

Protonation-state determination in proteins using high-resolution X-ray crystallography: effects of resolution and completeness

S. J. Fisher,^{a,b,*} M. P. Blakeley,^b
M. Cianci,^c § S. McSweeney^d and
J. R. Helliwell^a

^aSchool of Chemistry, University of Manchester, Brunswick Street, Manchester M13 9PL, England, ^bInstitut Laue–Langevin, 6 Rue Jules Horowitz, 38042 Grenoble, France, ^cSRS Daresbury Laboratory, Warrington WA4 4AD, England, and ^dESRF, 6 Rue Jules Horowitz, 38043 Grenoble, France

* Current address: Department of Molecular Biology, Faculty of Natural Sciences, University of Salzburg, 5020 Salzburg, Austria.

§ Current address: EMBL PETRA III, DESY, Notkestrasse 85, Hamburg, Germany.

Correspondence e-mail: fisher@ill.fr

A bond-distance analysis has been undertaken to determine the protonation states of ionizable amino acids in trypsin, subtilisin and lysozyme. The diffraction resolutions were 1.2 Å for trypsin (97% complete, 12% H-atom visibility at 2.5 σ), 1.26 Å for subtilisin (100% complete, 11% H-atom visibility at 2.5 σ) and 0.65 Å for lysozyme (PDB entry 2vb1; 98% complete, 30% H-atom visibility at 3 σ). These studies provide a wide diffraction resolution range for assessment. The bond-length e.s.d.s obtained are as small as 0.008 Å and thus provide an exceptional opportunity for bond-length analyses. The results indicate that useful information can be obtained from diffraction data at around 1.2–1.3 Å resolution and that minor increases in resolution can have significant effects on reducing the associated bond-length standard deviations. The protonation states in histidine residues were also considered; however, owing to the smaller differences between the protonated and deprotonated forms it is much more difficult to infer the protonation states of these residues. Not even the 0.65 Å resolution lysozyme structure provided the necessary accuracy to determine the protonation states of histidine.

Received 9 December 2011

Accepted 22 March 2012

PDB References: trypsin, 3unr; subtilisin Carlsberg, 3unx.

1. Introduction

The advancement of synchrotron X-ray facilities, and the consequent increase in flux currently available to users at such facilities, in combination with advances in protein crystallization means that it is becoming increasingly possible to collect protein crystallographic data sets at or close to atomic resolution. These data sets are often highly complete with large multiplicities, making them excellent candidates for protonation-state determination of amino-acid groups using bond-length analysis (Ahmed *et al.*, 2007; Fisher *et al.*, 2008; Wlodawer *et al.*, 2001; Deacon *et al.*, 1997). By carefully considering the difference in bond length between, for example, a C=O bond and a C–OH bond and the associated standard deviations, it is possible to determine whether these residues are in fact protonated at a statistical significance level (Ahmed *et al.*, 2007). Ahmed and coworkers also examined the resolution ranges 0.94–1.5 Å for concanavalin A and 1.26–1.38 Å for thrombin in order to estimate the diffraction resolution at which such a bond-length analysis breaks down. Another approach was used by Ramanadham *et al.* (1993) to assess the protonation states of Asp and Glu residues. By considering the hydrogen-bond networks of COO[−] and COOH, they concluded that each protonation state has a unique hydrogen-bond network and thus identification of these networks can be used to infer the relevant protonation states.

Here, we expand on these earlier studies and consider protonation-state determinations in trypsin at 1.2 Å resolution

(97% complete), subtilisin at 1.26 Å resolution (100% complete) and lysozyme at 0.65 Å resolution (PDB entry 2vb1; Wang *et al.*, 2007; 98% complete). The protonation states in cMyBP-C (cardiac myosin-binding protein C) have been determined previously (Fisher *et al.*, 2008) and these values are included here as a lower completeness case for comparison. The estimated coordinate errors obtained using the Cruickshank DPI (diffraction-component precision index; Cruickshank, 1999) and those determined from *SHELX* full matrix inversion have been compared in order to gauge the effects of both resolution and completeness. We find that together an increase in resolution of 0.1 Å and completeness of 16% yields a twofold improvement of the estimated coordinate errors, making the assignment of protonation states much more significant. The coordinate errors from *SHELX* full matrix inversion of the lysozyme data at 0.65 Å resolution are as small as 0.008 Å.

2. Background

Determination of protonation states in proteins can provide crucial details for the determination of enzyme mechanisms. For example, the protonation states of the key catalytic residues in serine proteases determine exactly how the enzyme functions at a mechanistic level. Such protonation states are generally determined directly from either ultrahigh-resolution X-ray crystallography and/or neutron crystallography experiments (Blakeley, 2009). Determination from X-ray data requires ultrahigh resolution (better than 1 Å in most cases), where H atoms are visible in the resulting electron-density maps provided that they are sufficiently well ordered (Howard, 2004). Determination of protonation states (as deuteriums) from neutron crystallography can provide usable results even at modest resolutions of between 2 and 3 Å; however, this method has larger sample-volume requirements and it is advantageous if the sample is perdeuterated, when samples of $\geq 0.15 \text{ mm}^3$ become feasible.

An alternative method of determining protonation states in proteins is *via* bond-length analysis (Ahmed *et al.*, 2007; Fisher *et al.*, 2008; Deacon *et al.*, 1997) using high-resolution X-ray data. This has clearer analysis options based on bond distances and angles, and their σ s rather than using electron-density peak heights, or integrated electron-density peak volumes, with their σ s. Bond-length analysis is based on the inference of the presence of an H atom from analysis of other localized bonding within the residue of interest, rather than by direct observation. For example, in the case of aspartate residues the C—O^{δ1} and C—O^{δ2} bond lengths differ considerably between the protonated and the deprotonated forms (see Fig. 1). Using the difference between the bond lengths, and the associated standard deviations, it is possible to determine whether there is a real difference in bond lengths and therefore to determine whether or not a residue is protonated.

The effectiveness of this technique in determining protonation states reliably is heavily dependent on the associated bond-length standard deviations. For a normal case of an aspartate residue the difference between the C—O^{δ1} and

C—O^{δ2} bond lengths in the protonated state is 0.1 Å, and therefore a combined standard deviation on the difference of the two bond lengths of less than 0.033 Å is required to positively infer that the residue is protonated. This means that a standard deviation on each bond length of 0.024 Å is required for a 3σ determination (as the combined error for the difference of two bond lengths is the root of the sum of the two bond-length errors squared). The standard deviations within a protein structure are dependent on a number of factors, including the data resolution and completeness, as well as how well ordered the residue of interest is. Split occupancy cases are an additional challenge not addressed here.

A similar case exists for histidine, in which the bond lengths within the imidazole ring change upon protonation. This has been evaluated by Fisher *et al.* (2009), together with the carboxylic acid case, using X-ray, neutron and NMR data as well as predicted protonation states. Berisio *et al.* (1999) used X-ray crystallography to follow the titration of a histidine residue in RNaseA. The protonation states were primarily followed by electron-density maps; however, they also tracked the protonation state using the change in bond angles within the histidine ring and analysed the impact on hydrogen bonding of the changes caused by protonation/deprotonation. The pH values that they studied were purposely chosen in order to induce a change in the protonation state of the catalytic histidine residues. In particular, Fig. 6 of Berisio *et al.* (1999) shows ‘the variation of the endocyclic bond angles at C^{δ2} and C^{ε1} of the catalytic residue His12 *versus* pH’. Their estimated uncertainties of these two angles, calculated *via* inversion of the least-squares matrix, were ‘of the order of 0.8°’.

In this article, the effects of resolution and completeness on bond-length standard deviations are assessed by considering the protonation-state determinations of aspartate and glutamate residues in four proteins, cMyBP-C, trypsin, subtilisin and lysozyme, at resolutions between 0.65 and 1.30 Å and with data completeness of between 80 and 100%.

cMyBP-C is a cardiac muscle protein implicated in hypertrophic cardiomyopathy (HCM; Govada *et al.*, 2008). The X-ray data for this study extended to a resolution of 1.3 Å with a completeness of 88.4%. This slightly lower than normal completeness was mainly owing to the presence of ice rings on the diffraction images. The protonation states of aspartate and glutamate residues in cMyBP-C have been examined

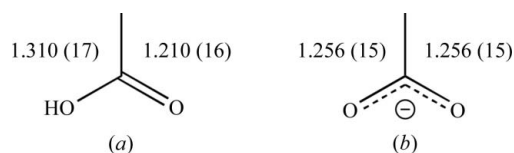


Figure 1

Bond lengths for (a) protonated and (b) deprotonated carboxylic acid groups. Data are from the Cambridge Structural Database (CSD; <http://www.ccdc.cam.ac.uk>). The values in brackets are largely a spread of values owing to the chemical microenvironment of the amino-acid side chain, but will contain a relatively small error contribution owing to crystal structure refinement.

previously (Fisher *et al.*, 2008) and the results are used here for easier comparison with the other cases tested.

The structure and function of lysozyme are well known and will not be reiterated here. This particular study (PDB entry 2vb1) is the highest resolution to date, extending to 0.65 Å. Although the full matrix inversions at this resolution have already been computed (Wang *et al.*, 2007), the bond lengths were not analysed. This data set provides a comparison case at ultrahigh resolution against the other protein structures studied.

3. Methods

3.1. Data collection and structure refinement

The data-collection strategies and structure refinements for the two new high-resolution X-ray data sets are discussed in turn below.

3.1.1. Trypsin. Trypsin was crystallized by sitting-drop vapour diffusion with a reservoir consisting of 25% PEG 8000, 0.2 M ammonium sulfate, 0.1 M Tris-HCl pH 8. Drops were composed of equal parts reservoir and protein solution. Data for the trypsin structure were collected on beamline ID23-2 at the ESRF in a quick pass and a slow pass, with the detector at 350 and 89 mm from the crystal, respectively. All data were collected at a wavelength of 0.98 Å. The slow-pass data extended to a nominal resolution of 1.2 Å and the quick pass was truncated at 2.5 Å, with 75° of data being collected for the slow pass and 85° for the quick pass. Both sets of data were independently indexed and integrated using *MOSFLM* (Winn *et al.*, 2011), giving the expected space group, namely $P2_12_12_1$, with unit-cell parameters $a = 54.57$, $b = 58.33$, $c = 67.18$ Å. The two data sets from *MOSFLM* were brought together using *CAD* and merged using *SCALA*, giving a final R_{merge} of 5.3%. Full data statistics are shown in Table 1.

Per-image R_{merge} factors were stable (~4–5%) throughout the entire slow-pass data collection. R_{merge} for the quick pass was very slightly higher (by 1–2%) as these data were collected second. Scaling B factors gently decreased by about 0.3 Å² during data collection. R_d (Diederichs, 2006) values calculated with *XDSSTAT* were stable for the first 50 images; they then showed a slight jump and remained stable again.

Table 1

Data statistics for the four X-ray data sets.

	cMyBP-C†	Trypsin	Subtilisin Carlsberg	Lysozyme†
PDB code	3cx2	3unr	3unx	2vb1
Space group	$I4_1$	$P2_12_12_1$	$P2_12_12_1$	$P1$
Unit-cell parameters				
a (Å)	48.85	54.57	52.21	27.07
b (Å)	48.85	58.33	55.21	31.25
c (Å)	95.13	67.18	75.54	33.76
α (°)	90	90	90	87.89
β (°)	90	90	90	108.00
γ (°)	90	90	90	112.11
X-ray source	SRS 10.1 (Cianci <i>et al.</i> , 2005)	ESRF ID23-2 (Flot <i>et al.</i> , 2010)	SRS 10.1 (Cianci <i>et al.</i> , 2005)	SBC-CAT 19ID (Rosenbaum <i>et al.</i> , 2006)
Wavelength (Å)	0.98	0.98	0.98	0.65
Resolution (Å)	1.30 (1.35–1.30)	1.20 (1.26–1.20)	1.26 (1.33–1.26)	0.65 (0.67–0.65)
Total reflections	98012 (11364)	212032 (28919)	518149 (58938)	1331953 (12764)
Unique reflections	22822 (3345)	65730 (9282)	59740 (7587)	187165 (6353)
Completeness (%)	88.4 (93.0)	97.4 (95.5)	100.0 (100.0)	97.6 (67.3)
Multiplicity	3.9 (2.7)	3.2 (3.1)	8.7 (6.9)	7.1 (2.7)
$\langle I/\sigma(I) \rangle$	14.0 (2.0)	6.3 (4.5)	5.5 (1.2)	36.2 (4.2)
R_{merge} (%)	4.6 (50.0)	5.3 (14.9)	9.1 (57.0)	4.5 (18.4)

† Details for cMyBP-C (Fisher *et al.*, 2008) and lysozyme (Wang *et al.*, 2005) are provided for ease of comparison.

Table 2

Refinement parameters and statistics for the four X-ray data sets.

	cMyBP-C†‡	Trypsin	Subtilisin Carlsberg	Lysozyme†
Resolution (Å)	1.30	1.20	1.26	0.65
Completeness (%)	88.4	97.4	100	97.6
No. of reflections	22822	65730	59740	187165
No. of R_{free} reflections	1184	3299	3011	9365
R (final) (%)	16.4	10.5	11.8	8.39
R_{free} (final) (%)	20.9	12.4	13.9	9.52
R (<i>SHELX</i>) (%)	—	11.7	12.8	—
R_{free} (<i>SHELX</i>) (%)	—	14.5	16.3	—
R.m.s.d.				
Bonds (Å)	0.025	0.011	0.011	0.016
Angles (°)	2.043	1.559	1.497	2.000
No. of atoms	1024	2253	2340	1323
No. of solvent atoms	176	337	265	170
B factors (Å ²)				
All atoms	24.4	10.0	12.4	7.1
Protein	18.2	7.6	10.5	6.0
Solvent	28.4	23.4	26.8	14.3
Cruickshank DPI on R (R_{free}) (Å)	0.068 (0.063)	0.027 (0.026)	0.032 (0.033)	0.005 (0.005)
Ramachandran				
Favoured	96.7	98.5	96.9	88.5
Additional	3.3	1.5	3.1	11.5
Outliers	0	0	0	0
Z	1	1	1	1

† Details for cMyBP-C (Fisher *et al.*, 2008) and lysozyme (Wang *et al.*, 2005) are provided for ease of comparison. ‡ Although the data extended to 1.20 Å into the corners of the detector, the final model was truncated to 1.30 Å, which was more appropriate for preserving completeness levels in the outer shell.

These values also indicated that little radiation damage had occurred.

Initial phases were calculated from PDB entry 1n6x (Ravelli *et al.*, 2003) after removal of the water molecules. The model coordinates were initially random atom shifted by 0.15 Å using *MOLEMAN2* (Kleywegt, 2003) in order to remove any model bias. Rigid-body refinement was then conducted to a resolution of 2.5 Å using *PHENIX* (Adams *et al.*, 2010), yielding an R factor of 29.4% and an R_{free} of 30.0%. Following this, restrained refinement was carried out to 1.5 Å resolution, yielding an R factor of 24.8% and an R_{free} of

27.4%. The resulting difference electron-density map was contoured at 3σ and used to identify 337 water molecules and a series of amino-acid side chains in multiple conformations. A further series of restrained refinements was carried out and used to assign further cases of multiple-conformation side chains as well as a number of sulfate ions. Restrained refinement was continued until convergence, resulting in a final R factor of 10.5% and an R_{free} of 12.4%. Full refinement statistics are shown in Table 2.

3.1.2. Subtilisin Carlsberg. Subtilisin Carlsberg was crystallized by taking 10 mg ml^{-1} subtilisin and adding 13% Na_2SO_4 . Prior to data collection, crystals were transferred into a cryobuffer containing 25% glycerol. Data for the subtilisin structure were collected on beamline 10 at the Daresbury Synchrotron Radiation Source (SRS; Cianci *et al.*, 2005) in a slow and a quick pass, with the detector at 90 and 140 mm from the crystal, respectively. All data were collected at a wavelength of 0.98 Å. The slow-pass data extended to a nominal resolution of 1.26 Å and the quick pass was truncated at 2 Å, with 180° of data being collected for both the slow and the quick passes. Both sets of data were independently indexed and integrated using *MOSFLM*, giving the same space group, namely $P2_12_12_1$, with unit-cell parameters $a = 52.21$, $b = 55.21$, $c = 75.54$ Å. The data sets were brought together using *CAD* and merged using *SCALA*, giving a final R_{merge} of 9.1%. Full data statistics are shown in Table 1.

The per-image R_{merge} factors were relatively well behaved for the first 90° of the slow pass ($\sim 8\%$), after which they began to rise steadily to a maximum of around 15%. The R_{merge} factors were constant for the quick pass, with values of around 8%. The per-image B factors gently decreased during data collection for the slow pass by about $0.3\text{--}4 \text{ \AA}^2$. The R_d values showed a general linear trend during data collection from around 8 to 16%. These values indicated that some radiation damage had occurred during data collection.

Initial phases were calculated from PDB entry 1c3l (Prangé *et al.*, 1998) after removal of Xe atoms and water molecules. Rigid-body refinement was conducted using *PHENIX* to a resolution of 2.5 Å, yielding an R factor of 33.0% and an R_{free} of 35.2%. Restrained refinement was then carried out and the resulting $F_o - F_c$ electron-density difference map was contoured at 3σ and used to assign 265 water molecules. Refinement was continued until convergence, resulting in the assignment of six split-occupancy amino acids. Further rounds of restrained refinement were conducted after adding H atoms in riding positions, adding further bound waters and assigning further multiple conformations, resulting in a final R factor of 11.8% and an R_{free} of 13.9%. Full refinement statistics are shown in Table 2.

3.2. Protonation-state determination using *SHELX97*

The coordinate-file format was converted to an input format suitable for *SHELX* (Sheldrick, 2008) and the $\text{C}-\text{O}^{\delta/\gamma 1}$ and $\text{C}-\text{O}^{\delta/\gamma 2}$ bonds of aspartic acid and glutamic acid were unrestrained for each of the four proteins. Ten cycles of conjugate-gradient least-squares (CGLS) refinement were

then conducted. The determined bond lengths were extracted using the *CCP4* program *DISTANG*. In order to be certain that the unrestrained Asp and Glu bond lengths were reproducible, a cyclic system of adding and removing restraints was used. Firstly, each structure was refined for ten cycles with restraints removed, allowing the determination of unrestrained bond lengths. The restraints were then reapplied, allowing the bond lengths to converge back to the restrained values. The restraints were then again removed, allowing convergence to new unrestrained values. This procedure was repeated five times. For the subtilisin structure the average standard deviation for all Asp and Glu residues was 0.005 Å. As the bond lengths converged back to very similar values, the procedure shows that the unrestrained values are consistent.

Following this, a least-squares full matrix inversion was computed using the *SHELX* L.S. 1 and ACTA keywords in order to determine the standard deviations associated with each bond length. The differences between the $\text{C}-\text{O}^{\delta/\gamma 1}$ and $\text{C}-\text{O}^{\delta/\gamma 2}$ bonds were calculated, and using the standard deviations a significance level for protonation of the residue was calculated using

$$\text{significance level} = \frac{|l_{\text{C}-\text{O}^{\delta/\gamma 1}} - l_{\text{C}-\text{O}^{\delta/\gamma 2}}|}{(\sigma_{l_{\text{C}-\text{O}^{\delta/\gamma 1}}}^2 + \sigma_{l_{\text{C}-\text{O}^{\delta/\gamma 2}}}^2)^{1/2}}, \quad (1)$$

where $l_{\text{C}-\text{O}^{\delta/\gamma 1}}$ and $l_{\text{C}-\text{O}^{\delta/\gamma 2}}$ are the lengths of the $\text{C}-\text{O}^{\delta/\gamma 1}$ and $\text{C}-\text{O}^{\delta/\gamma 2}$ bonds, respectively, and $\sigma_{l_{\text{C}-\text{O}^{\delta/\gamma 1}}}$ and $\sigma_{l_{\text{C}-\text{O}^{\delta/\gamma 2}}}$ are the standard deviations of these bond lengths.

A residue was considered to be protonated if the significance level was greater than 3σ . Finally, the bond lengths were compared against Engh & Huber (1991) dictionary values as a final check that the bond lengths were in reasonable agreement with those for protonated or deprotonated residues.

4. Considerations and limitations

4.1. Radiation damage

A significant factor in determining protonation states from high-resolution X-ray data is radiation damage. When collecting X-ray data at atomic resolution there is an increased risk of radiation damage. Aspartate and glutamate residues are some of the first to be affected by such radiation damage and can be decarboxylated. It is therefore important to assess whether radiation damage has occurred before considering the protonation states of these residues.

A number of experimental factors can indicate the occurrence of radiation damage. These include the per-image R_{merge} factor, which increases with radiation damage, the per-image B factor and per-image scale factors. A more direct measure of radiation damage is the Diederichs R factor R_d (Diederichs, 2006). This compares reflections measured against frame-number difference. One can also compare the B factor of the residue of interest (also known as the atomic displacement factor) with the average protein B factor, although this obviously includes thermal motion and possibly static disorder. Furthermore, analysis of disulfide bonds can be used to assess radiation damage. Disulfide bonds are readily

radiolysed by irradiation and their resulting increased separation can be an indicator of radiation damage. Large increases in separation indicate radiation damage, and in severe cases it is often possible to observe the effect of radiation damage directly in the difference electron-density maps. Full investigation of such effects involves keeping later diffraction images separate from earlier ones and analysing difference Fourier maps; an example of disulfide radiation damage in despentapeptide insulin can be found in Helliwell (1988).

4.2. *B* factors: correlation and increase under irradiation

The *B* factor of a residue has a clear effect on the associated standard deviation of the coordinates of its atoms. Residues in more mobile regions have increased *B* factors and hence an increased positional standard deviation. Furthermore, as stated above, the *B* factor increases in regions where radiation damage has occurred. It is therefore important to consider the per-residue *B* factor compared with the average protein *B* factor when considering the protonation states of residues. This can also be modelled as a decrease in atom occupancy.

5. Results

The following sections describe the results for the aspartate residues in the four proteins studied. Data for the glutamate residues have been omitted in order to keep the Results section succinct; these figures have instead been deposited as Supplementary Material¹. The catalytic Glu35 details for lysozyme are given in §5.4.

5.1. cMyBP-C

The standard deviations determined for cMyBP-C are all rather high, averaging around 0.09 Å. These result in relatively poor protonation-state determinations. Despite the high standard deviations, a number of the Asp bond lengths are well separated, indicating protonation. Fig. 2 shows the C—O bond lengths, standard deviations and *B* factors for each Asp residue in cMyBP-C. Asp151 is not considered as it is the terminal residue and hence is intrinsically disordered. All of the aspartate residues in cMyBP-C are on the surface of the protein and as such are more likely to have higher *B* factors. For example, Asp152 shows C—O^{δ1/2} bond lengths that are in almost perfect agreement with the Engh and Huber dictionary values. However, the large standard deviations result in a σ level of only 0.73 σ . The best determination is at 1.61 σ for Asp214, a σ level where the protonation state cannot be definitively inferred. The high standard deviations are partly a consequence of the incompleteness of the data set. Comparison of B_{avg} for the protein with the *B* factors of individual residues seem to indicate that some radiation damage may also have occurred. However, this could also arise from intrinsic thermal motion of the particular residues.

¹ Supplementary material has been deposited in the IUCr electronic archive (Reference: YT5040). Services for accessing this material are described at the back of the journal.

5.2. Trypsin

Owing to the high completeness of the data, even in the outer shell, and the resolution limit, the standard deviations from the full matrix inversion averaged 0.02 Å. These are excellent values for inferring the protonation states of residues with a high confidence level. However, the bond-length separations of the Asp and Glu residues are at best 0.064 Å, compared with the expected 0.1 Å for a fully protonated Asp or Glu residue. This means that despite the excellent standard deviations the best protonation-state determination is only at 2.26 σ . It could also mean that a proportion of the crystal has a protonated residue and a complementary proportion has a deprotonated residue. Fig. 3 shows the C—O bond lengths, standard deviations and *B* factors for the Asp residues in trypsin. Asp153 and Asp165 have been excluded from the analysis as both these residues have multiple conformations. Most of the aspartate residues in trypsin are located within the protein core. The exception is Asp71; it is therefore to be expected that this residue would have a slightly higher *B* factor, and hence the *B*-factor being above B_{avg} does not

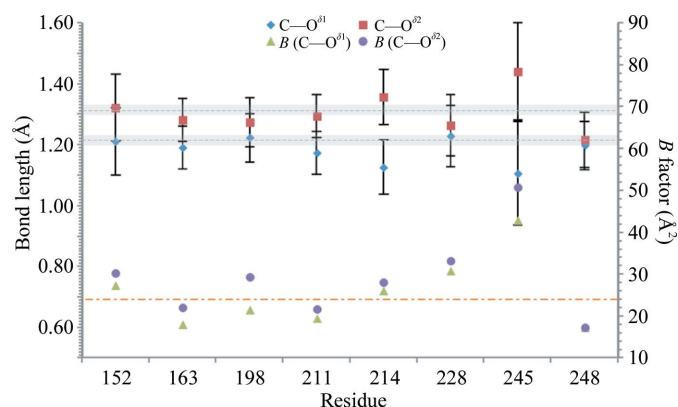


Figure 2
C—O bond lengths for Asp residues in the cMyBP-C C1 domain at 1.3 Å resolution. $B_{\text{avg}} = 24.4 \text{ Å}^2$, $\sigma_{l(\text{avg})} = 0.09 \text{ Å}$. Adapted from Fisher *et al.* (2008). Standard deviations are shown as error bars. Grey dotted lines show the Engh and Huber values for C=O and C—OH bond lengths and the grey region shows one standard deviation (natural spread) from these values. The orange dotted line shows the average *B* factor for the protein.

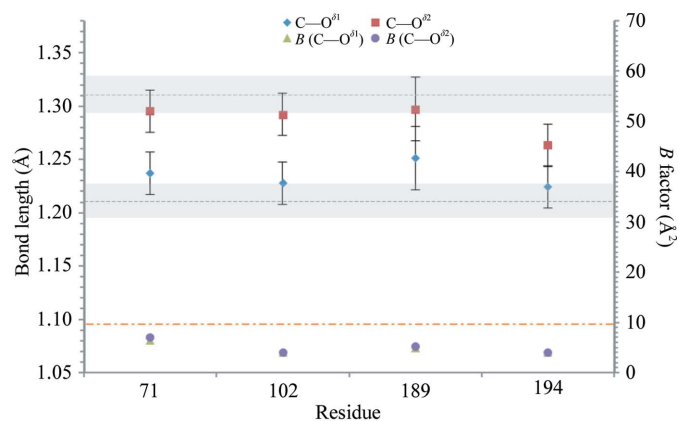


Figure 3
C—O bond lengths for Asp residues in bovine pancreatic trypsin. $B_{\text{avg}} = 10.0 \text{ Å}^2$, $\sigma_{l(\text{avg})} = 0.02 \text{ Å}$.

necessarily imply radiation damage. Comparing the B factors of the residues of interest with those of localized residues shows similar B -factor values.

5.3. Subtilisin Carlsberg

Once again the subtilisin data set has excellent completeness and extends to high resolution; the bond-length standard deviations average 0.04 \AA . This is a slightly higher value than for the trypsin data, primarily owing to the slightly lower resolution of the data set. Despite these slightly higher values, the Asp and Glu bond lengths are well separated for a number of residues, resulting in some 3σ protonation-state determinations. Asp172 does appear to show some signs of radiation damage, with elevated B factors and bond lengths that are far from ideal Engh and Huber values. However, this could also be a consequence of intrinsic thermal motion of this particular residue. The best determination is at 3.06σ for Asp181. Fig. 4 shows the C—O bond lengths, standard deviations and B factors for all of the Asp residues in subtilisin.

The variation in B factors for the Asp residues can also be explained in terms of their local environments. All of the aspartate residues in subtilisin Carlsberg are located on the surface of the protein, with the exceptions of Asp32, Asp41 and Asp60, which are buried. Therefore, the B factors of the buried residues would be expected to be lower than those of the surface residues. Therefore, for most of the surface aspartate residues the above-average B factors may arise from higher thermal motion rather than radiation damage. However, for Asp172 the bond-length values still seem to indicate that radiation damage may have occurred, as they are 4.3σ and 3.5σ , respectively, from the dictionary bond lengths. The B factors of the residues of interest are similar to those of other localized residues.

5.4. Lysozyme (Wang *et al.*, 2007)

Lysozyme is a well studied enzyme; its catalytic activity is targeted to bacterial cell walls and is related to general non-specific organism defence. Glu35 and Asp52 are involved in this mechanism and their protonation states are known from neutron protein crystallography experiments (Mason *et al.*,

1984; Bon *et al.*, 1999): Glu35 is protonated and Asp52 is deprotonated. The lysozyme data are an example of an ultrahigh-resolution X-ray data set and as such the protein is determined with extremely high precision. The bond-length standard deviations average at 0.008 \AA , which is about five times smaller than for the other protein structures. The majority of the Asp residues appear to be deprotonated, with differences in bonds lengths of around 0.01 \AA for Asp18, Asp66 and Asp119.

The catalytically active residue Asp52 shows a 3σ assignment, but the difference in the two bond lengths is only 0.05 \AA . As mentioned above, this may arise from a proportion of the crystal having this residue protonated and a complementary proportion not being protonated. The 3σ assignment arises because of the small standard deviations owing to the very high resolution of the structure used. In this case, it is also consistent with the view (see below) that the proton is positioned between the Asp and an adjacent residue, resulting in a slight lengthening of one of the bonds and a shortening of the other. Glu35 shows better signs of protonation, with bond lengths $\text{C—O}^{\delta 1} = 1.225 \text{ \AA}$ and $\text{C—O}^{\delta 2} = 1.310 \text{ \AA}$, resulting in a σ level of 15.6σ , which is in agreement with the neutron data (Mason *et al.*, 1984). Fig. 5 shows the C—O bond lengths, standard deviations and B factors for all of the Asp residues in lysozyme.

Asp101 is protonated, with a 14σ assignment and with bond lengths close to the dictionary values (1.178 \AA , 0.8σ from the dictionary C—OH bond length, and 1.318 \AA , 2.4σ from the C=O bond length).

In terms of B -factor analysis, all of the aspartate residues in lysozyme are located on the surface of the protein with the exception of Asp66. Therefore, it is not unexpected that these surface residues should have slightly higher than average B factors and hence these values do not necessarily indicate radiation damage.

5.5. Analysis of histidine residues

Histidine is another interesting case catalytically for determining protonation states. Berisio *et al.* (1999) showed how the protonation state of a histidine residue could be inferred from

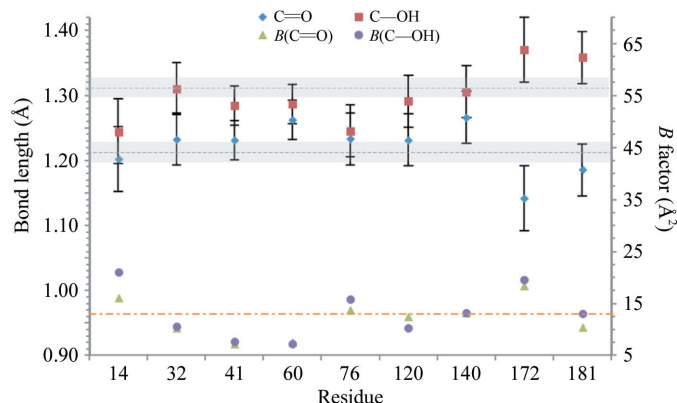


Figure 4
C—O bond lengths for Asp residues in subtilisin Carlsberg. $B_{\text{avg}} = 12.4 \text{ \AA}^2$, $\sigma_{l(\text{avg})} = 0.04 \text{ \AA}$.

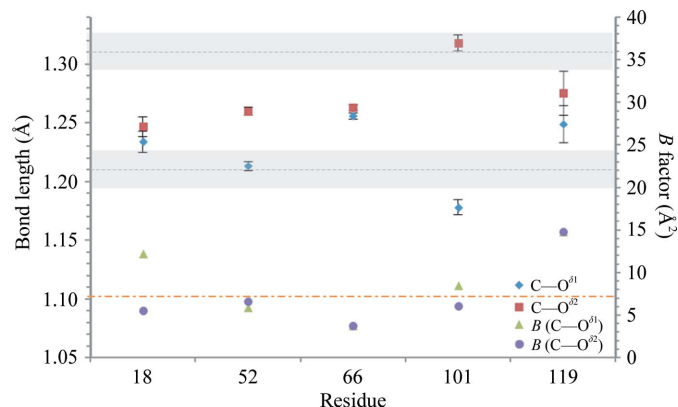


Figure 5
C—O bond lengths for Asp residues in lysozyme. $B_{\text{avg}} = 7.1 \text{ \AA}^2$, $\sigma_{l(\text{avg})} = 0.008 \text{ \AA}$.

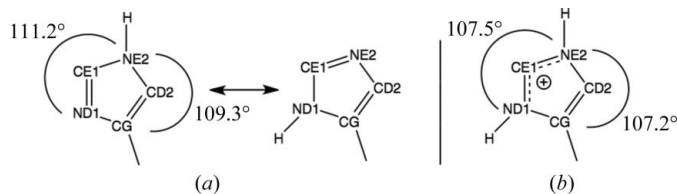
Table 3

Average σ and B values for each of the four proteins studied along with the best protonation-state determination.

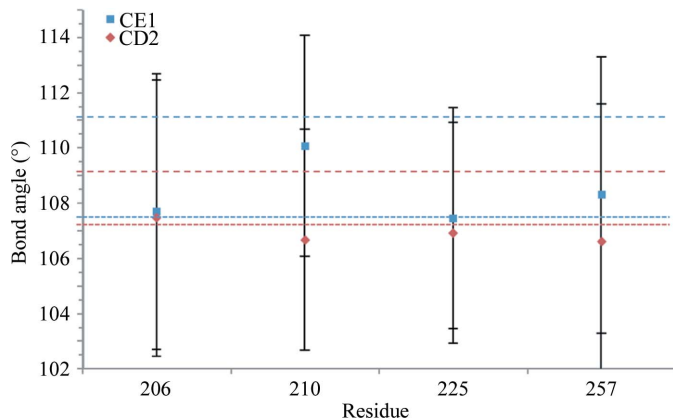
	cMyBP-C	Trypsin	Subtilisin Carlsberg	Lysozyme
Resolution (\AA)	1.3	1.2	1.26	0.65
Completeness (%)	88.4	97.7	100	97.6
$\sigma_{I(\text{avg})}$ (\AA)	0.09	0.02	0.04	0.008
B_{avg} (\AA^2)	24.4	9.95	12.4	7.1
Best determination	Asp214	Asp102	Asp181	Asp101
$\Delta(\text{C—O})$ (\AA)	0.217	0.064	0.173	0.140
$\sigma[\Delta(\text{C—O})]$ (\AA)	0.135	0.028	0.057	0.010
Significance level (σ)	1.61	2.26	3.06	14.5

the $\text{C}^{\delta 2}$ and $\text{C}^{\epsilon 1}$ bond angles. In the protonated form the two angles are virtually identical, with values of 107° . In the deprotonated form the values differ by around 2° . These values are in accordance with the Engh and Huber dictionary bond angles, as shown in Fig. 6.

Despite there being a 2.9° difference between the bond angles in the protonated and deprotonated forms, the precision required to positively infer the protonation state of these residues means that the standard deviations on each bond angle need to be around 0.65° for a 3σ protonation-state determination (assuming that both bond-angle standard deviations are the same). Furthermore in order to determine the orientation of the singly protonated histidine residue there is a difference of 1.9° between the two bond angles within the residue, thus requiring a standard deviation on each bond angle of 0.44° for a 3σ determination.

**Figure 6**

(a) Deprotonated and (b) protonated forms of histidine with Engh and Huber dictionary values.

**Figure 7**

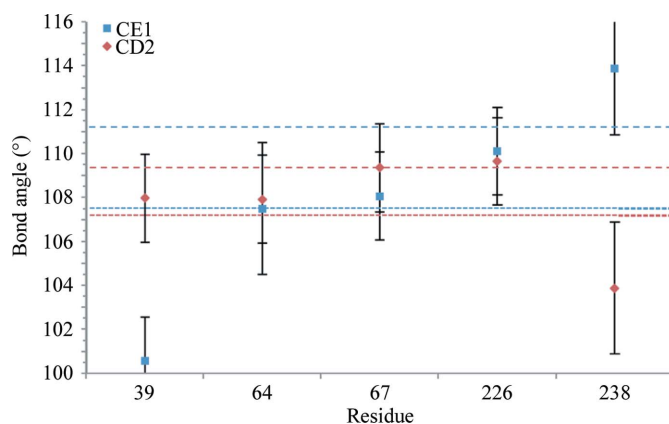
Histidine bond angles with associated standard deviations for cMyBP-C. $\sigma(\text{angle})_{\text{avg}} = 4^\circ$. The upper set of dashed lines indicate the CE1/CD2 bond angles for the deprotonated form of histidine and the lower set of lines indicate the values for the protonated form.

Figs. 7, 8 and 9 show the bond angles of histidine residues in cMyBP-C, subtilisin and trypsin, respectively. In lysozyme there is only a single histidine residue: His15. Its bond angles are 110.53° (1.34) and 107.18° (1.20), resulting in a σ level of 1.86σ .

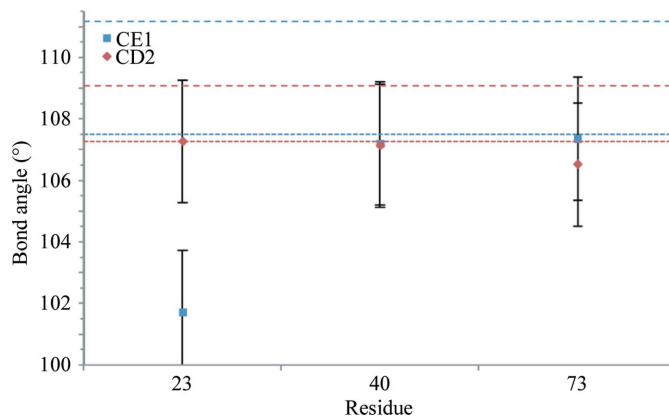
Unfortunately, at the diffraction resolutions of these data sets the precision of the corresponding angles is between 2° and 5° . This is not good enough to determine any of the protonation states definitively. Even at 0.65 \AA for lysozyme the precision of the angles in His15 are around 1.2° , which is still not good enough to definitively infer the corresponding protonation state. However, a number of residues do show bond angles that are in agreement with the protonated values.

6. Discussion

Comparing the four data sets, the resolution has the largest effect on the bond-length standard deviations. Table 3 shows the average standard deviation and B factor for each protein, together with the best protonation-state determination. An increase in resolution of 0.04 \AA (from 1.3 to 1.26 \AA) and an increase in completeness from 88.4 to 100% together result in a twofold reduction in the bond-length standard deviations

**Figure 8**

Histidine bond angles with associated standard deviations for subtilisin Carlsberg. $\sigma(\text{angle})_{\text{avg}} = 2^\circ$

**Figure 9**

Histidine bond angles with associated standard deviations for trypsin. $\sigma(\text{angle})_{\text{avg}} = 2^\circ$

from 0.09 to 0.04 Å in the case of cMyBP-C compared with subtilisin. Further increasing the resolution from 1.26 to 1.20 Å results in another twofold reduction in the standard deviations from 0.04 to 0.02 Å in the case of subtilisin compared with trypsin. It is difficult to compare the estimated standard deviations (e.s.d.s) obtained from different proteins at different resolutions, as other factors could also influence these values, such as the B factors and the location of particular residues with regard to the surface or the core of the protein. However, in general these results still indicate that the diffraction resolution remains the most important factor for determining protonation states from bond-length analysis.

Although four points (one from each protein) do not allow us to fit the 5/2 power law expected from Blow's rearrangement of the Cruickshank DPI, the results depicted in Figs. 10 and 11 and discussed in §6.1 do indicate that both the resolution and the completeness have a significant impact on the bond-length standard deviations. Cruickshank (1999) estimated that the DPI differs from the *SHELX* e.s.d.s by around 15%.

These results show that useful information regarding protonation states can be obtained from X-ray data at resolutions of around 1.2 Å. At a resolution worse than 1.3 Å the X-ray data-to-parameter ratio falls significantly, making it difficult to successfully calculate the full matrix inversion and hence to obtain any useful information, thus confirming the results of Ahmed *et al.* (2007). Minor increments in resolution around the 1.2 Å region result in a significant increase in protein structure precision. Clearly, when collecting data for such protonation-state determinations it is important to achieve the best possible completeness, preferably 100%, whilst considering the possibility of increased X-ray radiation damage.

6.1. Extrapolations

Cruickshank (1999) introduced the diffraction-component precision index (DPI), which provides an estimate of the average (*i.e.* overall) level of precision of a protein crystal

structure based on least-squares refinement. Cruickshank's form of the DPI based on the R_{free} factor is

$$\sigma(x, B_{\text{avg}}) = 1.0(N_i/n_{\text{obs}})^{1/2} C^{-1/3} R_{\text{free}} d_{\text{min}}, \quad (2)$$

where $\sigma(x, B_{\text{avg}})$ is the DPI for an atom with an average B factor (B_{avg}), N_i is the number of fully occupied atoms of type i , C is the completeness of the data, d_{min} is the diffraction resolution and $n_{\text{obs}} = Cn_{\text{int}}$, where n_{int} is the total number of independent intensities obtainable to the resolution limit d_{min} .

In this form, it is difficult to assess how the DPI is related to the resolution or the completeness owing to the dependencies of n_{obs} , C and d_{min} . Blow's rearrangement of the Cruickshank DPI (Blow, 2002) related it to parameters that are readily available to the experimental crystallographer,

$$\sigma(x, B_{\text{avg}}) = 0.18(1+s)^{1/2} V_{\text{M}}^{-1/2} C^{-5/6} R_{\text{free}} d_{\text{min}}^{5/2}. \quad (3)$$

The DPI now explicitly depends on the solvent content $s (= N_{\text{solvent}}/N_{\text{atoms}})$, the Matthews volume V_{M} (Matthews, 1968), the completeness, R_{free} and resolution. Direct comparisons against resolution and completeness can now be made. In this format, the precision of the protein is dependent on the resolution by a factor $d_{\text{min}}^{5/2}$ and on completeness by a factor $C^{-5/6}$.

The Cruickshank DPI provides an *average* estimate of the positional errors of individual atoms having a B factor of B_{avg} within a protein. When studying protonation states, the residues of interest have to have their atomic position precisions adapted according to their B factors *versus* the average B factor for the whole structure. (4) shows a way to correct $\sigma(x)$ for the change in B factor, where B_{avg} is the average B factor for the protein, B_{actual} is the B factor for the residue of interest and $\sigma(x_{\text{actual}})$ is the corrected $\sigma(x)$ value. It can be used to determine an extrapolation of the resolution and completeness required in order to determine specific protonation states to a 3σ significance level,

$$\sigma(x_{\text{actual}}) = \sigma(x)(B_{\text{actual}}/B_{\text{avg}})^{1/2}. \quad (4)$$

Here, we calculate the expected resolution and completeness extrapolations for the four proteins studied and give the

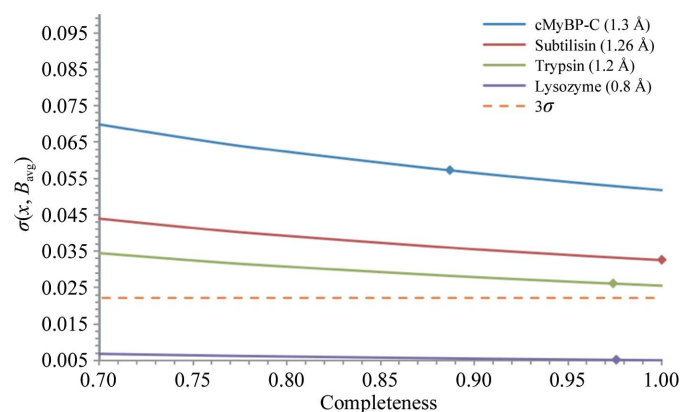


Figure 10
Completeness extrapolation for the four proteins studied using the Cruickshank DPI. The markers indicate the experimental completeness for each data set.

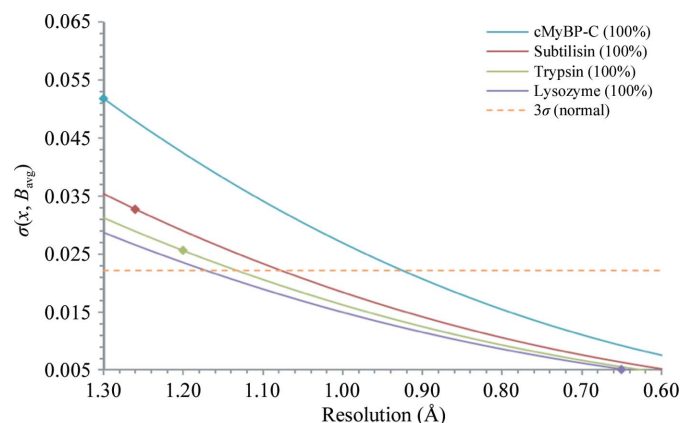


Figure 11
Resolution extrapolation for the four proteins studied based on 100% completeness using the Cruickshank DPI. The markers indicate the experimental diffraction resolution limits for each data set.

resolutions required in order to positively infer the protonation states of key catalytic residues to a 3σ significance level.

6.1.1. Completeness. Fig. 10 shows the completeness extrapolations for the four proteins studied. As the standard deviations depend on the completeness by a factor of $C^{-5/6}$, increasing the completeness alone does not allow sufficient improvement of the standard deviations to allow the determination of protonation states to a 3σ level.

6.1.2. Resolution. Fig. 11 shows the resolution extrapolations for the four proteins studied. The standard deviations are calculated according to Blow (2002) and are dependent on the resolution by a factor of $d_{\min}^{5/2}$. This has a much greater effect on the standard deviations compared with the completeness. Minor changes in resolution result in larger shifts in the estimated standard deviations. Therefore, even increasing the resolution by 0.1 Å can improve the standard deviations of the bond lengths significantly.

7. Mechanistic implications: key catalytic residues in serine proteases

The precise mechanism of serine proteases has been debated for over 50 years. Definitive proof that the catalytic triad histidine was the catalytic base was obtained when Kossiakoff and Spencer used neutron protein crystallography to identify that His57 was protonated (as deuterium) in trypsin

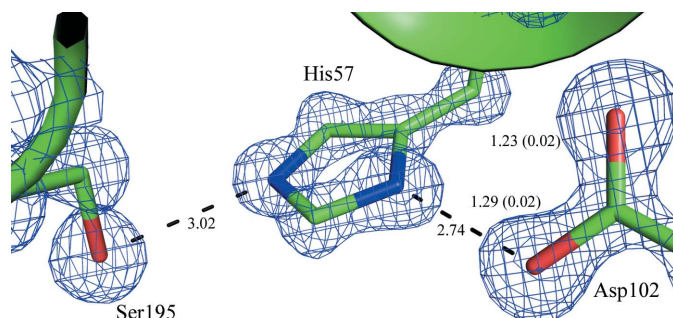


Figure 12
Active site of trypsin at 1.2 Å resolution with associated bond lengths. $2F_o - F_c$ electron-density map shown contoured at 2 r.m.s. $\sigma_{l(\text{avg})} = 0.02$ Å

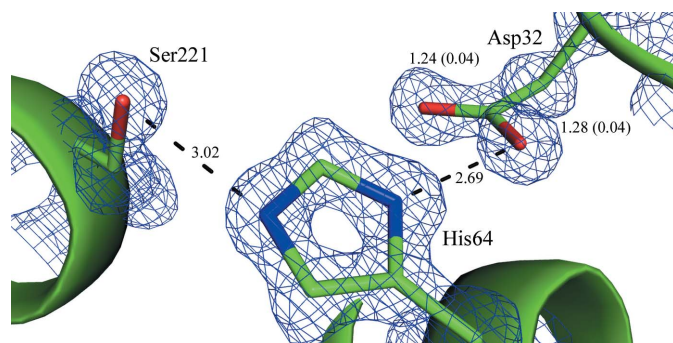


Figure 13
Active site of subtilisin Carlsberg at 1.26 Å resolution with associated bond lengths. $2F_o - F_c$ electron-density map shown contoured at 2 r.m.s. $\sigma_{l(\text{avg})} = 0.04$ Å

(Kossiakoff & Spencer, 1980). However, both NMR and crystallographic studies have challenged their results. NMR studies by Frey *et al.* (1994) showed that the proton may in fact be involved in a low-barrier hydrogen bond (LBHB) and as such is positioned between the two key catalytic residues rather than being fixed on the histidine. In more recent crystallographic studies, Kuhn *et al.* (1998) found the key proton to be positioned midway between His64 and Asp32 when studying subtilisin at 0.78 Å. They suggested that in the neutron experiment of Kossiakoff & Spencer (1980) the heavier D atom was fixed in a single position, whereas the lighter H atom could tunnel between two positions.

Analysis of Asp102 (trypsin) and Asp32 (subtilisin) shows bond lengths of around 1.25 Å for C—O^{δ1} and 1.29 Å for C—O^{δ2}, respectively. These results are consistent with the NMR and previous X-ray evidence that the key catalytic proton may be positioned between His and Asp, giving further evidence for the proposed LBHB. Alternatively, as mentioned above, it is possible that a proportion of the crystal is protonated and a proportion is not for these two catalytic Asps. Figs. 12 and 13 show the active sites of trypsin and subtilisin, respectively, with associated bond lengths.

8. Instrumentation and data measurement

With advances in crystal-growth and synchrotron technology, it is becoming increasingly possible to collect X-ray data to atomic resolution. With this and advancing technologies for neutron crystallography, it is becoming more feasible to collect **both** neutron and ultrahigh-resolution X-ray data. This combination provides key information in the determination of protonation states and also a cross-check between the two types of data.

It is also becoming commonplace within neutron protein crystallography to collect neutron and X-ray data from the same crystal, as neutrons cause no apparent radiation damage. This guarantees the same position of the atoms for the data sets, which is especially useful for joint refinement using neutron and X-ray data simultaneously (Afonine *et al.*, 2010).

These various advances make it increasingly possible to determine protonation states in proteins. A recent applicable example is that of HIV-1 protease (Adachi *et al.*, 2009), the protonation-state results of which have led to the development of novel inhibitors to tackle AIDS. Other examples include endothiapepsin (Coates *et al.*, 2008), type III anti-freeze protein (Howard *et al.*, 2011) and aldose reductase (Blakeley *et al.*, 2008).

9. Conclusions

These four studies confirm the previous methodological conclusions of Ahmed *et al.* (2007) and Fisher *et al.* (2008) and explore two new cases in detail, thus providing greater diffraction-resolution and completeness ranges (up to 0.65 Å and/or 100% completeness). Bond-length e.s.d.s as small as 0.008 Å provided an exceptional opportunity for bond-length analyses, which we have presented alongside the other cases.

The protonation states in histidine residues have also been considered; however, owing to the smaller differences between the protonated and deprotonated forms, even on bond angle changes it is much more difficult to infer the protonation states of these residues. Not even the 0.65 Å resolution lysozyme structure provides the necessary accuracy to positively infer the protonation states of histidine.

SJF is supported by FWF grant P22862 from the Austrian Science Fund. The EMBL Hamburg is thanked for general support (MC) and likewise the University of Manchester (JRH), the Institut Laue–Langevin (SJF) and the ESRF (SM). We are very grateful to Dr Zbyszek Dauter, Argonne National Laboratory for providing detailed bond distance and angle e.s.d.s relating to the triclinic lysozyme study.

References

- Adachi, M. *et al.* (2009). *Proc. Natl Acad. Sci. USA*, **106**, 4641–4646.
- Adams, P. D. *et al.* (2010). *Acta Cryst. D* **66**, 213–221.
- Afonine, P. V., Mustyakimov, M., Grosse-Kunstleve, R. W., Moriarty, N. W., Langan, P. & Adams, P. D. (2010). *Acta Cryst. D* **66**, 1153–1163.
- Ahmed, H. U., Blakeley, M. P., Cianci, M., Cruickshank, D. W. J., Hubbard, J. A. & Helliwell, J. R. (2007). *Acta Cryst. D* **63**, 906–922.
- Berisio, R., Lamzin, V. S., Sica, F., Wilson, K. S., Zagari, A. & Mazzarella, L. (1999). *J. Mol. Biol.* **292**, 845–854.
- Blakeley, M. P. (2009). *Crystallogr. Rev.* **15**, 157–218.
- Blakeley, M. P., Ruiz, F., Cachau, R., Hazemann, I., Meilleur, F., Mitschler, A., Ginell, S., Afonine, P., Ventura, O. N., Cousido-Siah, A., Haertlein, M., Joachimiak, A., Myles, D. & Podjarny, A. (2008). *Proc. Natl Acad. Sci. USA*, **105**, 1844–1848.
- Blow, D. M. (2002). *Acta Cryst. D* **58**, 792–797.
- Bon, C., Lehmann, M. S. & Wilkinson, C. (1999). *Acta Cryst. D* **55**, 978–987.
- Cianci, M. *et al.* (2005). *J. Synchrotron Rad.* **12**, 455–466.
- Coates, L., Tuan, H.-F., Tomanicek, S., Kovalevsky, A., Mustyakimov, M., Erskine, P. & Cooper, J. (2008). *J. Am. Chem. Soc.* **130**, 7235–7237.
- Cruickshank, D. W. J. (1999). *Acta Cryst. D* **55**, 583–601.
- Deacon, A., Gleichmann, T., Kalb (Gilboa), A. J., Price, H., Raftery, J., Bradbrook, G., Yariv, J. & Helliwell, J. R. (1997). *J. Chem. Soc. Faraday Trans.* **93**, 4305–4312.
- Diederichs, K. (2006). *Acta Cryst. D* **62**, 96–101.
- Engh, R. A. & Huber, R. (1991). *Acta Cryst. A* **47**, 392–400.
- Fisher, S. J., Helliwell, J. R., Khurshid, S., Govada, L., Redwood, C., Squire, J. M. & Chayen, N. E. (2008). *Acta Cryst. D* **64**, 658–664.
- Fisher, S. J., Wilkinson, J., Henschman, R. & Helliwell, J. R. (2009). *Crystallogr. Rev.* **14**, 231–259.
- Flot, D., Mairs, T., Giraud, T., Guijarro, M., Lesourd, M., Rey, V., van Brussel, D., Morawe, C., Borel, C., Hignette, O., Chavanne, J., Nurizzo, D., McSweeney, S. & Mitchell, E. (2010). *J. Synchrotron Rad.* **17**, 107–118.
- Frey, P. A., Whitt, S. A. & Tobin, J. B. (1994). *Science*, **264**, 1927–1930.
- Govada, L., Carpenter, L., da Fonseca, P. C., Helliwell, J. R., Rizkallah, P., Flashman, E., Chayen, N. E., Redwood, C. & Squire, J. M. (2008). *J. Mol. Biol.* **378**, 387–397.
- Helliwell, J. R. (1988). *J. Cryst. Growth*, **90**, 259–272.
- Howard, E. (2004). *Proteins*, **55**, 792–804.
- Howard, E. I., Blakeley, M. P., Haertlein, M., Petit-Haertlein, I., Mitschler, A., Fisher, S. J., Cousido-Siah, A., Salvay, A. G., Popov, A., Muller-Dieckmann, C., Petrova, T. & Podjarny, A. (2011). *J. Mol. Recognit.* **24**, 724–732.
- Kleywegt, G. (2003). *MOLEMAN2*. <http://xray.bmc.uu.se/usf/>.
- Kossiakoff, A. A. & Spencer, S. A. (1980). *Nature (London)*, **288**, 414–416.
- Kuhn, P., Knapp, M., Soltis, S. M., Ganshaw, G., Thoene, M. & Bott, R. (1998). *Biochemistry*, **37**, 13446–13452.
- Mason, S. A., Bentley, G. A. & McIntyre, G. J. (1984). *Neutrons in Biology*, edited by B. P. Schoenborn, pp. 323–334. New York: Plenum Press.
- Matthews, B. W. (1968). *J. Mol. Biol.* **33**, 491–497.
- Prangé, T., Schiltz, M., Pernot, L., Colloc'h, N., Longhi, S., Bourguet, W. & Fourme, R. (1998). *Proteins*, **30**, 61–73.
- Ramanadham, M., Jakkal, V. S. & Chidambaram, R. (1993). *FEBS Lett.* **323**, 203–206.
- Ravelli, R. B. G., Leiros, H.-K. S., Pan, B., Caffrey, M. & McSweeney, S. (2003). *Structure*, **11**, 217–224.
- Rosenbaum, G. *et al.* (2006). *J. Synchrotron Rad.* **13**, 30–45.
- Sheldrick, G. M. (2008). *Acta Cryst. A* **64**, 112–122.
- Wang, J., Dauter, M., Alkire, R., Joachimiak, A. & Dauter, Z. (2007). *Acta Cryst. D* **63**, 1254–1268.
- Wang, J., Kamtekar, S., Berman, A. J. & Steitz, T. A. (2005). *Acta Cryst. D* **61**, 67–74.
- Winn, M. D. *et al.* (2011). *Acta Cryst. D* **67**, 235–242.
- Wlodawer, A., Li, M., Gustchina, A., Dauter, Z., Uchida, K., Oyama, H., Goldfarb, N. E., Dunn, B. M. A. & Oda, K. (2001). *Biochemistry*, **40**, 15602–15611.