Title

Analysis of nucleoside-binding proteins by ligand-specific elution from dye resin: application to Mycobacterium tuberculosis aldehyde dehydrogenases

Permalink

https://escholarship.org/uc/item/9w05r8qr

Journal

Journal of Structural and Functional Genomics, 10(4)

ISSN

1570-0267

Authors

Kim, Chang-Yub
Webster, Cecelia
Roberts, Justin K.
et al.

Publication Date

2009-12-01

DOI

10.1007/s10969-009-9073-z

Peer reviewed

Analysis of nucleoside-binding proteins by ligand-specific elution from dye resin: application to *Mycobacterium tuberculosis* aldehyde dehydrogenases

Chang-Yub Kim · Cecelia Webster · Justin K. M. Roberts · Jin Ho Moon · Emily Z. Alipio Lyon · Heungbok Kim · Minmin Yu · Li-Wei Hung · Thomas C. Terwilliger

Abstract We show that Cibacron Blue F3GA dye resin chromatography can be used to identify ligands that specifically interact with proteins from *Mycobacterium tuberculosis*, and that the identification of these ligands can facilitate structure determination by enhancing the quality of crystals. Four native *Mtb* proteins of the aldehyde dehydrogenase (ALDH) family were previously shown to be specifically eluted from a Cibacron Blue F3GA dye resin with nucleosides. In this study we characterized the nucleoside-binding specificity of one of these ALDH isozymes (recombinant *Mtb* Rv0223c) and compared these biochemical results with co-crystallization experiments with different Rv0223c-nucleoside pairings. We found that the strongly interacting ligands (NAD and NADH) aided formation of high-quality crystals, permitting solution of the first *Mtb* ALDH (Rv0223c) structure. Other nucleoside ligands (AMP, FAD, adenosine, GTP and NADP) exhibited weaker binding to Rv0223c, and produced co-crystals diffracting to lower resolution. Difference electron density maps based on crystals of Rv0223c with various nucleoside ligands show most share the binding site where the natural ligand NAD binds. From the high degree of similarity of sequence and structure compared to human mitochondrial ALDH-2 (BLAST Z-score = 53.5 and RMSD = 1.5 Å), Rv0223c appears to belong to the ALDH-2 class. An altered oligomerization domain in the Rv0223c structure seems to keep this protein as monomer whereas native human ALDH-2 is a multimer.

C.-Y. Kim (✉) · E. Z. Alipio Lyon · H. Kim · T. C. Terwilliger
Advanced Measurement Science Group (B-9), Bioscience Division, MS M888, Los Alamos National Laboratory, Los Alamos, NM 87545, USA
e-mail: cykim@lanl.gov

C. Webster · J. K. M. Roberts
Department of Biochemistry, University of California, Riverside, CA 92521, USA

J. H. Moon
Institute of Life Sciences and Natural Resources, College of Life Sciences & Biotechnology, Korea University, Seoul, Korea

M. Yu
Physical Biosciences Division, Lawrence Berkeley National Laboratory, MS4R0230, Berkeley, CA 94720, USA

L.-W. Hung
Physics Division, MS D454, Los Alamos National Laboratory, Los Alamos, NM 87545, USA

Abbreviations

DNA	Deoxyribonucleic acid
RNA	Ribonucleic acid
AMP	Adenosine-5′-monophosphate
ADP	Adenosine-5′-diphosphate
ATP	Adenosine-5′-triphosphate
FAD	Flavin-adenine dinucleotide
FMN	Flavin mononucleotide
NMN	Nicotinamide mononucleotide
GTP	Guanosine-5′-triphosphate
IPTG	Isopropyl-β-D-thiogalactoside
NAD	Nicotinamide adenine dinucleotide

NADP   Nicotinamide adenine dinucleotide phosphate
LC   Liquid chromatography
MS   Mass spectrometry

Introduction

Nucleosides and their derivatives in cellular metabolism are well recognized as energy carriers in metabolic transactions, essential chemical links for intracellular signals, and constituents of nucleic acids (DNA and RNA) [1, 2]. The metabolism of nucleosides is vital to a cell's survival, and about half of all enzymes are nucleoside-dependent, representing one of the largest and most important classes of cellular proteins [3]. Determining structure–function relationships of nucleoside-binding proteins is therefore a significant component of the structural genomics of proteins that are crucial to cell function. Two major challenges in such studies are the significant number of target proteins, and the problem of sorting proteins according to their nucleoside specificity.

Nucleoside ligand-binding to individual proteins typically occurs with high specificity, and is often sensitive to slight changes at the protein interaction site [4–6]. Elucidating specific interactions between ligands and proteins helps to derive functional insights for many proteins [7, 8], complementing bioinformatics approaches, and may provide the sole source of functional information for hypothetical proteins.

The aldehyde dehydrogenases (ALDH) comprise a large family of proteins which metabolize various endogenous and exogenous substrates [9, 10]. The human genome contains 19 putative ALDH genes and three pseudogenes. Many of them are regulated in response to oxidative stress and over expressed in various tumors [11, 12]. ALDH enzymes have multiple catalytic and non-catalytic functions in ester hydrolysis, antioxidant properties, xenobiotic bioactivation and UV light absorption, and also play important roles in embryogenesis, development and neurotransmission [13, 14]. Mutations in these genes cause subsequent inborn errors in aldehyde metabolism, such as Sjögren-Larsson syndrome [15], Type II hyperprolinaemia and gamma-hydroxybutyric aciduria [16], and pyridoxine-dependent seizures [17]. Human ALDH-2 is important as a nitroglycerin reductase [18] and an activator of NADPH oxidases [9], and for the major function of elimination of toxic aldehydes which lead to lipid peroxidation, protein/enzyme dysfunction, structural damage and apoptosis in alcohol related disorders [19], such as alcohol liver disease [20], heart disease [21] and gastrointestinal cancer [22].

The *Mycobacterium tuberculosis* (*Mtb*) genome encodes ten putative ALDH proteins potentially associated with seven different ALDH classes (Table 1), suggesting that they have diversity comparable to human ALDHs [13, 14]. The presence of a Rossman consensus sequence [23] in four of ten predicted *Mtb* ALDHs (Table 1) further suggests diversity at the level of nucleotide binding at the cofactor binding site.

In a previous report, we used dye-ligand elution chromatography to screen for nucleoside-binding proteins in *Mtb* cell extracts and to analyze the specificity of nucleoside-protein interactions [24]. That study identified 26 native *Mtb* proteins binding to Cibacron Blue resin that were specifically eluted with nucleosides. Four of these 26 proteins were members of the ALDH family, as shown in Table 1. The large number of ALDH proteins in the relatively small *Mtb* genome, and the many essential functions of ALDH proteins in human cells, suggest potential critical roles of *Mtb* ALDHs for survival in its human host environment.

In this report, we purify one of these *Mtb* ALDHs (Rv0223c), characterize its nucleoside specificity, use ligands that interact with Rv0223c to improve its crystallization, present the first structure of an *Mtb* ALDH, and show its close structural similarity to human mitochondrial ALDH-2.

Materials and methods

Cloning and expression of putative *Mtb* ALDHs

Four *Mtb* ALDH genes (Rv0223c, Rv0458, Rv2858c, and Rv3293) were previously identified as interacting with nucleosides using dye-resin chromatography and ligand-specific elution [24]. Each targeted ALDH gene was amplified by PCR from a *M. tuberculosis* H37Rv COSMID library as the template with Pfu proof-reading DNA polymerase (Stratagene), using the 5′ NdeI primer, 5′-AGA-TATACATATG + (N-terminal 21 bases of target sequence)-3′, and the 3′ BamHI primer, 5′-AATTCG-GATCC + (C-terminal 23 bases of target sequence)-3′. The underlined bases represent the NdeI and BamHI sites, respectively. The PCR amplicon was digested with NdeI and BamHI restriction endonucleases (NEB), and cleaned using Qiaquick PCR spin column (Qiagen). The product was ligated into a modified pET-28 vector containing a C-terminal 6-His tag, in frame with the BamHI restriction site using T4 DNA ligase (New England BioLabs), and transformed into BL21(DE3) (Novagen). The expressed proteins contained the C-terminal tag GSHHHHHH, where GS is encoded by the BamHI restriction site (GGATCC). BL21(DE3) 3 ml cell culture was tested for the expression of heterologous protein by binding on a Cobalt-chelated Talon superflow bead slurry (Clontech) and SDS–PAGE analysis.

Table 1 Predicted ALDH family proteins in the *M. tuberculosis* Genome[a]

Locus tag	Protein name (probable)[b]	Putative class[c]	Rossmann motif[d]
Rv0147	Aldehyde dehydrogenase (NAD+ dependent)	ALDH-3	Yes
Rv0223c[e]	Aldehyde dehydrogenase	ALDH-1 or -2	No
Rv0234c	Succinate-semialdehyde dehydrogenase gabD1 (NADP+ dependent)	ALDH-5	No
Rv0458[e]	Aldehyde dehydrogenase	ALDH-1 or -2	Yes
Rv0753c	Methylmalonate-semialdehyde dehydrogenase mmsA	ALDH-6	Yes
Rv0768	Aldehyde dehydrogenase aldA (NAD+ dependent)	ALDH-1 or -2	No
Rv1187	Pyrroline-5-carboxylate dehydrogenase rocA	ALDH-4	No
Rv1731	Succinate-semialdehyde dehydrogenase gabD2 (NADP+ dependent)	ALDH-5	Yes
Rv2858c[e]	Aldehyde dehydrogenase	ALDH-1 or -2	No
Rv3293[e]	Piperideine-6-carboxilic acid dehydrogenase PCD (antiquintin)	ALDH-7	No

[a] Predicted ALDH proteins were identified based on the presence, and location, of the two ALDH active site signature sequences derived by Prosite (see Prosite document PDOC00068 and Fig. 5)

[b] Protein names are from the NCBI database

[c] Putative class is based on sequence homology to human ALDH proteins (using NCBI Blastp) and on available protein information in the NCBI EntrezGene database. For nomenclature of ALDH classes see [52]

[d] Rossmann motif indicates the presence or absence of the $GX_{1-2}GXXG$ sequence typically found in NAD(H)/FAD-binding proteins (see [23])

[e] Expressed ALDH proteins purified from *Mtb* lysates in this study

Cell culture was performed as described by Studier [25] with some modifications. Transformed cells were inoculated into 3 ml seed culture media (1 mM $MgSO_4$, 0.5% glucose, 17 amino acids of 100 µg/ml for each Na-Glu, Asp, Lys-HCl, Arg-HCl, His-HCl, Ala, Pro, Gly, Thr, Ser, Gln, Asn, Val, Leu, Ile Phe, Trp, metal mix of 50 µM Fe, 20 µM Ca, 10 µM Mn, 10 µM Zn, 2 µM for each Co, Cu, Ni, Mo, Se and B, 5 mM PO_4, 5 mM Na, 2.5 mM K, 2.5 mM NH_4 and 1.25 mM SO_4), and grown overnight at 37°C. From the seed culture, 500 µl was inoculated into 500 ml auto-induction media, containing 1 mM $MgSO_4$, metal mix (same as seed culture), 0.5% glycerol, 0.5% glucose, 0.2% α-lactose, NPS (same as seed culture), and 35 µg/ml kanamycin. After cells were grown at 37°C until OD_{600} reaches 0.5, the growth was continued at 20°C for approximately 16 h until the OD_{600} reached approximately 15. The cells were harvested and stored at −80°C.

The cell pellet was lysed by sonication in 10 ml of buffer A (20 mM Tris–HCl, pH 8.0, and 100 mM NaCl) per gram of cells for 10 min in 30 s pulses at 10°C. The cell debris was removed by ultra-centrifugation for 30 min at 38,000 rpm using a Ti 60 rotor (Beckman). The clear supernatant was filtered through a 0.45 µm pore membrane and loaded on a 5 ml Talon superflow affinity column equilibrated with buffer A. After washing with 30 ml buffer A and 20 ml buffer B (20 mM Tris–HCl, pH 8.0, 100 mM NaCl and 20 mM imidazole), the His-tagged Rv0223c (and the other ALDHs) was eluted from the cobalt affinity column using Buffer C (20 mM Tris–HCl, pH 8.0, 100 mM NaCl and 300 mM imidazole). The eluted fraction was dialyzed against Buffer D (20 mM Tris–HCl,

pH 8.0, 100 mM NaCl and 10 mM β-mercaptoethanol) and purified by gel filtration on a Superdex-75 column (GE Healthcare Inc.) using Buffer D for equilibration and elution. The peak fractions (monitored at OD_{280}) were analyzed by SDS–PAGE and the pooled protein fractions were concentrated using a Centricon Plus-20 (Millipore) up to 35 mg/ml, which was measured by Bradford assay with IgG (Bio-Rad) as a standard. The purity of each protein was estimated to be higher than 95% based on densitometry of SDS–PAGE gels [26].

Screening of Rv0223c for interactions with multiple nucleosides and nucleotides

Recombinant proteins were evaluated for their ligand-binding properties using a modified affinity elution chromatography protocol [27]. Individual proteins were diluted to 2 mg/ml in column buffer (CB, containing 50 mM potassium phosphate, pH 7.5, 1 mM $MgCl_2$ and 2 mM DTT) and adsorbed to multiple small aliquots of F3GA resin (100 µg protein per 10 mg resin) in 2 ml spin-columns (Costar, Fisher Scientific). Binding was for 1 h at 4°C with very gentle vortexing, followed by recovery of unbound protein (flow-through fraction) and washing of the resin (4 × 0.4 ml washes with CB); spin-columns were micro-centrifuged for 30 s at 10,000×g, to recover fractions and change solutions. Individual spin-columns containing resin-bound proteins were then incubated (as for protein binding, above) with 50 µl 1 mM test ligand in CB, and the elution fractions recovered by centrifugation. Protein which remained bound to the resin was recovered

by heating at 95°C for 5 min in 100 μl SDS sample buffer, and centrifugation (resin fraction). Aliquots of initial protein, spin-column flow-through and eluate fractions were diluted 1:1 with 2× SDS sample buffer, and loaded in equal proportion (equivalent to 1 μg input protein) on 15% gels and stained with silver.

Rv0223c protein–ligand crystallization and data collection

Crystallization experiments were carried out by the hanging-drop vapor diffusion method [28] at room temperature (25°C) using 24-well plates. Recombinant Rv0223c was tested for the presence of bound nucleotides (see Supplementary Methods); the results indicated that the protein has at most 0.1:1 bound NAD or other nucleotide (see Supplementary Fig. 1). Each protein–ligand solution was prepared by mixing protein (0.68 mM in solution of Buffer D) with the corresponding ligand (20 mM in H_2O) at a molar ratio of 1:2 protein:ligand. The mixtures were incubated at room temperature for 30 min prior to setting up crystallization experiments. The final concentration of protein in each protein–ligand mixture was between 0.60 and 0.63 mM. The ligands used were NAD, NADH, NADP, NADPH, adenosine, AMP, ADP, ATP, GTP, FAD, FMN and Cibacron Blue F3GA (free dye). Crystals were grown from drops consisting of 1 μl protein–ligand solution mixed with 1 μl of reservoir solution against a reservoir containing 0.1 M MES (pH 6.0) and 0.8 M ammonium sulfate, for 3 days at room temperature. For some ligands (e.g. ATP), the effect of Mg^{++} was tested for crystallization, but without any noticeable effect, crystallization experiments reported were executed without Mg^{++}. Native and SeMet Rv0223c-NAD complex crystals were flash-cooled in liquid N_2 with the addition of 10% glycerol in the crystallization buffer as cryoprotectant. Three-wavelength selenium multi-wavelength anomalous dispersion (MAD) data was collected at the beam line 5.0.2 at the Advanced Light Source (ALS). A native data set at a resolution of 1.8 Å was collected at the beam line 8.2.1 at the ALS. Both data sets were processed with the HKL2000 program suite [29].

Structure determination and data analysis of Rv0223c protein–ligand complexes

Initial phasing was carried out with the program SOLVE [30] using the MAD data set. The resulting experimental map was density modified and traced using the program RESOLVE [31]. The protein model was further improved and built with the ARP/WARP package [32, 33] against the 1.8 Å native data. Manual model rebuilding was carried out with programs COOT [34]. The final model of this complex is deposited in the Protein Data Bank (http://www.rcsb.org)

as entry 3B4W, and has R/R_{free} values of 0.18/0.20 at a resolution of 1.8 Å, after refinement with the phenix.refine program from the PHENIX software package [35]. Difference electron density maps were calculated for each Rv0223c-ligand complex by refining the structure of the Rv0223c protein (without ligands or solvent molecules) against the observed structure factor amplitudes for each complex. The resulting crystallographic phases were used to construct an $(mF_o - DF_\varphi)e^{i\varphi c}$ difference map [36]. The LigandFit algorithm for automated ligand-fitting in PHENIX [37] was used to identify the location of the largest contiguous regions of high density in the difference map. In this algorithm the contour level for identification of contiguous regions of density is set to a level such that the largest region is approximately the size of the anticipated ligand. In this way, the location of this largest region gives an indication of the location of the ligand. The difference electron density for each complex is shown in the region of the NAD in the Rv0223c protein-NAD complex and is displayed with PyMOL [38]. The overall structure comparison between *Mtb* Rv0223c and human mitochondrial ALDH-2 was performed by the DaliLite program [39].

Results

Sequence similarities of four *Mtb* ALDHs captured by dye-ligand chromatography to each other and to human ALDH proteins

In Table 2, we examine the degree of sequence identity among these four *Mtb* ALDHs and their closest human homologs in more detail. The sequence identities show that the four *Mtb* proteins have more sequence homology to specific human ALDHs than to each other; for example, Rv0223c, Rv0458, and Rv2858c show distinctively smaller *e*-values and higher percentages of identical residues when aligned with human ALDH class 1 and 2 protein (than when they are aligned with other *Mtb* ALDHs) by the NCBI Blastp program, and Rv3293 shows the same result with human ALDH class 7.

Analysis of the specificity of ligand–protein interactions with recombinant Rv0223c using dye-ligand chromatography

We used a modified version of our dye-resin/ligand-elution procedure to examine the specificity of ligand binding of one of the four *Mtb* ALDHs (see Materials and Methods). In this ligand-specific elution screen, recombinant Rv0223c was adsorbed on small aliquots of F3GA resin in spin-columns and assayed for elution by twelve nucleotides and nucleosides using one ligand per column. By using purified

Table 2 Homology analysis of *Mtb* ALDHs being compared with other *Mtb* ALDH or Human ALDH by the NCBI Blastp program

Mtb ALDHs	Rv0223c *e*-value[a] (seq. id/res)[b]	Rv0458 *e*-value (seq. id/res)	Rv2858c *e*-value (seq. id/res)	Rv3293 *e*-value (seq. id/res)
Rv0223c	0	1E-70 (36%/485)	2E-76 (37%/470)	6E-54 (32%/451)
Rv0458	1E-70 (36%/485)	0	9E-69 (35%/466)	7E-38 (26%/471)
Rv2858c	2E-76 (37%/470)	9E-69 (35%/466)	0	5E-51 (33%/441)
Rv3293	6E-54 (32%/451)	7E-38 (26%/471)	5E-51 (33%/441)	0
Hum ALDH1	**3E-82 (38%/477)**	**2E-100 (41%/489)**	**3E-84 (40%/459)**	4E-52 (30%/465)
Hum ALDH2	**1E-84 (39%/474)**	**2E-105 (44%/488)**	**1E-83 (37%/463)**	2E-47 (29%/467)
Hum ALDH3	2E-35 (29%/446)	9E-30 (27%/441)	2E-34 (28%/439)	4E-20 (25%/426)
Hum ALDH4	–	2E-21 (24%/443)	2E-29 (27%/457)	2E-22 (26%/440)
Hum ALDH5	1E-65 (32%/476)	7E-56 (30%/494)	9E-61 (35%/443)	6E-50 (30%/458)
Hum ALDH6	4E-50 (28%/479)	3E-40 (27%/454)	3E-39 (27%/443)	3E-44 (29%/418)
Hum ALDH7	–	4E-32 (24%/484)	9E-42 (28%/446)	**3E-125 (49%/469)**

Four *M. tuberculosis* ALDHs are, in general, more similar to their human homologs than to each other

[a] *Expected values* (*e*-values) for each pair were generated by NCBI Blastp, available at http://blast.ncbi.nlm.nih.gov/Blast.cgi, using each *Mtb* protein to query either the *Mtb* genome (upper grid) or human genome (lower grid), without sequence filters. *E* values calculations incorporate identity, similarity and gaps in the sequence alignment, and the distance over which proteins can be aligned. *E* values closer to 0 indicate higher homology, and the best matches are indicated in bold-face type

[b] seq. id/res shows percentage of identical residues from the number of aligned residues between two ALDH genes

recombinant protein and one ligand per column, we identify effects of individual ligands on the stability of the dye-Rv0223c protein complex. We expect that those ligands that cause the Rv0223c protein to elute from the resin are likely to bind specifically to the protein, although non-specific interactions could potentially also cause elution to occur [24].

Figure 1 shows that the Rv0223c protein was eluted by NAD and NADH but only weakly by NADP and NADPH. This result is consistent with the dinucleotide preferences of homologous NAD-dependent aldehyde dehydrogenases (E.C. 1.2.1.3) from other organisms [40] and is addressed further below in the context of structural data. Additionally, the elution of Rv0223c protein by AMP/ADP/ATP is consistent with adenylate-binding by other dehydrogenases (cf. [3]).

The differences in the release of Rv0223c from the affinity resin by various ligands were substantial (Fig. 1), and the elution of the purified recombinant *Mtb* ALDH with ATP is consistent with the observations from experiments with crude cytosolic extracts (Fig. 1; Table 3 of [24]. Comparing the methods used for recombinant proteins versus native proteins in cell extracts, we note that the method for cell extracts has some potential complications. As cell extracts are mixtures of many proteins of various abundances, binding and detection of minor proteins may be difficult. Moreover, some proteins retained on the column may be bound via other proteins in native complexes, rather than directly to the dye, while others may have increased or unchanged affinity for the dye when bound to

ligands [24]. Finally, the order of elution, especially when several related ligands are used, could potentially influence gel spot intensity and identification of some proteins due to depletion by earlier ligand(s). This method is therefore only useful as a positive screen; any protein eluted by a specific ligand is a possible in vivo target of that ligand, whereas failure to detect a given protein is uninformative. In contrast, our analysis of recombinant proteins eliminates most of these variables, and allows the direct comparison of a protein's affinity for an array of ligands at the dye-binding site.

Overall, Fig. 1 indicates that NAD, NADH, ATP, ADP, AMP and FAD all interact with Rv0223c protein, and that the highest affinity is for NADH and ADP. This pattern was identical in 3 replicate experiments. Further it indicates that NADP, NADPH, adenosine, GTP, FMN, and NMN interact only weakly with Rv0223c at the dye-binding site; ligand binding elsewhere in the protein cannot be ruled out.

Crystallization and X-ray diffraction of Mtb Rv0223c with multiple ligands

Ligands that bind specifically to proteins are potentially useful for protein crystallization and structure determination [41]. We therefore examined the effects of the nucleotide/nucleoside ligands shown in Fig. 1 on the crystallization of recombinant Rv0223c protein. We carried out a systematic screen for conditions that led to crystallization of Rv0223c in the presence of NAD, and found one

Fig. 1 Ligand-specific elution chromatography of recombinant Rv0223c, a putative *Mtb* aldehyde dehydrogenase. One-dimensional SDS–PAGE analysis of protein added (Rv0223c), not bound (FT, flow-through fraction) or released from Cibacron Blue F3GA resin by buffer or 12 different ligands shows the preferential release by NAD, NADH, ATP, and ADP
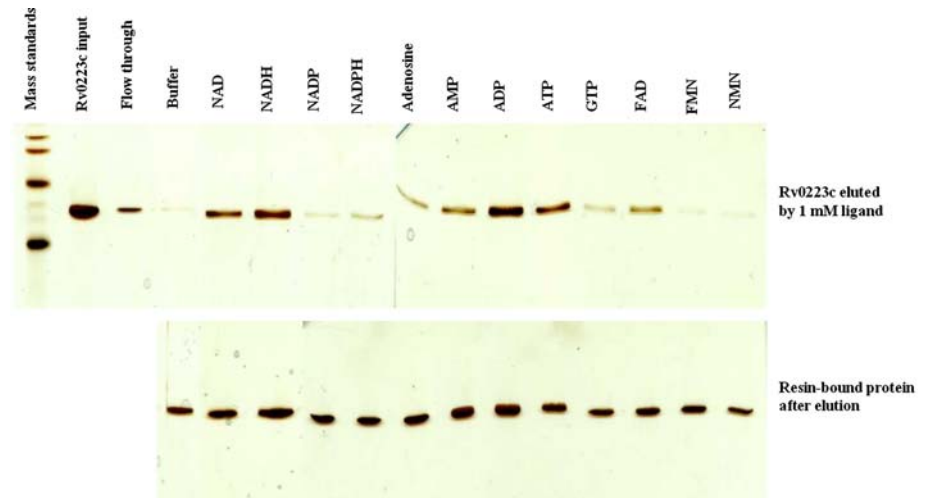
condition that led to crystals of the Rv0223c+NAD complex diffracting to high resolution (1.8 Å; see Materials and Methods; Table 3). We then carried out crystallization trials of the purified protein under this condition, but in the presence of either no ligand or the various nucleoside ligands used in the dye-ligand chromatography assay. As shown in Fig. 2, ligand-free protein generally yielded either very small or no crystals in repeated tests for crystallization (in one experiment, crystals were obtained that diffracted to a relatively low resolution of about 3 Å). In contrast, co-crystallization with several ligands (NAD, NADH, NADP, adenosine, AMP, GTP and FAD) enhanced protein crystal formation, yielding crystals diffracting to high resolution suitable for structural analysis. Other

ligands either hardly affected crystallization (NADPH, ADP and ATP) or promoted precipitation (FMN, and non-resin-bound F3GA). NMN was not tested. As summarized in Table 4, (although further screening of crystallization conditions may be required for the interactions between Rv0223c and ADP, ATP and F3GA) the strongly interacting ligands (NAD and NADH) enhanced crystal formation, leading to higher resolution data than the moderately or weakly interacting ligands (AMP, FAD, adenosine, GTP and NADP).

The structure of the Rv0223c+NAD complex was determined by MAD phasing at a resolution of 1.8 Å (PDB ID 3B4W; Table 3). The structures of all the complexes, Rv0223c+ligand (NADH, AMP, FAD, adenosine, GTP or

Table 3 Data collection and refinement statistics

	Native	MAD		
		λ_{peak}	$\lambda_{inflection}$	λ_{remote}
Data collection				
Space group	P4$_3$2$_1$2			
Unit cell parameters:	135.111, 135.111, 72.543			
a, b, c (Å)				
Resolution limits (Å)	40–1.80	50–1.9	50–2.1	50–2.0
	(1.86–1.80)[a]	(2.01–1.90)	(2.18–2.10)	(2.07–2.00)
Wavelength (Å)	1.0	0.9795	0.9800	0.9184
No. of unique reflections:	62,513	52,746	40,408	46,453
R$_{merge}^{b}$ (%)	9.1 (51.0)	9.0 (47.2)	8.6 (46.6)	8.2 (54.8)
Refinement				
Resolution limits (Å)	40–1.8			
R$_{cryst}^{c}$/R$_{free}^{d}$ (%)	16.8/19.6			
No. of protein atoms	3,577			
No. of solvent atoms	488			
No. of hetrogen atoms	54			
rmsd bond lengths (Å)	0.008			
rmsd bond angles (°)	1.21			

[a] Numbers in parentheses refer to the highest resolution shells

[b] $R_{merge} = \Sigma|I_{obs} - I_{avg}|/\Sigma I_{avg}$

[c] $R_{cryst} = \Sigma|F_o(hkl) - F_c(hkl)|/\Sigma|F_o(hkl)|$

[d] $R_{free} = R_{cryst}$, calculated for 10% randomly selected reflections that are not included in the refinement
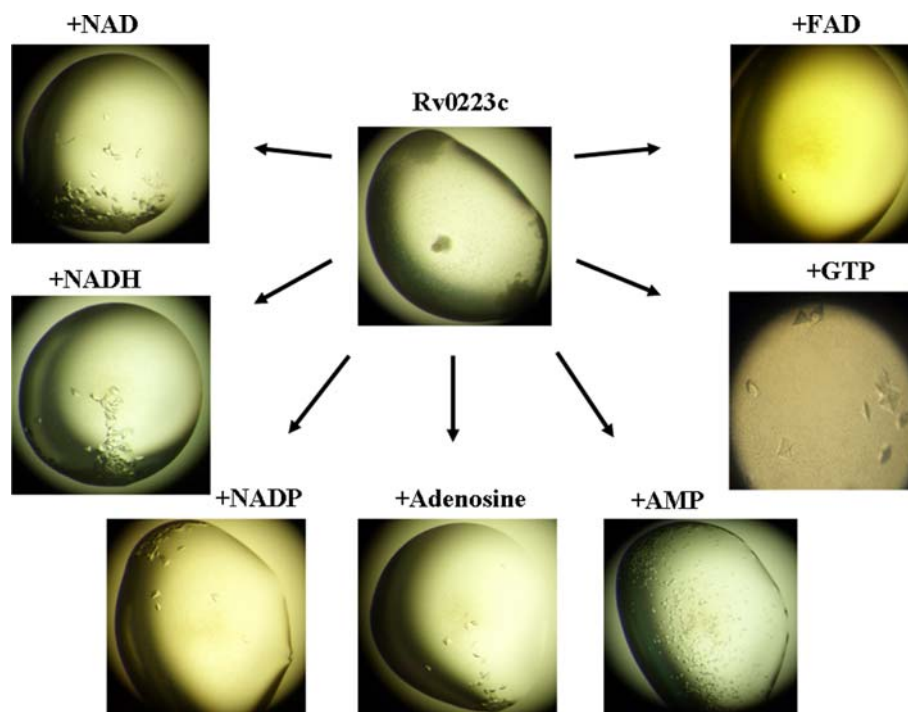
Fig. 2 The improvement of Rv0223c (putative aldehyde dehydrogenase) crystallization by addition of ligands identified by ligand-specific elution affinity chromatography method. Crystallization experiments were carried out by the hanging-drop vapor diffusion method at room temperature (295°K) using 24-well plates. Each protein–ligand solution was prepared by mixing with the corresponding ligand using molar ratio of 1:2 for protein vs. ligand, and incubated at room temperature for 30 min. In each protein–ligand mixture, the dilution factor of protein was kept less than 20% using the proper concentration of each ligand (NAD, NADH, NADP, adenosine, AMP, GTP and FAD) solution. The crystals were grown from drops consisting of 1 μL protein–ligand solution mixed with 1 μL of well solution against a reservoir containing 0.1 M MES (pH 6.0), 0.8 M ammonium sulfate, for 3 days at room temperature

Table 4 Summary of ligand elution from dye and co-crystallization of Rv0223c

Ligands	Dye-Rv0223c elution[a]	Rv0223c co-crystallization[b]
NAD	+++	+(1.8–2.0 Å)
NADH	++++	+(1.9–2.0 Å)
ADP	++++	–
ATP	+++	–
F3GA	++++	Precipitate
AMP	++	+(2.6 Å)
FAD	++	+(2.5–3.0 Å)
Adenosine	+	+(2.3–2.6 Å)
GTP	+	+(2.5–2.9 Å)
NADP	+	+(2.7–2.8 Å)
NADPH	+	–
FMN	–	Precipitate
NMN	–	N/A

[a] The estimation of ligand-specific elution of Rv0223c from Cibacron Blue F3GA dye resin in Fig. 1; very strong (++++), strong (+++), moderate (++), and weak (+)

[b] The observation for formation of crystals by co-crystallization with ligands; diffraction data is shown in parenthesis; N/A indicates no data available

NADP) were isomorphic with this one, with less than 1% changes in cell parameters. Consequently, we were able to use the structure of the Rv0223c protein (without ligand) from the Rv0223c+NAD complex to calculate phases and a difference electron density map for each of these structures (after refinement using the observed structure factors for each structure) using the diffraction data indicated in Table 4.

Location of nucleoside ligands in the Rv0223c structures

The crystals obtained with NAD, and with its reduced form NADH, were the largest and best-diffracting crystals (resolution range 1.8–2.0 Å) that we obtained in repeated experiments, and the ligand was most clearly defined in the difference electron density maps for these complexes (Fig. 3) although the crystals could potentially contain NAD because NADH showed ∼50% conversion to NAD after about 3 days in the crystallization condition used. These data suggest that Rv0223c forms the most well-defined complexes with NAD and NADH, consistent with the gene annotation (putative aldehyde dehydrogenase).

The other ligand-Rv0223c protein complexes diffracted to 2.3–3.0 Å, and for GTP, NADP and AMP, the strongest difference electron density corresponding to a bound ligand was found at the NAD-binding site (Fig. 3). In the case of adenosine-Rv0223c co-crystals, the second-largest region of contiguous density was at the NAD-binding site, but the density was weak. In a final case (FAD), no clear peaks of difference density were found. The varied electron density of ligands binding to a common site in Rv0223c suggests that the conformation of the ligands at this site may vary and that there may be considerable flexibility in binding at this site. It seems possible that the ligands that are normally involved in the function of Rv0223c might bind in a more specific fashion than those that are binding adventitiously, but we cannot be certain, as we do not know the ligand specificity of Rv0223c necessary for its function.

These structural studies provide independent evidence for various protein–ligand interactions detected using dye-ligand chromatography (Fig. 1). The crystallization data together with structural and biochemical analyses show strong and specific interactions between NAD (or NADH) and Rv0223c, while weaker interactions were also detected between Rv0223c and several other purine nucleosides/nucleotides (adenosine, AMP, GTP, NADP and FAD).

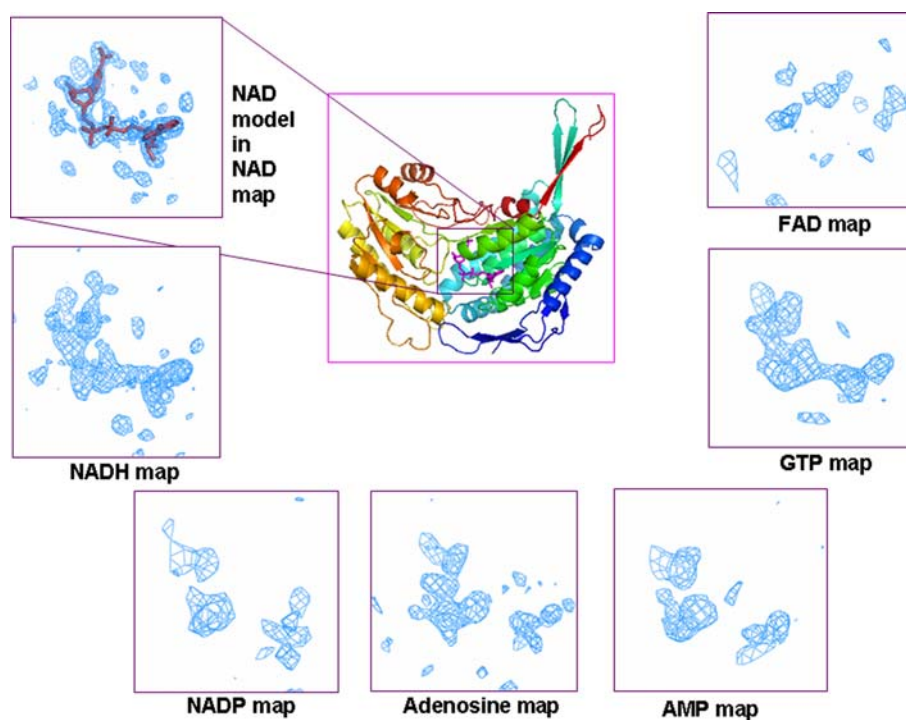Structural homology between His tagged recombinant Rv0223c and human mitochondrial ALDH

The structure of *Mtb* Rv0223c is the first reported structure from the *Mtb* ALDH family. We compared this structure with that of its human homologue, NAD-dependent mito-chondrial ALDH, ALDH-2 (PDB ID 1CW3) [42], using the DaliLite server [39] to align the structures, as shown in Fig. 4. The choice of human ALDH-2 for comparison was based on sequence similarity (Table 2), and because Rv0223c has generally been categorized as a class 2 enzyme [43, 44]. The structure of Rv0223c closely overlapped that of human ALDH-2, with an overall rmsd of 1.5 Å over 184 residues in common. Only minor regions of incongruity were found, and these were in peripheral regions of the protein (Fig. 4).

The orientation of NAD in Rv0223c was similar to human ALDH-2 (data not shown) [45]. As Perez-Miller and Hurley described, the adenine ring of NAD as located in the hydrophobic pocket formed by helices αD and αE in the coenzyme domain, with the nicotinamide end facing and interacting with residues of the catalytic domain. The amino acid residues within a 5 Å range of the NAD binding site (residues in red in Fig. 5) were substantially conserved. Of the 25 residues of Rv0223c ALDH within this range, 14 were identical to those in the human protein. The spatial locations of all 14 of these conserved residues were very close in the two structures with rmsd 1.1 Å (versus rmsd 1.5 Å for the structures overall), and had a similar orientation (data not shown), suggesting that these are critical amino acids for the interaction with NAD.

We noticed that *Mtb* Rv0223c is a monomer, whereas human mitochondrial ALDH is an octamer in crystal lattices [42, 45]. Both samples used for structural analysis were expressed in *E. coli*, but the *Mtb* Rv0223c protein has

Fig. 3 Structure of *Mtb* Rv0223c with NAD (in center; resolution range 1.8–2.0 Å), and the difference electron density map of other nucleoside ligands that promoted crystallization of Rv0223c. NADH (resolution range 1.9–2.0 Å), NADP (2.7–2.8 Å), adenosine (2.3–2.6 Å), AMP (2.6 Å), GTP (2.5–2.9 Å) and FAD (2.5–3.0 Å) identified at the binding site of NAD

Fig. 4 Comparison of crystal structure of *Mtb* Rv0223c (*green*) with NAD (*red*) against human mitochondria aldehyde dehydrogenase-2 (*orange*) structure. The homology of two structures is very high with RMSD value 1.5 Å. In Rv0223c structure, the miss-aligned sequences (*in magenta*) are minor, consisting of residues; Pro360, Glu361, Gly362, and Leu363 in the catalytic domain, Asp9, Lys10, and Leu11 in the coenzyme domain, and Gly135, Ser136, Tyr137, Gly138, Gln139, and Ser140 in the oligomerization domain. The three domains are sectioned by *purple lines*. The adenine ring of NAD (and NADH) is located in the hydrophobic pocket formed by helices αD (*gray*) and αE (*blue*) in the coenzyme domain

Fig. 5 Protein sequence alignment based on secondary structure of *Mtb* protein Rv0223c and human ALDH-2. The protein sequence alignment based on secondary structure of *Mtb* protein Rv0223c and human ALDH-2 was obtained using the DaliLite server [39]. Residues within 5 Å of NAD(H) are shown in red. The two ALDH signature sequences given by Prosite (PDOC00068) are highlighted in *yellow*. The identical amino acids between Rv0223c and human ALDH-2 are indicated with amino acid letter code and the similar amino acids are marked by '+' in the line labeled with 'Ident'. The secondary structure of each protein was indicated based on the assignment by DSSP program (H, helix; E, extended strand; L, loop or irregular) [50]

a His-tag at its C-terminus. In the human mitochondrial ALDH, the C-terminus of the protein is part of the oligomerization domain. It is possible that the His-tag of Rv0223c may interfere with oligomerization. However, gel filtration chromatography data with His-tagged and non-tagged Rv0223c proteins indicated that the His-tag does not influence the oligomerization state of the protein as reflected in the monomeric elution times of both proteins from the column (data not shown). It is more likely that the Rv0223c protein is monomeric for another reason, as the C-terminal residues of *Mtb* Rv0223c (Val486, Thr485, and Tyr484) are on the opposite side of the oligomerization domain from the side which is involved in monomer–monomer contacts based on the human ALDH-2 structure (PDB ID 1CE3). (Note that the His-tag is not visible in the electron density map and is therefore not shown in Fig. 4).

Sequence homology between *Mtb* ALDH Rv0223c and human mitochondrial ALDH

The partial amino acid sequences of Rv0223c and human mitochondrial ALDH were aligned based on secondary structure for sequence homology analysis, as shown in Fig. 5. The full length *Mtb* and human sequences are shown (487 and 518 amino acids, respectively), and were determined by NCBI Blastp to be 39% identical and 53% similar. The two ALDH signature sequences identified by Prosite [46], and including cysteine and glutamic acid active site residues, were conserved in both proteins and found at identical locations.

Discussion

The dye-resin chromatography method generates data about the interaction of protein and specific nucleoside ligand(s), and the identified ligand(s) may help stabilize the protein and improve the target protein's crystallization (Fig. 2). Structural data obtained from such crystals may also show the interactions of the ligand with the protein. In conjunction with functional information derived from the identification of ligands, this may help elucidate the biochemical role of the protein. Additionally, since many

drugs are analogues of nucleosides, proteins bound by nucleoside analogue drugs can be identified through this approach as a biomedical application (Fig. 5 of [24].

The native *Mtb* Rv0223c protein did not crystallize readily. As shown in Fig. 2, co-crystallization with NAD and NADH, which showed strong interaction by dye-ligand chromatography (Fig. 1), greatly improved the crystal quality and yielded structures of the complexes at resolutions of 1.8–2.0 Å (Fig. 3; Table 3). In contrast, NADP, adenosine, GTP, AMP and FAD, which showed relatively weak interactions (Fig. 1), generated lower resolution (2.3 Å or higher) structural data. We note that the ligands ADP and ATP, which showed strong interactions with Rv0223c, did not generate crystals (Table 4). Figure 3 reveals that ligands share a common interaction site within Rv0223c, and that Rv0223c contains a hydrophobic pocket and adenine recognition motif (composed of Gln217 and Glu239) in the coenzyme domain, which is a common motif for binding adenine-derivative coenzymes [47]. This information, in combination with our biochemical data on protein–ligand interactions (Fig. 1) and crystallization data from protein–ligand mixtures (Fig. 2), supports the conclusion that the binding of each ligand to Rv0223c occurs via contacts of the ligand to both the adenine recognition motif and hydrophobic pocket. Further, the stability of different protein–ligand crystals and their quality for high resolution structure determination depends on the degree of stabilization of protein by the interaction of the nicotinamide moiety of NAD(H) with residues in the catalytic domain.

A major bottleneck in structural genomics projects is the production of crystals suitable for analysis by X-ray diffraction. The demonstration in this work that information on ligands generated by dye-ligand chromatography can be used to identify ligands binding to nucleoside-binding proteins and to improve crystallization may be a significant contribution to overcoming this bottleneck.

There are several extensions to our method that we will make the process higher throughput. Using LC-MS (or LC-MS-MS) systems to identify native proteins in cell extracts that are eluted by specific ligands [48] may allow the protein identification process to be finished immediately after the ligand elution. In addition to Cibacron Blue F3GA dye resin, which was used in our dye-ligand chromatography approach, there are several other dye resins which are known to interact with a specific group of proteins [49]. These resins may be useful in binding groups of related proteins in a high throughput manner. The techniques described here and in Roberts et al. [24] could be applied to newly sequenced organisms to identify the metabolically important proteins that interact with nucleoside ligands, and to complement the annotation of each gene based on the interacting ligand(s) [51].

References

1. Nelson LD, Cox MM (2005) Lehninger principles of biochemistry, 4th edn. W.H. Freeman & Co
2. Voet D, Voet JG (2004) Biochemistry, vol. 1, 3rd edn. Wiley, NJ
3. Scopes RK (1994) Protein purification: principles and practice, 3rd edn. Springer, NY
4. Eppink MH, Overkamp KM, Schreuder HA, Van Berkel WJ (1999) Switch of coenzyme specificity of p-hydroxybenzoate hydroxylase. J Mol Biol 292:87–96
5. Goward CR, Nicholls DJ (1994) Malate dehydrogenase: a model for structure, evolution, and catalysis. Protein Sci 3:1883–1888
6. Wang X, Kemp RG (1999) Identification of residues of Escherichia coli phosphofructo- kinase that contribute to nucleotide binding and specificity. Biochemistry 38:4313–4318
7. George RA, Spriggs RV, Bartlett GJ, Gutteridge A, MacArthur MW, Porter CT, Al-Lazikani B, Thornton JM, Swindells MB (2005) Effective function annotation through catalytic residue conservation. Proc Natl Acad Sci USA 102:12299–12304
8. Edgar RC, Batzoglou S (2006) Multiple sequence alignment. Curr Opin Struct Biol 16:368–373
9. Kono H, Rusyn I, Yin M, Gabele E, Yamashina S, Dikalova A, Kadiiska MB, Connor HD, Mason RP, Segal BH, Bradford BU, Holland SM, Thurman RG (2000) NADPH oxidase-derived free radicals are key oxidants in alcohol-induced liver disease. J Clin Invest 106:867–872
10. Szöcs K, Lassègue B, Wenzel P, Wendt M, Daiber A, Oelze M, Meinertz T, Műnzel T, Baldus S (2007) Increased superoxide production in nitrate tolerance is associated with NAD(P)H oxidase and aldehyde dehysrogenase 2 downregulation. J Mol Cell Cardiol 42:1111–1118
11. Jelski W, Szmitkowski M (2008) Alcohol dehydrogenase (ADH) and aldehyde dehydrogenase (ALDH) in the cancer diseases. Clin Chim Acta 395:1–5
12. Wenzel P, Müller J, Zurmeyer S, Schuhmacher S, Schulz E, Oelze M, Pautz A, Kawamoto T, Wojnowski L, Kleinert H, Műnzel T, Daiber A (2008) ALDH-2 deficiency increases cardiovascular oxidative stress—evidence for indirect antioxidative properties. Biochem Biophys Res Commun 367:137–143
13. Marchitti SA, Brocker C, Stagos D, Vasilou V (2008) Non-P450 aldehyde oxidizing enzymes: the aldehyde dehydrogenase superfamily. Expert Opin Drug Metab Toxicol 4:697–720
14. Vasilou V, Nebert DW (2005) Analysis and update of the human aldehyde dehysrogenase (ALDH) gene family. Hum Genomics 2:138–143

15. Rizzo WB (2007) Sjögren-Larsson syndrome: molecular genetics and biochemical pathogenesis of fatty aldehyde dehydrogenase deficiency. Mol Genet Metab 90:1–9

16. Gibson KM (2005) Gamma-hydroxybutyric aciduria: a biochemistry's education from a heritable disorder of GABA metabolism. J Inherit Metab Dis 28:247–263

17. Kanno J, Kure S, Narisawa A, Kamada F, Takayanagi M, Yamamoto K, Hoshino H, Goto T, Takahashi T, Haginoya K, Tsuchiya S, Baumeister FA, Hasegawa Y, Aoki Y, Yamaguchi S, Matsubara Y (2007) Allelic and non-allelic heterogeneities in pyridoxine dependent seizures revealed by ALDH7A1 mutational analysis. Mol Genet Metab 91:384–389

18. Chen Z, Zhang J, Stamler JS (2002) Identification of the enzymic mechanism of nitroglycerin bioactivation. Proc Natl Acad Sci USA 99:8306–8311

19. Cederbaum AI (1989) Role of lipid peroxidation and oxidative stress in alcohol toxicity. Free Radic Biol Med 7:537–539

20. Tuma DJ (2002) Role of malondialdehyse-acetaldehyde adducts in liver injury. Free Radic Biol Med 32:303–308

21. Zhang X, Li SY, Brown RA, Ren J (2004) Ethanol and acetaldehyde in alcohol cardiomyopathy: from bad to ugly en route to oxidative stress. Alcohol 32:175–186

22. Seitz HK, Maurer B, Stickel F (2005) Alcohol consumption and cancer of the gastrointestinal tract. Dig Dis 23:297–303

23. Kleiger G, Eisenberg D (2002) GXXXG and GXXXA motifs stabilize FAD and NAD(P)-binding Rossman folds through C^α-H···O hydrogen bonds and van der Waals interactions. J Mol Biol 323:69–76

24. Roberts JKM, Webster C, Terwilliger TC, Kim C-Y (2009) High-throughput analysis of nucleoside- and nucleotide-binding by proteins. In: Hicks MG, Kettner C (eds) Proceedings of Systems Chemistry Symposium, Beilstein Institute, 26–30 May 2008, Bozen, Italy, pp 207–219 (http://www.beilstein-institut.de/index.php?id=222&L=3)

25. Studier FW (2005) Protein production by auto-induction in high-density shaking cultures. Protein Expr Purif 41:207–234

26. Watanabe T (1995) Analysis of a polyacrylamide gel electorphoretogram of beagle serum protein by laser densitometer. Lab Anim Sci 45:295–298

27. Kim C-Y, Takahashi K, Nguyen TB, Roberts JKM, Webster C (1999) Identification of a nucleic acid binding domain in eukaryotic initiation factor eIFiso4G from wheat. J Biol Chem 274:10603–10608

28. McPherson A (1999) Crystallization of biological macromolecules. Cold Spring Harbor Laboratory Press, Plainview

29. Otwinowski Z, Minor W (1997) Processing of X-ray diffraction data collected in oscillation mode. Methods Enzym 276:307–326 Academic Press

30. Terwilliger TC, Berendzen J (1999) Automated MAD and MIR structure solution. Acta Cryst D55:849–861

31. Terwilliger TC (1999) Reciprocal-space solvent flattening. Acta Cryst D55:1863–1871

32. Lamzin VS, Wilson KS (1993) Automated refinement of protein models. Acta Cryst D49:129–147

33. Perrakis A, Sixma TK, Wilson KS, Lamzin VS (1997) wARP: improvement and extension of crystallographic phases by weighted averaging of multiple-refined dummy atomic models. Acta Cryst D53:448–455

34. Emsley P, Cowtan K (2004) Coot: model-building tools for molecular graphics. Acta Cryst D60:2126–2132

35. Adams PD, Grosse-Kunstleve RW, Hung L-W, Ioerger TR, McCoy AJ, Moriarty NW, Read RJ, Sacchettini JC, Sauter NK, Terwilliger TC (2002) PHENIX: building new software for automated crystallographic structure determination. Acta Cryst D58:1948–1954

36. Read RJ (1986) Improved fourier coefficients for maps using phases from partial structures with errors. Acta Cryst A42:140–149

37. Terwilliger TC, Klei H, Adams PD, Moriarty NW, Cohn JD (2006) Automated ligand fitting by core-fragment fitting and extension into density. Acta Cryst D62:915–922

38. DeLano WL (2002) The PyMol molecular graphics system, http://www.pymol.org

39. Holm L, Park J (2000) DaliLite workbench for protein structure comparision. Bioinformatics 16:566–567 (http://www.ebi.ac.uk/Tools/dalilite/)

40. Perozich J, Kuo I, Lindahl R, Hempel J (2001) Coenzyme specificity in aldehyde dehydrogenase. Chem Biol Interact 130–132:115–124

41. Vedadi M, Niesen FH, Allali-Hassani A, Fedorov OY, Finerty JPJ, Wasney GA, Yeung R, Arrowsmith C, Ball LJ, Berglund H, Hui R, Marsden BD, Nordlund P, Sundstrom M, Weigelt J, Edwards AM (2006) Chemical screening methods to identify ligands that promote protein stability, protein crystallization, and structure determination. Proc Natl Acad Sci USA 103:15835–15840

42. Zhou NLJ, Hurley TD, Weiner H (1999) Human liver mitochondrial aldehyde dehydrogenase: three-dimensional structure and the restoration of solubility and activity of chimeric forms. Protein Sci 8:2784–2790

43. Ni L, Zhou J, Hurley TD, Weiner H (1999) Human liver mitochondrial aldehyde dehydrogenase: three-dimensional structure and the restoration of solubility and activity of chimeric forms. Protein Sci 8:2784–2790

44. Yoshida A, Rzhetsky A, Hsu LC, Chang C (1998) Human aldehyde dehydrogenase gene family. Eur J Biochem 251:549–557

45. Perez-Miller SJ, Hurley TD (2003) Coenzyme isomerization is integral to catalysis in aldehyde dehydrogenase. Biochemistry 42:7100–7109

46. Hulo N, Bairoch A, Bulliard V, Cerutti L, Cuche B, De Castro E, Langendijk-Genevaux PS, Pagni M, Sigrist CJA (2006) The PROSITE database. Nucleic Acids Res 34:D227–D230

47. Denessiouk KA, Rantanen V-V, Johnson MS (2001) Adenine recognition: a motif present in ATP-, CoA-, NAD-, NADP-, and FAD-dependent proteins. Proteins 44:282–291

48. Wysocki VH, Resing KA, Zhang Q, Cheng G (2005) Mass spectrometry of peptides and proteins. Methods 35:211–222

49. Denizli A, Pişkin E (2001) Dye-ligand affinity systems. J Biochem Biophys Methods 49:391–416

50. Kabsch W, Sander C (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22:2577–2637

51. Brylinski M, Skolnick J (2008) A threading-based method (FINDSITE) for ligand-binding site prediction and functional annotation. Proc Natl Acad Sci USA 105:129–134

52. Vasiliou V, Nebert DW (2005) Analysis and update of the human aldehyde dehydrogenase (ALDH) gene family. Hum Genomics 2:138–143