	0

X-ray data were reduced with *XDS* and *XSCALE*. Details of the data collections are summarized in Table 2.

2.4. Structure solution and refinement

The structures were solved by molecular replacement using *Phaser* (McCoy *et al.*, 2007). PDB entry 1xqu (Southeast Collaboratory for Structural Genomics, unpublished work) was used as the search model for the *Eh*HIT + sulfate structure;

the *Eh*HIT + sulfate structure (PDB entry 3oj7) was used as the model for *Eh*HIT + AMP, and the *Eh*HIT + AMP structure (PDB entry 3omf) was used as the model for *Eh*HIT + GMP. Reciprocal-space refinement was performed with *REFMAC* using TLS (Murshudov *et al.*, 2011) in iterative cycles of model building using *Coot* (Emsley & Cowtan, 2004). The models contained almost the entire length of the protein: residues 4–113 for *Eh*HIT + sulfate, residues 3–113 for *Eh*HIT + AMP and residues −2 to 113 for *Eh*HIT + GMP. PDB codes for the three structures are given in Table 3.

3. Results and discussion

The *E. histolytica* protein EnhA.01296.a (*Eh*HIT) is a classic example of a HIT protein with the signature motif H ϕ H ϕ H ϕ ϕ , in this case His-Ile-His-Phe-His-Ile-Leu, with the first histidine located 17 amino acids from the C-terminus at residue 97. By sequence analysis this protein appears to belong to the HINT branch of the HIT family. *Eh*HIT is similar in amino-acid sequence to proteins that have been given the designation protein kinase C inhibitor-1 (PKCI-1). This HINT branch has also been identified to contain purine nucleoside- and nucleotide-binding proteins (Brenner *et al.*, 1997; Brenner, 2002). This target was predicted by *XtalPred* (Slabinski *et al.*, 2007) to have average or slightly better than average crystallizability, with no other physical or sequence elements thought to inhibit crystallization. With this in mind, we set up initial

Figure 1
Overlay of three *Eh*HIT structures. The models are presented as cartoon renderings with the sulfate-bound structure (PDB entry 3oj7) in green, the AMP-bound structure (PDB entry 3omf) in cyan and the GMP-bound structure (PDB entry 3oxk) in magenta. The AMP in PDB entry 3omf is shown as sticks and the Zn²⁺ ion is shown as a grey sphere. The pairwise r.m.s.d.s are 0.111 Å for 3oj7 *versus* 3omf, 0.133 for 3oj7 *versus* 3oxk and 0.064 Å for 3omf *versus* 3oxk. The overlay and pairwise scores were generated in *PyMOL*.

Figure 2

Key interactions of molecules bound to *EhHIT*. (a) Coordination of Zn^{2+} in all three structures, (b) coordination of sulfate in PDB entry 3oj7, (c) coordination of AMP in PDB entry 3omf, (d) coordination of GMP in PDB entry 3oxk. The protein coloration is the same as in Fig. 1. This figure was generated in *PyMOL*.

apo crystallization trials using only two commercial crystallization screens: JCSG+ and PACT. Based upon the proposed sequence similarity to the HINT branch, the initial apo trials were followed up with co-crystallizations with the purine monophosphate nucleotides AMP and GMP. The first crystal structure of *EhHIT* to be solved contained a bound sulfate that came from the crystallization condition, JCSG+ condition A1, which contained 200 mM Li_2SO_4 . The model was determined by molecular replacement using the *Clostridium thermocellum* HIT structure (PDB entry 1xqu) as a search model. The nucleotide-bound structures, *EhHIT* + AMP and *EhHIT* + GMP, were subsequently determined by molecular replacement using PDB entries 3oj7 and 3omf as search models, respectively. The structures of all three were determined at an atomic resolution of between 1.40 and 1.80 Å. The pairwise C^α r.m.s.d.s are 0.064 Å for *EhHIT* + AMP versus *EhHIT* + GMP, 0.111 Å for *EhHIT* + sulfate versus *EhHIT* + AMP and 0.133 Å for *EhHIT* + sulfate versus *EhHIT* + GMP.

The three models overlay with very small r.m.s.d.s (Fig. 1). In all three structures the Zn^{2+} ion is coordinated by Cys5, Cys8, His47 and His97 (Fig. 2a). We identified the metal ion as

Zn^{2+} based on other HINT structures such as PDB entry 1kpa (Lima *et al.*, 1996). Additionally, we used *CheckMyMetal* (http://csgid.org/csgid/metal_sites/) to confirm the assignment of Zn^{2+} . In the sulfate-bound structure the sulfate is coordinated by Asn86, Thr94, His99 and His101 (Fig. 2b). In the AMP-bound structure the phosphate is coordinated by the same residues as the sulfate, Asn86, Thr94, His99 and His101, while the ribose and adenine are further coordinated by Asp31 and Asn32 (Fig. 2c). Similarly, for the GMP-bound structure the phosphate, ribose and guanosine of GMP make the same contacts as AMP, with one additional contact between the backbone carbonyl of Lys30 and the 2-amino group of guanosine (Fig. 2d). In both the AMP-bound and GMP-bound structures the Watson–Crick face of the purine makes water-mediated contacts with the protein, the opposite face is solvent-exposed and the sugar edge makes direct contacts with the protein.

Acknowledgements

The authors wish to thank the entire SSGCID team. This research was funded by the National Institute of Allergy and

Infectious Diseases, National Institute of Health, Department of Health and Human Services under Federal Contract Nos. HHSN272201200025C and HHSN272200700057C. The expression plasmids and surplus protein samples can be obtained from either the contact authors or through <http://www.ssgcid.org> and the raw X-ray diffraction data can be obtained through <http://www.csgid.org>.

References

- Brenner, C. (2002). *Biochemistry*, **41**, 9003–9014.
- Brenner, C., Garrison, P., Gilmour, J., Peisach, D., Ringe, D., Petsko, G. A. & Lowenstein, J. M. (1997). *Nature Struct. Mol. Biol.* **4**, 231–238.
- Choi, R., Kelley, A., Leibly, D., Nakazawa Hewitt, S., Napuli, A. & Van Voorhis, W. (2011). *Acta Cryst. F* **67**, 998–1005.
- Emsley, P. & Cowtan, K. (2004). *Acta Cryst. D* **60**, 2126–2132.
- Hofmann, K. & Stoffel, W. (1993). *Biol. Chem. Hoppe Seyler*, **374**, 166.
- Lima, C. D., Klein, M. G., Weinstein, I. B. & Hendrickson, W. A. (1996). *Proc. Natl Acad. Sci. USA*, **93**, 5357–5362.
- Loftus, B. *et al.* (2005). *Nature (London)*, **433**, 865–868.
- McCoy, A. J., Grosse-Kunstleve, R. W., Adams, P. D., Winn, M. D., Storoni, L. C. & Read, R. J. (2007). *J. Appl. Cryst.* **40**, 658–674.
- Murshudov, G. N., Skubák, P., Lebedev, A. A., Pannu, N. S., Steiner, R. A., Nicholls, R. A., Winn, M. D., Long, F. & Vagin, A. A. (2011). *Acta Cryst. D* **67**, 355–367.
- Myler, P. J., Stacy, R., Stewart, L., Staker, B. L., Van Voorhis, W. C., Varani, G. & Buchko, G. W. (2009). *Infect. Disord. Drug Targets*, **9**, 493–506.
- Phan, I. Q. H., Stacy, R. & Myler, P. J. (2014). *Methods Mol. Biol.* **1140**, 53–59.
- Slabinski, L., Jaroszewski, L., Rychlewski, L., Wilson, I. A., Lesley, S. A. & Godzik, A. (2007). *Bioinformatics*, **23**, 3403–3405.
- Stacy, R., Begley, D. W., Phan, I., Staker, B. L., Van Voorhis, W. C., Varani, G., Buchko, G. W., Stewart, L. J. & Myler, P. J. (2011). *Acta Cryst. F* **67**, 979–984.
- Studier, F. W. (2005). *Protein Expr. Purif.* **41**, 207–234.
- Wishart, D. S., Knox, C., Guo, A. C., Cheng, D., Shrivastava, S., Tzur, D., Gautam, B. & Hassanali, M. (2008). *Nucleic Acids Res.* **36**, D901–D906.