research papers

Validation and correction of Zn–Cys$_x$His$_y$ complexes

Wouter G. Touw,[a,b] Bart van Beusekom,[b] Jochem M. G. Evers,[a] Gert Vriend[a] and Robbie P. Joosten[b]*

[a]Centre for Molecular and Biomolecular Informatics, Radboud University Medical Center, Geert Grooteplein-Zuid 26-28, 6525 GA Nijmegen, The Netherlands, and [b]Department of Biochemistry, Netherlands Cancer Institute, Plesmanlaan 121, 1066 CX Amsterdam, The Netherlands. *Correspondence e-mail: r.joosten@nki.nl
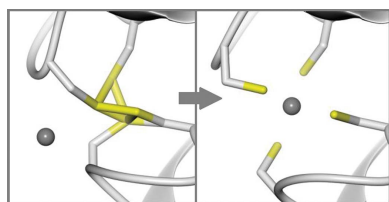
Many crystal structures in the Protein Data Bank contain zinc ions in a geometrically distorted tetrahedral complex with four Cys and/or His ligands. A method is presented to automatically validate and correct these zinc complexes. Analysis of the corrected zinc complexes shows that the average Zn–Cys distances and Cys–Zn–Cys angles are a function of the number of cysteines and histidines involved. The observed trends can be used to develop more context-sensitive targets for model validation and refinement.

1. Introduction

Many efforts have been directed towards improving the identification of ion types in macromolecular structures (see, for example, Sodhi *et al.*, 2004; Hsin *et al.*, 2008; Andreini *et al.*, 2009, 2013; Hemavathi *et al.*, 2010; Brylinski & Skolnick, 2011; Echols *et al.*, 2014; Zheng *et al.*, 2014; He *et al.*, 2015; Morshed *et al.*, 2015). The geometry of ion-binding sites often needs to be improved as well. The bond-valence method (Brown & Altermatt, 1985; Brese & O'Keeffe, 1991; Brown, 2009) that is generally used to identify ion types (Hooft, Vriend *et al.*, 1996; Nayal & Di Cera, 1996; Müller *et al.*, 2003; Zheng *et al.*, 2014) requires that the modelled geometry of the binding site accurately represents the crystallographic data.

Zinc ions (Zn^{2+}) are the most common transition-metal ions in protein crystal structures in the Protein Data Bank (PDB; Berman *et al.*, 2007; Gutmanas *et al.*, 2014) and are the second most common metal ions overall after magnesium. Zn^{2+} ions can play a largely catalytic role or a largely structural role in proteins (see, for example, Alberts *et al.*, 1998; Lee & Lim, 2008; Sousa *et al.*, 2009; Laitaoja *et al.*, 2013), but they are sometimes also found to have nonbiological functions as crystal-packing mediators. The zinc finger is the most commonly observed zinc-binding motif in the PDB (Krishna *et al.*, 2003). It is present in protein domains with diverse functions such as binding DNA, RNA, proteins or lipids (Laity *et al.*, 2001).

Structural zinc sites typically consist of four Cys and/or His ligands (see, for example, Torrance *et al.*, 2008; Laitaoja *et al.*, 2013; Daniel & Farrell, 2014) that coordinate Zn^{2+} in a tetrahedral fashion (see, for example, Simonson & Calimet, 2002; Dudev & Lim, 2003; Lee & Lim, 2008; Torrance *et al.*, 2008). Cysteines that coordinate Zn^{2+} tend to be deprotonated (Dudev & Lim, 2002; Simonson & Calimet, 2002) and are often stabilized by hydrogen bonds to backbone H^N protons

(Maynard & Covell, 2001). In some protein families anionic zinc environments are stabilized by the positive charges of arginine and lysine (Maynard & Covell, 2001).

Several studies have reported on the $Zn^{2+}-S$ and $Zn^{2+}-N$ distances observed in crystal structures in the PDB or the Cambridge Structural Database (CSD; Groom & Allen, 2014). These studies, summarized in Supplementary Table S1, indicate that Zn^{2+}-coordination geometries are rather complex and depend, for example, on the combination of ligand types (see, for example, Simonson & Calimet, 2002; Daniel & Farrell, 2014). The stereochemical restraint targets that are commonly used to refine Zn^{2+} complexes, however, still tend to be simple and undifferentiated.

We recently reported on the inaccuracies and severely distorted geometries observed in crystallographic structure models in the PDB around tetrahedral complexes in which Zn^{2+} is coordinated by four cysteines (Evers *et al.*, 2015), and the impossible chemistry that one could naively derive from such distorted complexes was described. Although the article was published in jest on April 1st, the underlying problem we described was rather serious. Many Zn^{2+} sites in the PDB poorly describe the experimental data and show structural features that are not supported by known chemistry. This can lead to misinterpretation of the protein and incorrect answers to biological questions (Touw *et al.*, 2016).

It is easy to accidentally introduce errors during the model building and refinement of zinc sites because the use of geometric restraints between Zn^{2+} and the coordinating amino acids is not yet the default in today's refinement programs, which, of course, is especially a problem at low resolution. The *PDB_REDO* databank (Joosten & Vriend, 2007) contained several entries in which distorted Zn^{2+} sites were accidentally introduced. Automatic detection of disulfide bonds can draw two Zn^{2+}-binding cysteine side chains into a cysteine bridge, leading to the aforementioned impossible chemistry. There is currently no systematic validation of distorted metal-binding sites in the PDB validation pipeline (Read *et al.*, 2011; Gore *et al.*, 2012), which leaves distorted Zn^{2+} sites mostly undetected.

We present a method to validate Zn^{2+} complexed by cysteine and histidine ligands. The validation is based on parameters that characterize the geometry of zinc complexes and is available at the *WHAT IF* (Vriend, 1990) web server and through *WHAT_CHECK* (Hooft, Vriend *et al.*, 1996). A method to improve the geometry of zinc complexes by re-refinement, and side-chain rebuilding if required, has been implemented in *PDB_REDO* (Joosten, Salzemann *et al.*, 2009) and was applied to all PDB entries with $Zn-Cys_xHis_y$ sites.

In the resulting structure models, it was observed that the ideal ion–ligand distance is not a constant, but rather a function of at least the chemical identity of the other ligands. The ideal $Zn^{2+}-S^{\gamma}$ distance, for example, shortens when more of the ligands are histidines (and thus fewer are cysteines). The ideal $S^{\gamma}-Zn^{2+}-S^{\gamma}$ angle widens when more cysteines are replaced by histidines. These observations confirm, in protein structure models, the observations made by Simonson & Calimet (2002; Supplementary Table S1) on small-molecule data and provide a starting point from which more sophisticated, context-specific, geometric restraints for Zn^{2+}-coordination sites can be developed.

2. Methods

2.1. Geometric restraint generation

The present study considered Cys or His side chains coordinating zinc in a tetrahedral fashion. These zinc-binding sites will be referred to as $ZnCys_xHis_y$, with x and y in {0, 1, 2, 3, 4} and $x + y = 4$. The ligand atoms are S^{γ} for Cys and either $N^{\delta 1}$ or $N^{\varepsilon 2}$ for His. For brevity, the latter two will be referred to as N^{δ} or N^{ε}, respectively. The Zn^{2+} double positive charge will be implicit in notations such as $Zn-N^{\varepsilon}$. With tetrahedral complexes we mean the collection of both tetrahedral and nearly tetrahedral complexes.

An automated method to properly refine metal complexes ideally includes the identification of the ion, the ligands and the preferred coordination number and geometric arrangement. The program *Zen* was created to perform all of the tasks necessary for preparing refinement scripts and parameters. *Zen* identifies putative $ZnCys_xHis_y$ complexes in PDB entries and assumes that the ion is indeed Zn and that the ligands are arranged tetrahedrally. The reader is referred to *WHAT_CHECK* (Hooft, Vriend *et al.*, 1996) or *CheckMy-Metal* (Zheng *et al.*, 2014) for validating the identity of ions when the ligands are not S^{γ}, N^{δ} or N^{ε} atoms.

Zen searches around Zn for S^{γ} atoms within 4.8 Å and $N^{\delta}/N^{\varepsilon}$ atoms within 3.8 Å. Dixon's Q-test (Dean & Dixon, 1951) is performed on the Zn–ligand distances when five or more potential coordinating atoms are found. If four ligands are left after outlier rejection, they are assumed to constitute a $ZnCys_xHis_y$ site. Complexes are discarded if (i) a different type of ligand (neither Cys S^{γ} nor His $N^{\delta}/N^{\varepsilon}$) is found close to Zn (2.9 Å or closer) and (ii) a $S^{\gamma}/N^{\delta}/N^{\varepsilon}$ ligand is found 3.25 Å or further away from Zn. In order to prevent the detection of octahedral Zn sites, such as the Zn site observed in the polyketide cyclase RemF (PDB entry 3ht2; Silvennoinen *et al.*, 2009), $ZnHis_4$ complexes are also discarded if only requirement (i) is satisfied. Additionally, all sites with at least three His ligands require all ligand atoms to be present within 3.0 Å of Zn. Clusters of tetrahedral Zn complexes in which individual S^{γ} atoms coordinate more than one Zn ion are also detected by *Zen*. The abovementioned distance cutoffs were optimized empirically to minimize the number of false positives (for example $ZnHis_6$ sites detected as $ZnHis_4$ sites) and false negatives (undetected $ZnCys_xHis_y$ sites).

The fact that many PDB file headers have missing or spurious LINK records for distorted sites as well as SSBOND records between cysteines coordinating a zinc ion (Evers *et al.*, 2015) poses a problem for the refinement program *REFMAC* (Murshudov *et al.*, 2011) which is used in *PDB_REDO*. Incorrect annotation of the covalent and metal-coordination bonds causes *REFMAC* to generate incorrect geometry restraints. The authors have contacted the developers of *REFMAC* to prevent the activation of cysteine-bridge restraints when at least one of the cysteines is also involved in

a zinc-coordination LINK record. The annotation of $ZnCys_x$-His_y complexes, however, still has to be correct and complete to prevent refinement problems. Therefore, all SSBOND and LINK records involving $ZnCys_xHis_y$ complexes are corrected by *Zen*, resulting in so-called Cys-cleaned PDB files.

Based on the re-annotated LINK records, *REFMAC* imposes distance and angle restraints during refinement. The distance-restraint targets presently are 2.340 ± 0.020 Å for $Zn-S^\gamma$, 2.057 ± 0.064 Å for $Zn-N^\delta$ and 2.058 ± 0.073 Å for $Zn-N^\varepsilon$. $Zn-S^\gamma-C^\beta$ angles are restrained to $109.000 \pm 3.000°$. $Zn-N^\delta-C^\gamma$, $Zn-N^\delta-C^\varepsilon$, $Zn-N^\varepsilon-C^\delta$ and $Zn-N^\varepsilon-C^\varepsilon$ angles are restrained to $125.350 \pm 3.000°$. The Zn–Cys distance and angle targets were already present in the *REFMAC* dictionary (Vagin *et al.*, 2004). The Zn–His distance targets were obtained from tetrahedral complexes in the MESPEUS database (Hsin *et al.*, 2008) solved at 1.6 Å resolution or better and were added to the *REFMAC* refinement dictionary. The associated $Zn-N^\delta-C^\gamma$, $Zn-N^\delta-C^\varepsilon$, $Zn-N^\varepsilon-C^\delta$ and $Zn-N^\varepsilon-C^\varepsilon$ angle targets were set to the same as the values for the $H^{\varepsilon2}$ and $H^{\delta1}$ atoms. The numeric precision in the new restraints described above is kept consistent with the existing restraints, but the significant digits do not represent the accuracy at which bond angles are determined.

The *REFMAC* dictionary currently does not provide a mechanism to add angle restraints that involve three separate compounds (*i.e.* the Zn and two coordinating residues). Therefore, the (ligand 1)–Zn–(ligand 2) angles cannot be restrained automatically. The absence of these restraints allows Zn sites to depart from tetrahedral geometry without severely violating the available geometric restraints. Additionally, without these restraints it is difficult to recover, by

refinement only, from the distorted geometries that we have described previously (Evers *et al.*, 2015). *Zen* therefore creates specific angle restraints that can be applied in refinement using the external restraints mechanism in *REFMAC* (Nicholls *et al.*, 2012). The target for $S^\gamma-Zn-S^\gamma$ angles was set to the ideal tetrahedral value of $109.5 \pm 3.0°$. Angles involving histidine are not restrained because the position of histidine side chains in Zn sites is much better defined than those of cysteine side chains because of the size and rigidity of the imidazole group.

2.2. Updates to *PDB_REDO*

The *PDB_REDO* pipeline (Joosten, Salzemann *et al.*, 2009) was extended to include the refinement of $ZnCys_xHis_y$ complexes. In the initial stage, *Zen* is run when a model contains at least one Zn ion. The *PDB_REDO* program *extractor* (Joosten, Womack *et al.*, 2009) was updated to add Zn ions to the TLS (Schomaker & Trueblood, 1968) group of the coordinating residues, provided that they are all part of the same macromolecular chain. This applies only to the TLS-group selections created by *extractor*; TLS-group selections provided by the user or extracted from the header of the PDB file are purposely left unchanged. During the initial re-refinement with *REFMAC*, the external restraints generated by *Zen* are applied with default weights. For the sake of this study, automated disulfide-bond detection in *REFMAC* was switched off to prevent *REFMAC* from generating erroneous disulfide-bond restraints when cysteine side chains are too close. As a result of our findings, *REFMAC* was updated to not generate disulfide-bond restraints if one of the cysteine S^γ atoms is involved in a LINK record. Automated cysteine-bridge detection in *REFMAC* is therefore switched back on again in the latest version of *PDB_REDO*.

Re-refinement and subsequent model rebuilding (Joosten *et al.*, 2011) can change the structure model to such an extent that previously undetected $ZnCys_xHis_y$ complexes can be identified. If this is the case, *Zen* updates the model annotation and external restraints and the second round of model refinement is extended to increase the probability of convergence. For example, the $ZnCys_4$ complex around Zn A2456 in RNA polymerase II in PDB entry 2b63 (Kettenberger *et al.*, 2006) is not detected because the $Zn-S^\gamma$ distance for Cys107 is above the detection threshold (5.70 Å). After re-refinement the distance is just below (4.73 Å) the detection threshold. Consequently, the $ZnCys_4$ complex is
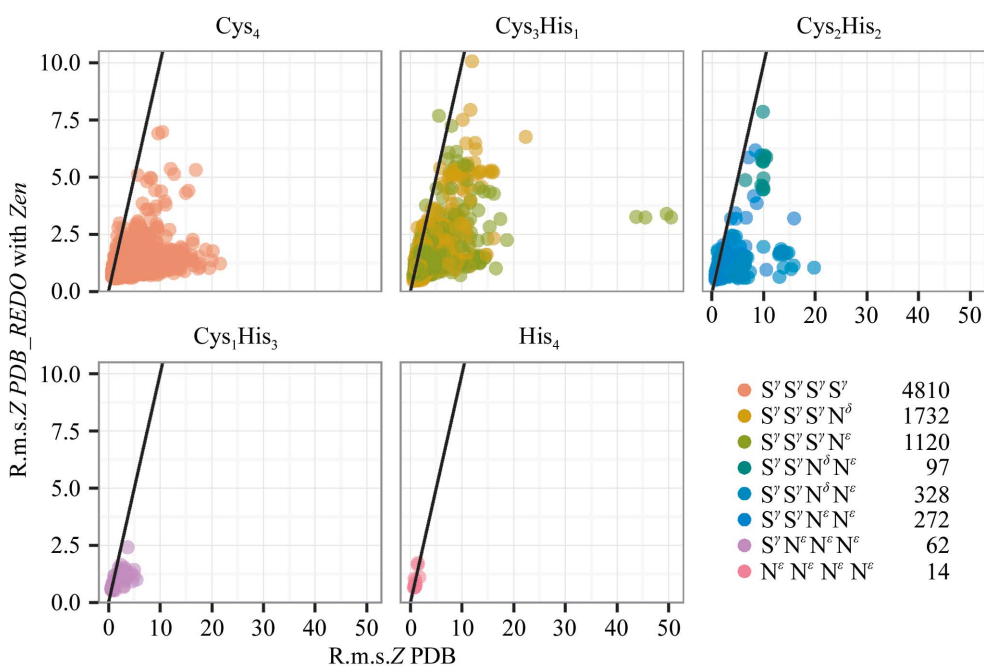
Figure 1
R.m.s.*Z* for the five possible $ZnCys_xHis_y$ site types. The scales on the two axes are different; black lines indicate the situation where the r.m.s.*Z* is the same for complexes in the PDB and after *Zen* remediation and re-refinement in *PDB_REDO*. Ligand atoms and site counts are indicated in the legend.

recognized by *Zen* and during a second round of refinement the distance decreases to 2.35 Å.

The updated *PDB_REDO* pipeline was used to replace all entries of the *PDB_REDO* databank (Joosten & Vriend, 2007) containing $ZnCys_xHis_y$ sites.

2.3. $ZnCys_xHis_y$ geometry validation

Features characterizing the $ZnCys_xHis_y$ coordination complexes were determined using *WHAT IF* (Vriend, 1990). These features included bond distances, angles, torsion angles, point charge distributions, the presence and apparent multiplicity of cysteine bridges, the Zn position in the tetrahedron, and atom occupancies and B factors. His side-chain flips (Hooft, Sander *et al.*, 1996) and crystallographic symmetry (Hooft *et al.*, 1994) can be taken into account by the validation routines. The sample mean and standard deviation of each
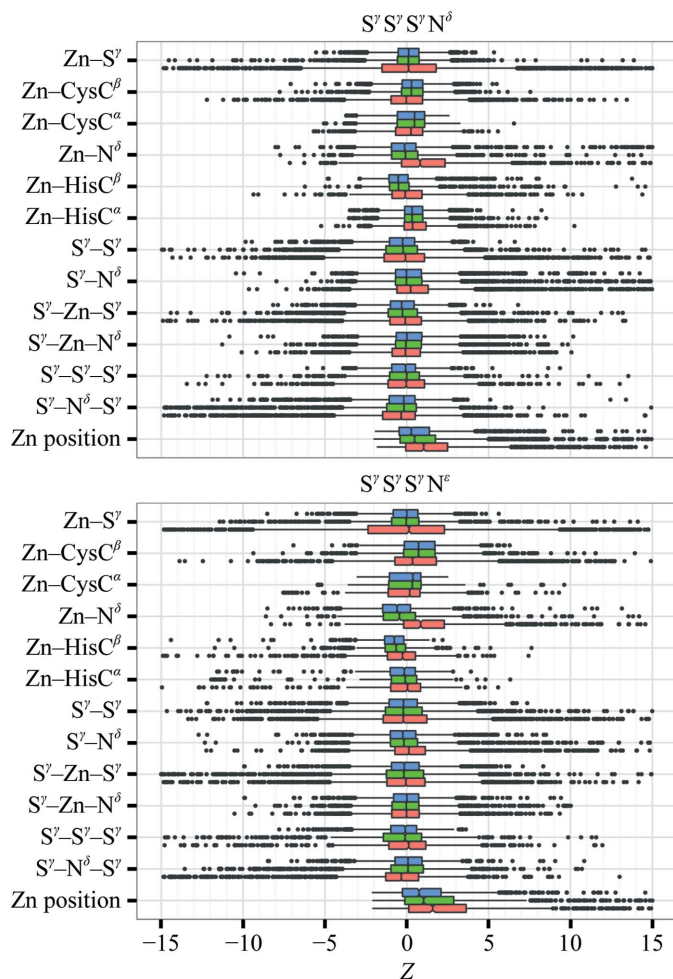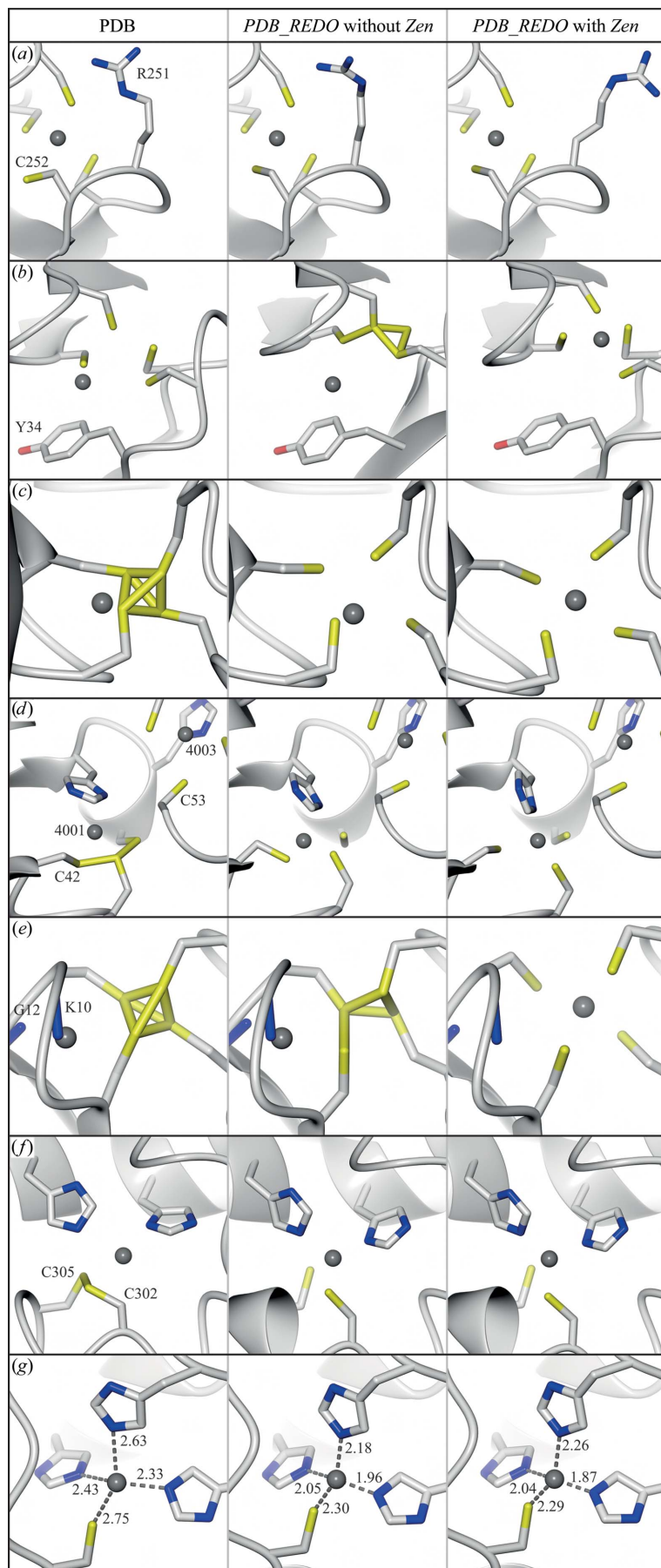

Figure 2
Box-and-whisker plots of the Z-scores characterizing $ZnCys_3His_1$ complexes in *PDB_REDO* with *Zen* remediation (blue), *PDB_REDO* without *Zen* remediation (green) and original PDB (red) structure models. The whiskers extend to the nearest value that is within 1.5 times the inter-quartile range; outliers are marked as dots. The Z score for 'Zn position' indicates the deviation from the expected Zn position in the tetrahedron. 1411 outliers with a Z-score outside $(-15, +15)$ are not shown for clarity. 891 of these outliers are from PDB structure models, while 476 and 44 outliers are from *PDB_REDO* entries without and with *Zen* remediation, respectively.

feature were determined as a function of the ligand composition. In order to prevent bias from different refinement strategies, these statistics were not derived from original sites but from sites that had been re-refined with *PDB_REDO* using the abovementioned undifferentiated restraint targets. Z-scores were calculated for the distances, angles and Zn position in the tetrahedron because manual inspection showed that these features were most indicative of the quality of the $ZnCys_xHis_y$ complex. A combined quality metric was constructed by calculating the root-mean-square Z-score (r.m.s.Z). The optimal value of an r.m.s.Z statistic varies between 0.0 at low resolution and 1.0 at high resolution (Tickle, 2007).

3. Results

3.1. The geometric quality of $ZnCys_xHis_y$ complexes is improved

8610 $ZnCys_xHis_y$ complexes were detected in 3110 PDB entries (April 20th 2016) and subjected to optimization by *PDB_REDO* with and without *Zen* remediation. The validation routines detected that 170 sites contained Zn ligands next to a chain break and that five PDB complexes [in PDB entries 4hoo (Krishnan & Trievel, 2013), 4tvr (Structural Genomics Consortium, unpublished work) and 5etx (Soumana *et al.*, 2016)] contained incompletely built Zn ligands that had been completed by *PDB_REDO*. These outliers were removed from the subsequent analyses. The 8435 tetrahedral $ZnCys_xHis_y$ complexes resulted in nearly all cases in a higher overall tetrahedral coordination geometry quality after processing by *Zen* and optimization by *PDB_REDO* (Fig. 1 and Supplementary Fig. S1). The average r.m.s.Z was 2.65 ± 9.89 for PDB complexes, 1.78 ± 2.07 after optimization without *Zen* remediation and 1.14 ± 0.60 after optimization with *Zen* remediation. The median r.m.s.Z was 1.58, 1.15 and 1.00, respectively. A median decrease of 5.59 was observed for the 10% most improved complexes. 217 complexes had an r.m.s.Z that was above 1.00 in the PDB (average 1.33 ± 0.43, median 1.20) and lower than the r.m.s.Z after *Zen* remediation (average 1.49 ± 0.60, median 1.33). Only 58 complexes had an r.m.s.Z below 1.00 (0.91 ± 0.06) in the PDB and above 1.00 in *PDB_REDO* (1.10 ± 0.10). In line with our treatment of bond-length and bond-angle r.m.s.Z scores on the *PDB_REDO* server (Joosten *et al.*, 2014), we regard these 275 complexes (3.3% of the total number of complexes) as deteriorated.

Generally, the individual Z-score components of r.m.s.Z also improved. *PDB_REDO* models after *Zen* remediation have Z-score distributions that cluster more tightly around the expected values and have fewer outliers than PDB models (to a smaller extent this is also observed for *PDB_REDO* models that have not been processed by *Zen*). This is exemplified for the features capturing the geometric quality of $ZnCys_3His_1$ complexes in Fig. 2. As expected, parameters that were directly targeted because they had been restrained (*e.g.* Zn—S^γ, Zn—N^δ and Zn—N^ε distances and S^γ—Zn—S^γ angles) or

PDB	PDB_REDO without Zen	PDB_REDO with Zen
(a) R251, C252		
(b) Y34		
(c)		
(d) 4003, C53, 4001, C42		
(e) G12, K10		
(f) C305, C302		
(g) 2.63, 2.33, 2.43, 2.75	2.18, 1.96, 2.05, 2.30	2.26, 1.87, 2.04, 2.29

Cys-cleaned ($S^\gamma - S^\gamma$ distances) on average improved most. Notably, the $Zn - S^\gamma$ Z-score distribution is essentially symmetric in the PDB, *i.e.* $Zn - S^\gamma$ distances are either too long or too short, whereas $Zn - N^\delta$ or $Zn - N^\varepsilon$ distances in the PDB are typically too long. This may be caused by the absence of a standard target in the restraint dictionaries, but, at

Figure 3

$ZnCys_xHis_y$ complexes before (left) and after *PDB_REDO* without (middle) and with (right) *Zen* remediation. Side chains are coloured by atom type; grey spheres are Zn ions. Figures were prepared with *CCP4mg* (McNicholas *et al.*, 2011). Electron-density maps were omitted for clarity and are available from the *PDB_REDO* databank. (*a*) Zn300, chain *A*, from the 8-oxoguanine DNA glycosylase MutM (PDB entry 1l1z; 1.7 Å; Fromme & Verdine, 2002). Cys252 points away from the Zn ion. The LINK between Cys252 and Zn was not annotated in the PDB model. In the *PDB_REDO* models Cys252 S^γ has moved 2.7 Å. Arg251 was refitted to a more plausible conformation only after *Zen* detected the $ZnCys_4$ site. (*b*) Zn203, chain *I*, from the RNA polymerase II–transcription factor IIB complex (PDB entry 1r5u; 4.5 Å; Bushnell *et al.*, 2004). Zn203 is modelled far away from the centre of the four S^γ ligands. The presence of a LINK record between Zn and $C^{\delta 2}$ of Tyr34 and the absence of three $S^\gamma - Zn$ LINK records in the PDB file precludes complex formation in a standard (re-)refinement. Correction of the Zn site required the Zn to move more than 5 Å. (*c*) Zn313, chain *B*, from aspartate transcarbamoylase (PDB entry 3d7s; 2.8 Å; Stieglitz *et al.*, 2009). Several types of cysteine-bridge problems exist in the PDB (Evers *et al.*, 2015), and the four cysteines next to Zn313 form an extreme example. Only three of the four necessary LINK records are specified in the original PDB file and at the same time superfluous SSBOND records are present for three of the six bridges shown. The cysteine clashes are almost resolved even without *Zen* processing thanks to the adaptations that were made to *REFMAC* as a result of our work. The additional restraints generated by *Zen* were necessary to refine the Zn position correctly. (*d*) Zn4001, chain *D*, from the DDB1–Cul4A–Rbx1–SV5V complex (PDB entry 2hye; 3.1 Å; Angers *et al.*, 2006). The three cysteines and the histidine are not arranged tetrahedrally around Zn4001 and the three cysteines appear to form one big cysteine bridge. Without *Zen* remediation the r.m.s.Z is 9.69. The correct Cys42 rotamer was found during re-refinement after processing with *Zen*, allowing better refinement of the Zn and ligand positions (final r.m.s.Z of 1.09). The Zn4003 site is located close to the Zn4001 site and has a tetrahedral conformation. In the PDB entry the distance from the C^β atom of Cys53 to Zn4001 is 4.38 Å, whereas the distance to Zn4003 is 4.20 Å. *Zen* detected correctly that Cys53 only coordinates Zn4003. (*e*) Zn61, chain *B*, from the box H/ACA ribonucleoprotein protein particle–RNA complex (PDB entry 3lwq; 2.7 Å; Zhou *et al.*, 2010). Four cysteines are tightly connected near the Zn. In the PDB entry SSBOND records are present for these cysteines, while LINK records for the Zn are found to the backbone N atoms of Gly12 and Lys10. Normal $ZnCys_4$ geometry is obtained in the *Zen*-processed *PDB_REDO* model. The ion has moved 3.5 Å. (*f*) Zn6, chain *C*, of the *Simian virus 40* large T-antigen–human p53 complex (PDB entry 2h1l; 3.2 Å; Lilyestrom *et al.*, 2006). For 12 of the 24 chains in the PDB model SSBOND records are specified between Cys302 and Cys305, while these two residues actually coordinate the Zn together with two histidines. The complex was refined correctly with and without processing by *Zen*. (*g*) Zn4, chain *B*, from the catalytic domain of human AMSH (PDB entry 3rzu; 2.5 Å; Davies *et al.*, 2011). The coordination distances are too large. The distances in the *PDB_REDO* models were closer to the expected values.

least for structure models refined by *REFMAC*, also by the presence of 'riding' H atoms on the N^{δ} or N^{ε} atoms during refinement in the absence of LINK records (that describe a bond-length target plus the explicit deprotonation of these N atoms). These H atoms push the Zn ions and the histidine N atoms apart. The median *PDB_REDO* ZnCys$_3$His$_1$ Zn—N distance is smaller than expected, most likely because the undifferentiated restraint target distances (see §2) are much shorter than the ZnCys$_3$His$_1$-specific validation targets: at 1.6 Å resolution the average overall Zn—N distance is 2.074 ± 0.056 (see below). On a more detailed level, Zn—N^{δ} distances are 2.076 ± 0.057 and Zn—N^{ε} distances are 2.065 ± 0.050 on average. Zn—C^{β} distances are not directly restrained (although Zn—C^{β} distances are influenced by Zn—S^{γ}—C^{β} angle restraints) and their median deviates more from the expected values in *PDB_REDO* complexes than in PDB complexes. The number of Zn—C^{β} distance outliers in *PDB_REDO* complexes is reduced at the same time.

The changes in geometric parameters for the other four ZnCys$_x$His$_y$ complexes are shown in Supplementary Fig. S2 and follow similar patterns.

Visual inspection showed that a lower r.m.s.*Z* corresponds to a more plausible geometry and that most of the severely distorted ZnCys$_x$His$_y$ complexes improved dramatically upon re-refinement. Special, complicated cases such as the Cys$_3$–Zn–Cys$_1$–Zn–Cys$_2$His$_1$ complex in the UBR box of E3 ubiquitin ligase (PDB entry 3nih; Choi *et al.*, 2010) and the ZnCys$_4$ site between the two Get3 chains in the Get3–Get1 complex (PDB entry 3sjb; Stefer *et al.*, 2011) were handled correctly by our method. Fig. 3 shows several examples of complex problems that were solved satisfactorily.

Taken together, it was observed that *PDB_REDO* optimization without *Zen* remediation leads to a tighter distribution of geometry scores and that the extra *Zen* processing step further improves the average geometric quality by removing additional outliers (without significantly changing the average *B* factor; see Supplementary Fig. S3). Supplementary Fig. S4 shows examples of the classes of outliers that were still observed in our data set. These challenges include false-positive detection of ZnCys$_x$His$_y$ complexes when one of the true Zn ligands is not Cys or His (Supplementary Fig. S4*a*), spurious LINKs between Zn ligands (Supplementary Fig. S4*b*; most of these problems have been resolved in the most recent version of *Zen*) and undetected His side-chain flips (Supplementary Fig. S4*c*).

The fully automated detection of missing waters is a long-standing problem in crystallography and is particularly challenging in the vicinity of metal ions (Supplementary Fig. S5).

### 3.2. ZnCys$_x$His$_y$ refinement targets are context-dependent

The Zn—S^{γ} distances and S^{γ}—Zn—S^{γ} angles were calculated as a function of ligand identity for the set of re-refined complexes from which 5σ outliers were iteratively removed. Fig. 4 shows that the refined distances and angles are different from their refinement targets and that the refined distances

and angles are not constant but are a function of the ligand composition of the ZnCys$_x$His$_y$ complex.

4. Discussion

4.1. Automated restraint generation

The feasibility of fully automatically generating refinement restraints for metal sites depends on the quality of the structure model and the prior knowledge of the correct geometry. The effect of errors in the atomic coordinates on structural interpretation of a metal site for restraint generation is less
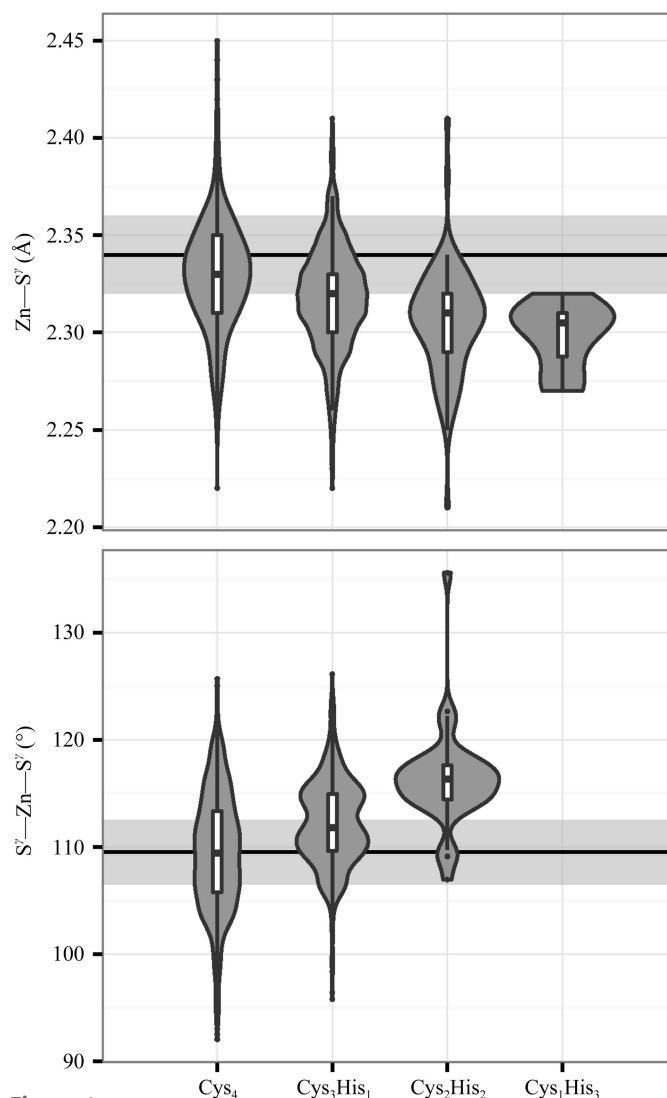
Figure 4
Zn—S^{γ} distance (top) and S^{γ}—Zn—S^{γ} angle (bottom) distributions as a function of the number of cysteines and histidines in ZnCys$_x$His$_y$ complexes determined at 1.6 Å resolution or better. The contours of the violin plots are kernel density estimates and the box plots are shown as in Fig. 2. The light grey background areas show one standard deviation around the refinement targets for the Zn—S^{γ} distance (2.340 ± 0.020 Å) and the S^{γ}—Zn—S^{γ} angle (109.5 ± 3.0°). The difference between the types of ZnCys$_x$His$_y$ complexes is significant (see Table 1). When Zn is coordinated by N^{δ} in ZnCys$_3$His$_1$ complexes, the S^{γ}—Zn—S^{γ} angle distribution is somewhat bimodal and partly depends on the rotameric state and backbone conformation of the cysteines.

severe if accurate prior knowledge is available from other experiments or data mining. Here, we show that effective restraints can be generated for Zn sites with predicted tetrahedral geometry, even when the input model is severely distorted. $ZnCys_xHis_y$ complexes have better r.m.s.Z scores after optimization by *Zen* and *PDB_REDO*. These scores are a combined measure of geometric variables in the context of an entire $ZnCys_xHis_y$ complex. The Z-score distributions seem to indicate that the total quality sometimes improves at the cost of a worse score for an individual r.m.s.Z component. This might for example be caused by incorrect restraint targets (see below), the effect of which is only problematic at low resolution, or, more generally, by difficulty in escaping local refinement minima. At the same time, however, the number of outliers decreased for all geometric variables.

If not all Zn ligands are modelled, the site will remain undetected and no restraints are generated. For catalytic Zn sites it is difficult to predict the geometry, and restraints must be made manually. Alternatively, refinement can be performed using computationally more expensive methods based on quantum mechanics (QM), such as the semi-empirical QM refinement in *PHENIX/DivCon* (Borbulevych *et al.*, 2014). Metal sites may be refined without restraints when crystallographic data are of sufficient quality and resolution.

The methods developed here can, when sufficient examples are available in the PDB, be extended to other ligand compositions of tetrahedral zinc complexes, *e.g.* Zn sites that involve water, but also to other geometries and other ion types, such as octahedral magnesium sites that are often observed in nucleic acid structures.

4.2. Validation using electron density

Improvement of a crystallographic structure model generally leads to an improvement of the corresponding electron-density map (EDM). The real-space correlation coefficient (RSCC) measures the fit of the atoms to the EDM, but correlates strongly with metrics of model precision such as the atomic B factors (Tickle, 2012). Particularly at low resolution, the RSCC metric becomes less reliable. Tickle (2012) suggested the real-space difference density Z-score (RSZD) as an EDM metric that only correlates with model accuracy and not with model precision. We did not observe a clear correlation between the geometric quality of $ZnCys_xHis_y$ complexes and their fit to the EDM measured by either the RSCC or RSZD. It was observed that a complex can have reasonable EDM metrics even when it is very bad in terms of geometry, and *vice versa*. In our hands these EDM metrics therefore were not very helpful in determining whether re-refinement of $ZnCys_xHis_y$ complexes was successful or not. The validation was therefore solely based on geometric parameters. We did observe in many cases, though, that re-refinement with inclusion of anisotropy for just the Zn ions led to visually more pleasing EDMs with less difference density around the Zn (see Fig. 5 for an example). Anisotropic atomic displacement can be partially modelled using the TLS formalism and this is currently implemented in *PDB_REDO*. Zn

and other heavy atoms may be refined with anisotropic B factors systematically in a future implementation, provided that the data-to-parameter ratio is not severely affected. This implementation may also need to include and optimize B-factor sphericity restraints in order to balance residual difference density and B-factor anisotropy.

4.3. Context-specific refinement targets

The original Engh and Huber parameters (Engh & Huber, 1991, 2001) are targets for bond lengths and angles and are averages for all conceivable situations. The very large number of high-resolution structures available from the PDB today allows fine-detailing of these parameters, as has, for example, been shown in a study on the angle τ, the $N-C^{\alpha}-C$ angle (Touw & Vriend, 2010). This large volume of data allows us to start determining better parameters for restraints for distances and angles in $ZnCys_xHis_y$ complexes. Clearly, these parameters are also determined by the local environment. For example, the $Zn-S^{\gamma}$ distance is shorter when the number of coordinating cysteines is smaller. QM calculations have suggested that this trend partly correlates with a smaller electrostatic repulsion between the thiolate S atoms and that steric and stabilizing electrostatic interactions from the secondary coordination sphere have an effect on zinc-site geometry (Simonson & Calimet, 2002; Daniel & Farrell, 2014). These findings imply that further fine-detailing will be possible as a function of the presence of nearby positive or negative groups. We indeed observe an excess of positively charged amino acids close to many, but not all, $ZnCys_xHis_y$ complexes. Counting statistics presently still preclude taking such details into account. Only when more data become available, especially at high resolution, will we be able to express target values as a function of more environmental factors and determine which environmental factors influence the target values most. The $Zn-S^{\gamma}$, $S^{\gamma}-Zn-S^{\gamma}$, $Zn-N$ and $N-Zn-N$ parameters for tetrahedral $ZnCys_xHis_y$ complexes that we

(a)   (b)
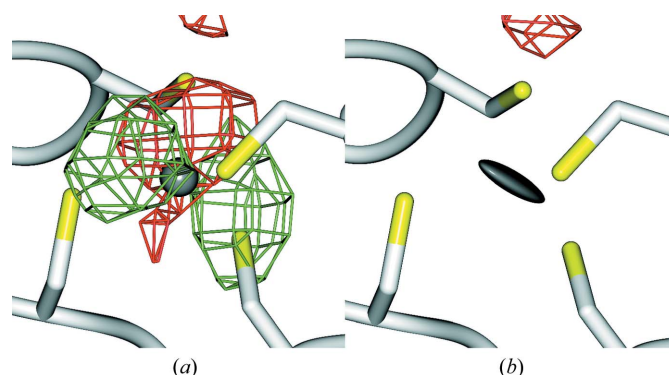
Figure 5
Zn1702, chain *B*, from jumonji H3K27 demethylase (PDB entry 4eyu; Kruidenier *et al.*, 2012). $mF_o - DF_c$ difference electron-density maps after a *PDB_REDO* run with (*a*) an isotropic B factor for Zn^{2+} (grey sphere) or (*b*) an anisotropic B factor for Zn^{2+} (grey thermal ellipsoid). The maps (positive, green mesh; negative, red mesh) are contoured at 3σ, are rendered with a grid size of 0.77 Å and for clarity are shown only in the vicinity of the Zn. The largest atomic displacement between any atom in this $ZnCys_4$ complex between (*a*) and (*b*) is 0.16 Å.

Table 1
Suggested refinement targets for the five possible $ZnCys_xHis_y$ complex types.

The targets have been derived from crystallographic structures determined at a resolution of 1.6 Å or better and are listed as mean ± standard deviation. Numbers in parentheses indicate the number of observations. For all targets a significant difference between means was observed across the types of $ZnCys_xHis_y$ complexes [one-way ANOVA with a Welch correction for nonhomogeneity of variances (Welch, 1951): $Zn-S^\gamma$ distance, $F_{(3, 49.5)} = 50.7$, $p = 4.1 \times 10^{-15}$; $S^\gamma-Zn-S^\gamma$ angle, $F_{(2, 100.3)} = 124.7$, $p \ll 10^{-15}$; $Zn-N$ distance, $F_{(2, 86.9)} = 45.5$, $p = 3.1 \times 10^{-14}$; $N-Zn-N$ angle, $F_{(1, 71.6)} = 16.6$, $p = 1.2 \times 10^{-4}$]. The same parameters derived from crystallographic structures determined at a resolution of 2.5 Å or better are given in Supplementary Table S2.

$Zn-S^\gamma$ (Å)	$S^\gamma-Zn-S^\gamma$ (°)	$Zn-N$ (Å)	$N-Zn-N$ (°)	$ZnCys_xHis_y$
2.330 ± 0.029 (1033)	109.45 ± 5.46 (1553)	n/a	n/a	Cys_4
2.318 ± 0.027 (912)	112.15 ± 3.96 (912)	2.074 ± 0.056 (303)	n/a	Cys_3His_1
2.306 ± 0.029 (76)	116.23 ± 4.58 (38)	2.040 ± 0.050 (65)	102.38 ± 5.44 (38)	Cys_2His_2
2.298 ± 0.017 (12)	n/a	2.002 ± 0.045 (36)	107.23 ± 4.78 (36)	Cys_1His_3
n/a	n/a	Insufficient data	Insufficient data	His_4

observe in the *PDB_REDO* databank in the subset of structures solved at a resolution of 1.6 Å or better are listed in Table 1.

There are not yet enough data to treat N^δ and N^ε separately and there are limited data available for $ZnCys_1His_3$ and $ZnHis_4$ sites. The parameters in Table 1 depend significantly on the type of $ZnCys_xHis_y$ complex. However, the data show signs of an underlying multimodality that we cannot yet fully resolve (Fig. 4). Nevertheless, these parameters provide a starting point for making more sophisticated sets of restraints, and the growth of the PDB and the *PDB_REDO* databank will provide more reliable statistics over time. Like many other geometric values (see, for example, Touw & Vriend, 2010), the $ZnCys_xHis_y$ values are a function of crystallographic resolution. The values that we observe for structures solved at a resolution of 2.5 Å or better (Supplementary Table S2) are slightly different from those in Table 1 but follow the trends described above.

Extracting restraints from the *PDB_REDO* databank and subsequently applying them in the *PDB_REDO* pipeline introduces circularity. This important practical issue can be avoided by only applying these restraints to low-resolution structure models (where the restraints are most needed) and not to the high-resolution structure models that will be used to derive new refinement targets. In this way, future data sets will remain unbiased. Restraint targets ideally are derived from unrestrained Zn sites, but the number of available $ZnCys_xHis_y$ complexes solved at atomic resolution will preclude the extraction of statistically significant targets from unrestrained structure models for some time to come.

5. Conclusion

The geometry of both moderately and severely distorted $ZnCys_xHis_y$ sites in the PDB could be improved substantially by restraining the sites to tetrahedral coordination geometry using both Zn–ligand distance restraints and tetrahedral $S^\gamma-Zn-S^\gamma$ angle restraints. Correcting geometry using refinement with restraints based on prior chemical knowledge and validating the results require that accurate refinement targets are known. Geometric trends in systematically re-refined $ZnCys_xHis_y$ sites show that current restraint targets may be replaced by context-specific targets. Context-specific angle

restraint targets will soon be implemented in *PDB_REDO* and context-specific distance targets will follow subject to the availability of a suitable framework for these in *REFMAC*. Geometric targets for $ZnCys_xHis_y$ sites may be further detailed once sufficient data are available.

6. Availability

The functionality to improve the refinement of $ZnCys_xHis_y$ sites is available through the *PDB_REDO* web server (Joosten *et al.*, 2014). *Zen* is distributed with *PDB_REDO* and the source code is available upon request. The *WHAT IF* web servers and web services are freely available and *WHAT IF* is shareware. *WHAT_CHECK* and *PDB_REDO* will become part of the *CCP*4 software suite (Winn *et al.*, 2011) soon. A large .csv file that contains all of the data used for analysing the 8435 tetrahedral $ZnCys_xHis_y$ complexes is available as supplementary data.

7. Related literature

The following references are cited in the Supporting Information for this article: Chung *et al.* (2005), Duan *et al.* (2009), Harding (2006), LaPlante *et al.* (2014), Ma *et al.* (2015), Samara *et al.* (2012) and Tamames *et al.* (2007).

References

Alberts, I. L., Nadassy, K. & Wodak, S. J. (1998). *Protein Sci.* **7**, 1700–1716.
Andreini, C., Bertini, I., Cavallaro, G., Holliday, G. L. & Thornton, J. M. (2009). *Bioinformatics*, **25**, 2088–2089.
Andreini, C., Cavallaro, G., Lorenzini, S. & Rosato, A. (2013). *Nucleic Acids Res.* **41**, D312–D319.
Angers, S., Li, T., Yi, X., MacCoss, M. J., Moon, R. T. & Zheng, N. (2006). *Nature (London)*, **443**, 590–593.
Berman, H., Henrick, K., Nakamura, H. & Markley, J. L. (2007). *Nucleic Acids Res.* **35**, D301–D303.

research papers

Borbulevych, O. Y., Plumley, J. A., Martin, R. I., Merz, K. M. & Westerhoff, L. M. (2014). *Acta Cryst.* D**70**, 1233–1247.

Brese, N. E. & O'Keeffe, M. (1991). *Acta Cryst.* B**47**, 192–197.

Brown, I. D. (2009). *Chem. Rev.* **109**, 6858–6919.

Brown, I. D. & Altermatt, D. (1985). *Acta Cryst.* B**41**, 244–247.

Brylinski, M. & Skolnick, J. (2011). *Proteins*, **79**, 735–751.

Bushnell, D. A., Westover, K. D., Davis, R. E. & Kornberg, R. D. (2004). *Science*, **303**, 983–988.

Choi, W. S., Jeong, B.-C., Joo, Y. J., Lee, M.-R., Kim, J., Eck, M. J. & Song, H. K. (2010). *Nature Struct. Mol. Biol.* **17**, 1175–1181.

Chung, S. J., Fromme, J. C. & Verdine, G. L. (2005). *J. Med. Chem.* **48**, 658–660.

Daniel, A. G. & Farrell, N. P. (2014). *Metallomics*, **6**, 2230–2241.

Davies, C. W., Paul, L. N., Kim, M.-I. & Das, C. (2011). *J. Mol. Biol.* **413**, 416–429.

Dean, R. B. & Dixon, W. J. (1951). *Anal. Chem.* **23**, 636–638.

Duan, J., Li, L., Lu, J., Wang, W. & Ye, K. (2009). *Mol. Cell*, **34**, 427–439.

Dudev, T. & Lim, C. (2002). *J. Am. Chem. Soc.* **124**, 6759–6766.

Dudev, T. & Lim, C. (2003). *Chem. Rev.* **103**, 773–788.

Echols, N., Morshed, N., Afonine, P. V., McCoy, A. J., Miller, M. D., Read, R. J., Richardson, J. S., Terwilliger, T. C. & Adams, P. D. (2014). *Acta Cryst.* D**70**, 1104–1114.

Engh, R. A. & Huber, R. (1991). *Acta Cryst.* A**47**, 392–400.

Engh, R. A. & Huber, R. (2001). *International Tables for Crystallography*, Vol. *F*, edited by M. G. Rossmann & E. Arnold, pp. 382–392. Dordrecht: Kluwer Academic Publishers.

Evers, J. M. G., Touw, W. G. & Vriend, G. (2015). *Evidence for Novel Quantum Chemistry to Form Triple and Quadruple Cysteine Bridges.* http://onlinelibrary.wiley.com/journal/10.1002/(ISSN)1097-0134/homepage/PROTAprilFool2015.pdf.

Fromme, J. C. & Verdine, G. L. (2002). *Nature Struct. Biol.* **9**, 544–552.

Gore, S., Velankar, S. & Kleywegt, G. J. (2012). *Acta Cryst.* D**68**, 478–483.

Groom, C. R. & Allen, F. H. (2014). *Angew. Chem. Int. Ed.* **53**, 662–671.

Gutmanas, A. *et al.* (2014). *Nucleic Acids Res.* **42**, D285–D291.

Harding, M. M. (2006). *Acta Cryst.* D**62**, 678–682.

He, W., Liang, Z., Teng, M. & Niu, L. (2015). *Bioinformatics*, **31**, 1938–1944.

Hemavathi, K., Kalaivani, M., Udayakumar, A., Sowmiya, G., Jeyakanthan, J. & Sekar, K. (2010). *J. Appl. Cryst.* **43**, 196–199.

Hooft, R. W. W., Sander, C. & Vriend, G. (1994). *J. Appl. Cryst.* **27**, 1006–1009.

Hooft, R. W. W., Sander, C. & Vriend, G. (1996). *Proteins*, **26**, 363–376.

Hooft, R. W. W., Vriend, G., Sander, C. & Abola, E. E. (1996). *Nature (London)*, **381**, 272.

Hsin, K., Sheng, Y., Harding, M. M., Taylor, P. & Walkinshaw, M. D. (2008). *J. Appl. Cryst.* **41**, 963–968.

Joosten, R. P., Joosten, K., Cohen, S. X., Vriend, G. & Perrakis, A. (2011). *Bioinformatics*, **27**, 3392–3398.

Joosten, R. P., Long, F., Murshudov, G. N. & Perrakis, A. (2014). *IUCrJ*, **1**, 213–220.

Joosten, R. P., Salzemann, J. *et al.* (2009). *J. Appl. Cryst.* **42**, 376–384.

Joosten, R. P. & Vriend, G. (2007). *Science*, **317**, 195–196.

Joosten, R. P., Womack, T., Vriend, G. & Bricogne, G. (2009). *Acta Cryst.* D**65**, 176–185.

Kettenberger, H., Eisenführ, A., Brueckner, F., Theis, M., Famulok, M. & Cramer, P. (2006). *Nature Struct. Mol. Biol.* **13**, 44–48.

Krishna, S. S., Majumdar, I. & Grishin, N. V. (2003). *Nucleic Acids Res.* **31**, 532–550.

Krishnan, S. & Trievel, R. C. (2013). *Structure*, **21**, 98–108.

Kruidenier, L. *et al.* (2012). *Nature (London)*, **488**, 404–408.

Laitaoja, M., Valjakka, J. & Jänis, J. (2013). *Inorg. Chem.* **52**, 10983–10991.

Laity, J. H., Lee, B. M. & Wright, P. E. (2001). *Curr. Opin. Struct. Biol.* **11**, 39–46.

LaPlante, S. R., Nar, H., Lemke, C. T., Jakalian, A., Aubry, N. & Kawai, S. H. (2014). *J. Med. Chem.* **57**, 1777–1789.

Lee, Y.-M. & Lim, C. (2008). *J. Mol. Biol.* **379**, 545–553.

Lilyestrom, W., Klein, M. G., Zhang, R., Joachimiak, A. & Chen, X. S. (2006). *Genes Dev.* **20**, 2373–2382.

Ma, Y., Wu, L., Shaw, N., Gao, Y., Wang, J., Sun, Y., Lou, Z., Yan, L., Zhang, R. & Rao, Z. (2015). *Proc. Natl Acad. Sci. USA*, **112**, 9436–9441.

Maynard, A. T. & Covell, D. G. (2001). *J. Am. Chem. Soc.* **123**, 1047–1058.

McNicholas, S., Potterton, E., Wilson, K. S. & Noble, M. E. M. (2011). *Acta Cryst.* D**67**, 386–394.

Morshed, N., Echols, N. & Adams, P. D. (2015). *Acta Cryst.* D**71**, 1147–1158.

Müller, P., Köpke, S. & Sheldrick, G. M. (2003). *Acta Cryst.* D**59**, 32–37.

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.

Nayal, M. & Di Cera, E. (1996). *J. Mol. Biol.* **256**, 228–234.

Nicholls, R. A., Long, F. & Murshudov, G. N. (2012). *Acta Cryst.* D**68**, 404–417.

Read, R. J. *et al.* (2011). *Structure*, **19**, 1395–1412.

Samara, N. L., Ringel, A. E. & Wolberger, C. (2012). *Structure*, **20**, 1414–1424.

Schomaker, V. & Trueblood, K. N. (1968). *Acta Cryst.* B**24**, 63–76.

Silvennoinen, L., Sandalova, T. & Schneider, G. (2009). *FEBS Lett.* **583**, 2917–2921.

Simonson, T. & Calimet, N. (2002). *Proteins*, **49**, 37–48.

Sodhi, J. S., Bryson, K., McGuffin, L. J., Ward, J. J., Wernisch, L. & Jones, D. T. (2004). *J. Mol. Biol.* **342**, 307–320.

Soumana, D. I., Kurt Yilmaz, N., Prachanronarong, K. L., Aydin, C., Ali, A. & Schiffer, C. A. (2016). *ACS Chem. Biol.* **11**, 900–909.

Sousa, S. F., Lopes, A. B., Fernandes, P. A. & Ramos, M. J. (2009). *Dalton Trans.*, pp. 7946–7956.

Stefer, S., Reitz, S., Wang, F., Wild, K., Pang, Y.-Y., Schwarz, D., Bomke, J., Hein, C., Löhr, F., Bernhard, F., Denic, V., Dötsch, V. & Sinning, I. (2011). *Science*, **333**, 758–762.

Stieglitz, K. A., Xia, J. & Kantrowitz, E. R. (2009). *Proteins*, **74**, 318–327.

Tamames, B., Sousa, S. F., Tamames, J., Fernandes, P. A. & Ramos, M. J. (2007). *Proteins*, **69**, 466–475.

Tickle, I. J. (2007). *Acta Cryst.* D**63**, 1274–1281.

Tickle, I. J. (2012). *Acta Cryst.* D**68**, 454–467.

Torrance, J. W., MacArthur, M. W. & Thornton, J. M. (2008). *Proteins*, **71**, 813–830.

Touw, W. G., Joosten, R. P. & Vriend, G. (2016). *J. Mol. Biol.* **428**, 1375–1393.

Touw, W. G. & Vriend, G. (2010). *Acta Cryst.* D**66**, 1341–1350.

Vagin, A. A., Steiner, R. A., Lebedev, A. A., Potterton, L., McNicholas, S., Long, F. & Murshudov, G. N. (2004). *Acta Cryst.* D**60**, 2184–2195.

Vriend, G. (1990). *J. Mol. Graph.* **8**, 52–56.

Welch, B. L. (1951). *Biometrika*, **38**, 330–336.

Winn, M. D. *et al.* (2011). *Acta Cryst.* D**67**, 235–242.

Zheng, H., Chordia, M. D., Cooper, D. R., Chruszcz, M., Müller, P., Sheldrick, G. M. & Minor, W. (2014). *Nature Protoc.* **9**, 156–170.

Zhou, J., Liang, B. & Li, H. (2010). *Biochemistry*, **49**, 6276–6281.