Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

His-tag impact on structure

Mike Carson,* David H. Johnson, Heather McDonald, Christie Brouillette and Lawrence J. DeLucas

Center for Biophysical Sciences and Engineering, University of Alabama at Birmingham, 251 CBSE, 1025 18th Street South, Birmingham, AL 35294-4400, USA

Correspondence e-mail: carson@uab.edu

Received 28 September 2006
Accepted 1 December 2006

Crystallographers are increasingly determining structures of protein constructs that include His tags. Many have taken for granted that these tags have little effect on the native structure. This paper surveys and compares crystal structures with and without His tags. It is observed that actual refined tag residues fitted into density occur in less that 10% of the tagged sequences. However, higher resolution crystals are observed when this occurs. It is shown that these purification tags generally have no significant effect on the structure of the native protein. Resolution and *R* factors are not affected, but the overall *B* factors are slightly higher. Additional annotation in the PDB format to make tag definition explicit is suggested.

1. Introduction

His tags (Smith *et al.*, 1988) have gained great popularity over the last decade as a purification tool for recombinant proteins. Cloning vectors generally introduce six consecutive histidines and an optional protease-cleavage site or linker to the N- or C-terminus of the protein of interest. These His tags facilitate selective binding of the expressed protein to a nickel-affinity column. The tag may then optionally be removed by a protease, requiring another purification step. The often tacit assumption is that these tags have no effect on the structure and function of the protein (Chant *et al.*, 2005).

There may be drawbacks to these tags. They may alter the solubility or increase the aggregation of the purified protein. Many crystallographers believe these tags hinder crystallization and thus will cleave the tag before screening for crystallization conditions (Derewenda, 2004; Waugh, 2005). However, crystallographers are increasingly determining structures of the entire cloning construct. We were concerned that these tags might have an effect on the native structure, particularly at the terminus where the tag is attached (McDonald *et al.*, 2007). Here, we survey crystal structures with and without His tags. We show that His tags are not necessarily hazardous to your structure and may actually be helpful.

2. Methods and results

All results are based on protein crystal structures deposited in the Protein Data Bank (PDB; Berman *et al.*, 2000). The structural information is generally entered by scientists using the PDB's *ADIT* tool, producing the well known PDB-formatted file of keywords and data. For example, the 'ATOM' records contain the coordinates and temperature factors of protein residues fitted into density. Various 'REMARK'

Table 1
Number of PDB protein entries with crystallization records (all) and number of His tags (tags) in the SEQRES records by year.

The percentage is given as 100 × tags/all.

Year	Tags	All	Percentage
1995	0	16	0.00
1996	2	206	0.97
1997	5	604	0.83
1998	18	804	2.24
1999	15	1053	1.42
2000	44	1426	3.09
2001	79	1507	5.24
2002	103	1602	6.43
2003	268	2160	12.41
2004	362	2324	15.58
2005	246	1346	18.28

records contain a wealth of refinement information such as resolution, *R* factors and solvent content. The 'SEQRES' records contain the sequence. The *ADIT* tool requests one-letter-code sequences for each macromolecule 'including cloning artifacts, purification tags, modifications and any residues not modeled due to missing density'.

2.1. Defining His tags and extracting data

The PDB, as of late 2005, had recently been mined for crystallization records (M. Carson & L. J. DeLucas, unpublished work) much in the manner of the recent publication by Peat *et al.* (2005). The PDB format 2.0, which includes crystallization records, was formally introduced in 1996 after beta testing. Protein structures were selected only if they contained the 'REMARK 280' crystallization records, these records could be parsed and there was only one unique protein sequence among the 'SEQRES' records. Custom Python scripts were developed to extract all information. A total of 21 547 PDB header files were examined and a total of 13 048 structures were successfully parsed.

Six consecutive histidine (H) residues found in the sequence define a His tag. Examining the 13 048 structures successfully parsed, 1142 contain His tags. Of the tags, 680 were at the N-terminus, 457 at the C-terminus and five were engineered in internal loops. Table 1 gives the number of structures per year from 1995, when the 'REMARK 280' records first appeared, through to 2005. No His tags are seen until 1996, but their percentage of the PDB protein structures has steadily increased to nearly 20% per year today.

In addition to parsing sequence and refinement information in the PDB header, the 'ATOM' records for each protein chain were split into separate files and analyzed for sequence, secondary structure and average temperature factors using utilities of the *ribbons* program (Carson, 1997). A key point here is that the sequence of the structure may not contain coordinates for all atoms. Any residue without both a C^{α} and carbonyl O atom is considered to be missing.

Only 65 of the 1142 His-tag structures had main-chain coordinates defined for four or more consecutive histidines. This is less than 6% of all the structures with crystallized tags.

Table 2
PDB entries with well defined His-tag structures are sorted on #res, the number of residues.

Entries with the same #res have the same or a very similar sequence.

N-terminal		C-terminal	
Code	#res	Code	#res
1mhx	65	1u6t	121
1hz5	72	1v30	124
1jml	72	1zhv	134
1k51	72	2axw	134
1k52	72	1sxr	183
1k53	72	1q3i	214
1uj8	77	1z97	266
1m16	146	1rwi	270
1rlh	173	2bv9	290
1oi4	193	1gkl	297
1w3o	216	1wb4	297
1qwz	235	1wb5	297
1ici	256	1ii0	589
1t9h	307	1ii9	589
1vlc	366		

Only 29 of these, 15 N-terminal and 14 C-terminal, had five or more consecutive histidines with complete side-chain coordinates and average *B* factors for the tag atoms under the arbitrary cutoff of 40 Å^2. This small group were taken as the only structures with well defined His tags and are listed in Table 2.

2.2. Statistical analysis

The data were divided into two groups: tagged for the 1142 with His tags and all others (11 906). Statistics were gathered and histograms plotted for the number of residues, *R* factor, $R_{\rm free}$, resolution, $V_{\rm M}$, percentage solvent and average temperature factor of the protein chain. The results are shown in Table 3. Histograms of the distributions are shown in Fig. 1.

The number of residues is shifted higher for the tagged structures, which is expected as the extra residues of the tag have been added by the cloning vector. No differences are seen in the *R* values or resolution. The histograms show a very slight shift towards higher *B* factors and solvent content for the tagged structures. Analysis of the pairs of distributions through *t*-tests shows the degree of similarity of the distributions. The smaller the *t* value plotted in Fig. 1, the more similar the pair. The $R_{\rm free}$ values are the most similar of the pairs and the *B* factors are the most dissimilar. While all the differences (except $R_{\rm free}$) are statistically significantly different, the differences are not scientifically meaningful. The very low *p* values are an artifact of the large numbers of observations in the distributions. The difference in the mean *B* factors is slightly over 4 Å^2, while the standard deviation of each distribution is about 15 Å^2. It should again be noted that less than 6% of the structures have any atomic coordinates fitted to the tag residues.

The analysis was repeated comparing the 680 N-terminal tags with the 457 C-terminal tags, as well as comparing the 65 structures with partial tag coordinates to those without. No differences were seen (data not shown). The 29 structures with

well defined tags have significantly better refinement statistics, with resolution and R factors shifted by roughly one standard deviation; however, these tend to be significantly smaller proteins. This is addressed further in §3.

2.3. Structural comparisons

A list of sequences in the PDB with both a native and a His-tagged form was generated. All sequences with tags were compared with all sequences without tags, requiring an exact match except for a reasonable extension at one end. This generated 216 pairs of PDB codes, of which 52 were unique sequence pairs. For each set of PDB codes with the same sequence, the PDB file with the lowest $R_{\rm free}$ value was selected. This created a list of 52 pairs of PDB files, one having a native sequence and one having a His tag, presented in Table 4. By inspection of the 'AUTHOR' records, 33 of the pairs appeared to be solved in the same laboratory. The differences in R values, resolution and B factors were analyzed. No differences were seen, only a roughly normal distribution about zero as shown in Fig. 2.

The tagged structure was superimposed on the untagged structure using common C^{α} atoms. Plots were generated illustrating the regions of missing atoms, pairs of C^{α} more than 1 Å apart, temperature factors and secondary structure. Interactive examination of all pairs as ribbon models revealed little out of the ordinary. A typical result is shown in Fig. 3, which is the only pair where the His-tag residues are well

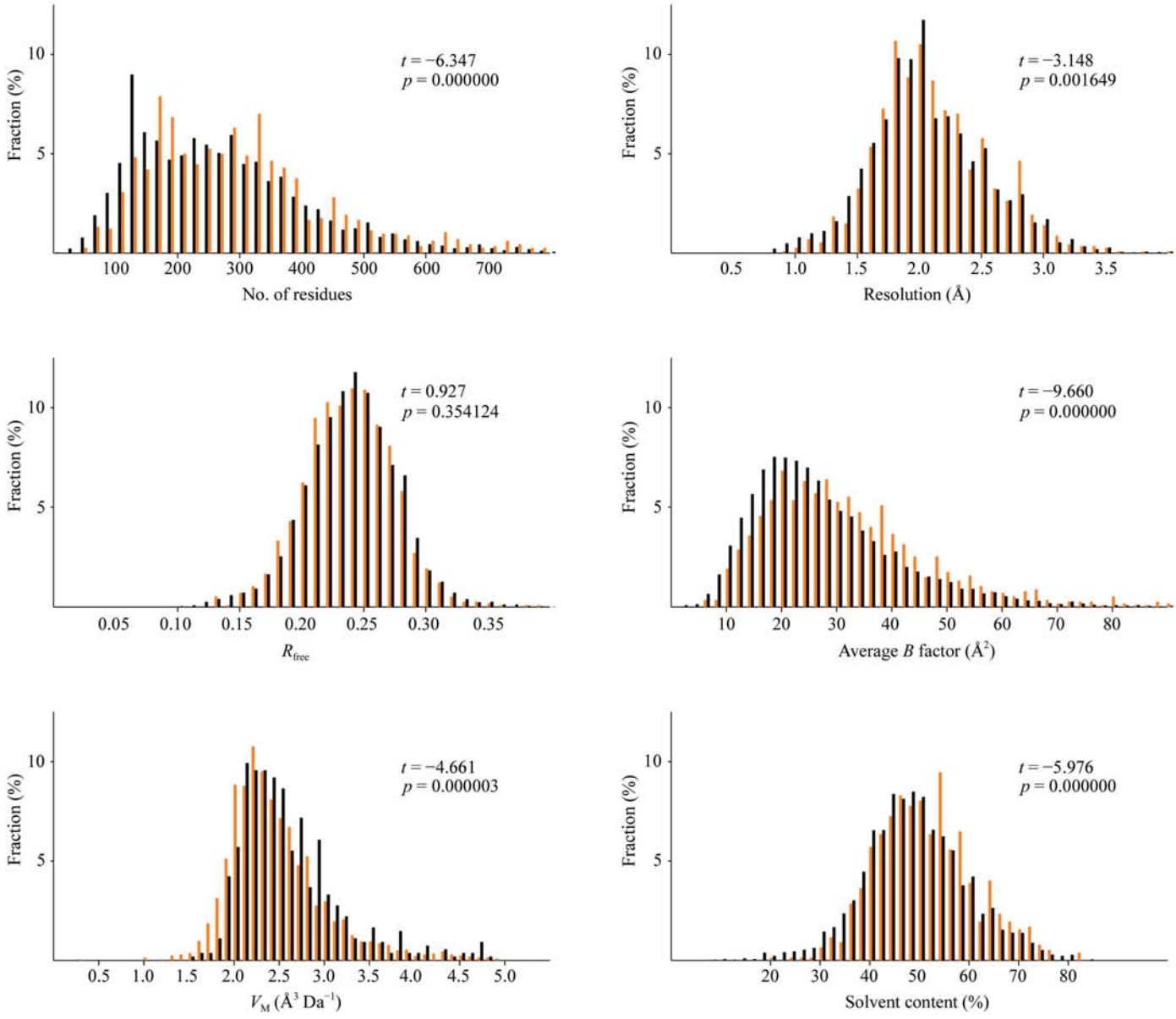

Figure 1
Histograms of refinement values for tagged and untagged structures. The values for the untagged structures are in black and those for tagged structures in orange. The fraction in each bin is plotted against the labeled value. The data corresponds to Table 1. The t-test statistic comparing the distributions is shown.

Table 3
Average refinement statistics for structures with (tagged) and without (no tag) His tags and for those with well defined His-tag atoms (H6 fit).

The average, standard deviation, minimum and maximum values are given for each distribution of the listed variable. The number of observations depends on the successful parsing of the refinement statistic. The average *B* factor was calculated over the 'ATOM' records of the protein.

Variable	Average	Std dev.	Min.	Max.	No. of observations
No. of residues					
No tag	325.7	165.0	6	1520	11904
Tagged	349.1	154.4	65	1053	1142
H6 fit	213.8	139.6	65	589	29
R value					
No tag	0.200	0.032	0.05	0.47	11824
Tagged	0.202	0.031	0.10	0.33	1142
H6 fit	0.184	0.034	0.11	0.25	29
R_{free}					
No tag	0.245	0.038	0.06	0.48	11104
Tagged	0.244	0.036	0.13	0.40	1139
H6 fit	0.214	0.038	0.13	0.29	29
Resolution (Å)					
No tag	2.127	0.488	0.54	9.50	11906
Tagged	2.166	0.448	0.95	4.00	1142
H6 fit	1.794	0.361	1.05	2.60	29
Percentage solvent					
No tag	50.41	10.55	9.80	91.00	6239
Tagged	52.68	9.83	21.10	82.00	772
H6 fit	49.75	10.57	33.79	71.20	16
V_M (Å^3 Da^{-1})					
No tag	2.544	0.574	0.32	5.00	6044
Tagged	2.672	0.598	0.97	4.98	761
H6 fit	2.579	0.629	1.86	4.31	16
Average *B* factor (Å^2)					
No tag	28.53	14.75	0.49	309.2	11905
Tagged	32.97	15.52	7.03	132.3	1142
H6 fit	22.42	9.54	10.17	45.38	29

resolved. There are slight differences at the first few residues of the N-terminus, in a loop in the middle of the structure and at the last few residues of the C-terminus before the ordered His tag extends along and then away from the native protein.

The 29 structures with well defined tags were all examined with the molecular-graphics program *Coot* (Emsley & Cowtan, 2004), which has excellent facilities to view crystal packing. Almost all of the resolved tags are involved in packing contacts. Fig. 5 shows a typical result. About a third of the tags nestle against the native structure; the remainder extend out into the solvent region.

Table 4
Pairs of PDB files for direct untagged/tagged comparisons.

The first PDB file listed in each pair has the native sequence, while the second contains a His tag in the SEQRES records.

1ah6/1a4h	1am1/1bgq	1amw/1bgq	1cv2/1mj5
1di9/1a9u	1di9/1oz1	1exx/3lbd	1f91/1ek4
1frw/1e5k	1hd2/1oc3	1hkn/1rg8	1hzp/1m1m
1jt6/1rkw	1jv4/1znd	1k5h/1q0q	1ljt/1gd0
1m7j/1rk6	1mo2/1kez	1mq7/1six	1mq7/1sjn
1o7q/1g93	1omo/1vll	1oz4/1r7r	1p4a/1o57
1pvf/1hzt	1pvf/1x84	1q54/1hx3	1rxf/1w28
1rzm/1vr6	1s7n/1s7f	1t6k/1u1w	1tb7/1y2b
1thz/1m9n	1tj2/1tiw	1u2k/1u2j	1up3/1uoz
1uwf/1tr7	1wa3/1vlw	1x7g/1w4z	1xf0/1s1p
1yrd/1p2y	1z0i/1yzt	1z6y/1zj6	1z9u/1s7f
1zly/1meo	1zw5/1yv5	2axm/1rg8	2az3/2az1
2b8q/2b8p	2b8t/1xmr	2bvd/2bv9	2mat/4mat

2.4. Tag sequences and disorder

There is no automatic way to discern the 'native' sequence from a PDB file and thus to determine the sequence of the tag. The 1142 sequences identified with His tags from the 'SEQRES' records were subjected to *BLAST* (Altschul *et al.*, 1990) searches against all nonredundant sequences. A tag sequence could be automatically extracted for most sequences by comparing the top two *BLAST* hits. Questionable results were flagged and corrected by inspection.

Additional filters were performed to select 816 sequences with very high confidence that the actual tag sequence was correctly identified. Many crystallographers use a residue-numbering convention giving cloning artifacts at the N-terminal negative residue numbers, with the first native residue numbered as '1'. Any tag that positioned the first native residue as number 1 was selected, as well as any tag that made the first native residue a 'MET'. Any tag that was equal to one of these previously selected tags was allowed. For C-terminal tags, there is no such numbering convention. Tags were allowed if four or more instances occurred. The tags observed

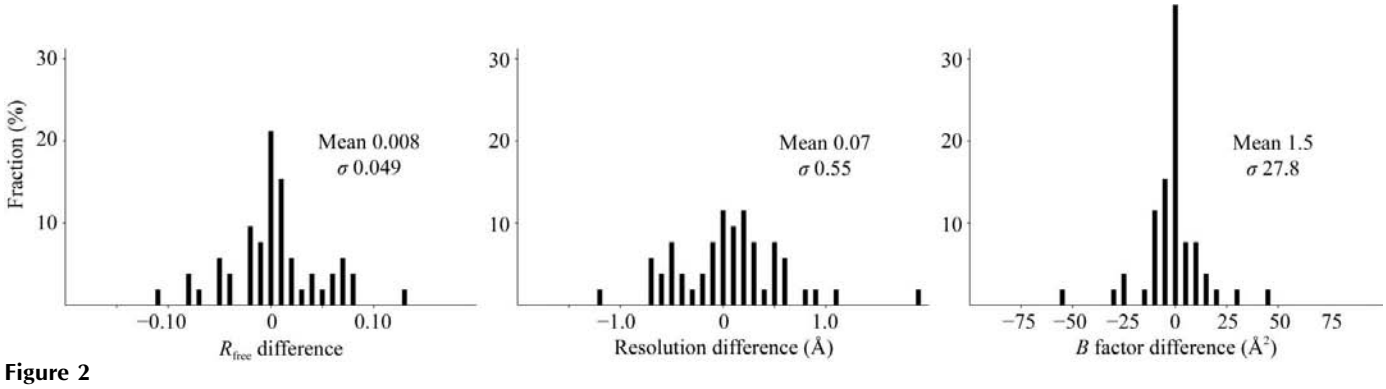

Figure 2
Differences in selected refinement statistics over 52 matched pairs of structures. Histograms give the values of the untagged PDB file minus that of the corresponding tagged PDB file. Thus, if the untagged structures had better R_{free} values, the histogram would be shifted to the left. The mean and standard deviation of the distributions are shown above.

Table 5
Most common high-confidence tag sequences for the N- and C-termini, sorted by count.

Count gives the number of times the sequence is observed; length is the length of the tag.

Count	Length	N-tag sequence	Count	Length	C-tag sequence
78	12	MGSDKIHHHHHH	353	6	HHHHHH
63	20	MGSSHHHHHHSSGLVPRGSH	17	8	LEHHHHHH
60	7	MHHHHHH	8	7	SHHHHHH
38	12	MRGSHHHHHHGS	6	8	VDHHHHHH
37	6	HHHHHH	5	7	EHHHHHH
23	19	MGSSHHHHHHSSGLVPRGS	4	8	HHHHHHHH
16	8	MAHHHHHH	4	7	HHHHHHH
11	25	MGSSHHHHHHDYDIPTTENLYFQGH			
9	14	MGGSHHHHHHGMAS			
6	22	MGSSHHHHHHSSGRENLYFQGH			
5	21	SYYHHHHHHLESTSLYKKAGL			
5	13	MRGSPHHHHHHGS			
5	9	GSSHHHHHH			
3	17	MRGSHHHHHHGLVPRGS			
3	18	MRGSHHHHHHGIPLPGRA			
3	20	MGSSHHHHHHSSENLYFQGH			
3	10	HHHHHHHHHH			

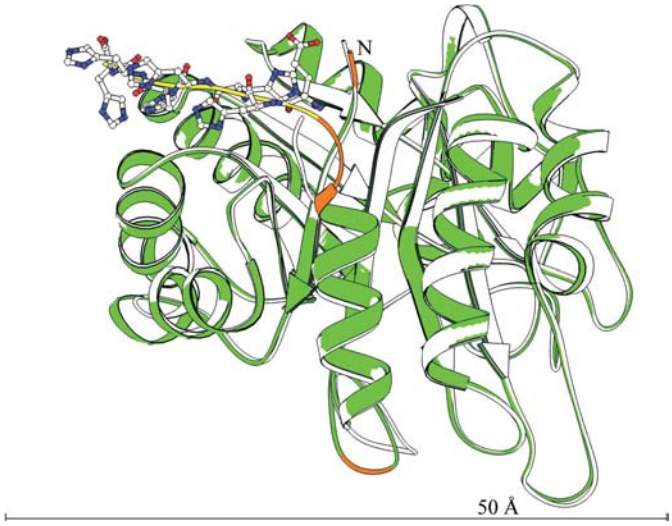

Figure 3
Comparison of an untagged and tagged structure with a resolved tag. The native structure (2bvd) is shown as a white ribbon. The superimposed tagged structure (2bv9) is colored green, except orange where the distance between corresponding C^{α} atoms exceeds 1 Å and yellow for the EHHHHHH tag at residues 284–290 of the C-terminus. The atomic model is shown for the tag. The first seven residues at the labeled N-terminus are missing.

in this high-confidence set are shown in Table 5. The average length of the 419 N tags was 13.6 residues and the average length of the 397 C tags was 6.2 residues.

Each 'SEQRES' sequence was compared with the observed sequence in the solved structure. The positions of the first and last observed residues relative to the entire sequence were determined. Most of the untagged structures start at the first residue and end at the last residue. However, one or more residues are often unresolved at the ends of crystal structures (Li *et al.*, 1999). The distribution of these unresolved or disordered residues at the N- and C-termini is shown in Fig. 4(*a*). Similar results were seen for the termini of tagged structures opposite the tag, *e.g.* comparing disorder at the C-terminus for all structures with an N-terminal tag. As discussed previously, over 90% of the His tags are disordered. The length of the tag was subtracted and the distribution of the number of disordered termini residues repeated for all tagged structures. The results are shown in Fig. 4(*b*). In this case, a negative value is possible, being equal to the number of residues in the cloning tag actually observed in density.

To monitor the effects of disorder alone, the 11 906 structures without His tags were divided into two groups. Structures were selected where either the first ten or more or the last six or more residues were missing to mimic the structures with His tags. This disordered group contains 2601 structures. Statistics were gathered and histograms plotted for the number of residues, R factor, $R_{\rm free}$, resolution, $V_{\rm M}$, percentage solvent and average temperature factor as in Table 3 and Fig. 1. The values and distributions for the disordered groups were virtually identical to the tagged values. Those structures with ordered termini have only marginally different values from the complete set of untagged proteins (data not shown).

3. Discussion

Derewenda (2004) reported that 90% of crystal structures are based on recombinant methods and nearly 60% use some sort of His tag. Crystallographers are increasingly solving the structures of constructs that include His tags, with nearly 20% of the proteins reported from last year containing these tags, as seen in Table 1. The survey presented here is dependent on the integrity of the data reported. There is anecdotal evidence that some crystallographers are unaware of the *ADIT* requirement to enter the tagged sequence into the database. Some may purify and crystallize with His tags (which are not usually observed in the structure), but report only the native sequence. One wonders how common this practice is. The instructions given to crystallographers regarding the sequence information to be submitted should be emphasized.

Fig. 1 and Table 3 show that His tags have very little effect on refinement statistics on aggregate. This might be expected as the average length of these tagged sequences is about 350 residues, while the average tag length is only ten residues. No His-tag histidine coordinates are reported in over 94% of the structures. This might be expected as Dunker and coworkers (Li *et al.*, 1999) have shown histidine composition to be one of the top four predictors for disorder at both the N-temini and C-termini, but not in interior regions. They do not mention His tags in their paper, but when their work was performed in the late 1990s only 1–2% of structures had tags.

Comparison of tagged with corresponding untagged structures in Figs. 2 and 3 reveal only minor structural differences

of the type that might be observed when comparing two identical sequences solved in different space groups. Determining the exact sequence of the tag is problematic. We recommend that the PDB provide more annotation to this point. The data presented in Table 5 and Figs. 3 and 4 indicate that the presence of a tag has a negligible effect on the structure or lack of structure at the terminus to which the tag is attached.

A small increase in average *B* factors is seen for both His-tagged structures (generally disordered) and the untagged proteins with disordered ends. Yet the disordered tags are shown to have a minimal effect on the observed structure. We observe that corresponding residues in the untagged and His-tagged structures have essentially the same coordinates and we suspect that the atoms have essentially the same mean displacements. We speculate that the observed refined coordinates must serve as a sink to absorb the displacement values modeled by the *B* factors in the refinement protocol. Perhaps a bulk-disorder correction could be incorporated into refinement programs analogous to those for bulk-solvent modeling.

Structures with well defined His tags are rare and generally unremarkable. The secondary structure is helical in two cases and extended in all others, with the residues involved in crystal packing. Only one example exists of a pair of structures where one is the native sequence and the other has a well ordered His tag, 2bvd and 2bv9 (Taylor *et al.*, 2005), as shown in Fig. 3. Taylor and coworkers note that their two structures are virtually identical. The tagged 2bv9 has a slightly better resolution and $R_{\rm free}$ (1.5 Å, 0.157) than 2bvd (1.6 Å, 0.185). The tag packs in the cleft of a symmetry-related molecule. The packing is shown in Fig. 5 and is typical, with polar and aromatic contacts more prevalent than charge–charge interactions.

The most dramatic structural shifts involving proteins with ordered His tags occurs where domain-swapping is involved. The high-resolution crystal structure of the bacterial invasin DraD (Jędrzejczak *et al.*, 2006) was determined from a construct with a C-terminal extension of 13 residues consisting of a linker and a His tag. This long extension forms a swapped strand in the symmetrical dimers and is thought to be a model for the fiber formation involving this adhesin. Baker and coworkers conducted a computer-based protein-design study that generated five of the 29 well defined His-tag structures (Kuhlman *et al.*, 2001). The small protein L, consisting of a single helix packed on a four-stranded sheet, was converted into a domain-swapped dimer in which a turn straightens and the C-terminal strand inserts itself into the sheet of its partner. The sequences and structures have an N-terminal His tag that forms no secondary structure. The mutations in the structures (1jml, 1mhx, 1k51, 1k52, 1k53) are all in the hinge region away from the tag.

There have been several reports of the influence of the type and position of His tags on expression (Doray *et al.*, 2001; Mast *et al.*, 2004; Woestenenk *et al.*, 2004). This is obviously a crucial point in the production of material for crystallization, but

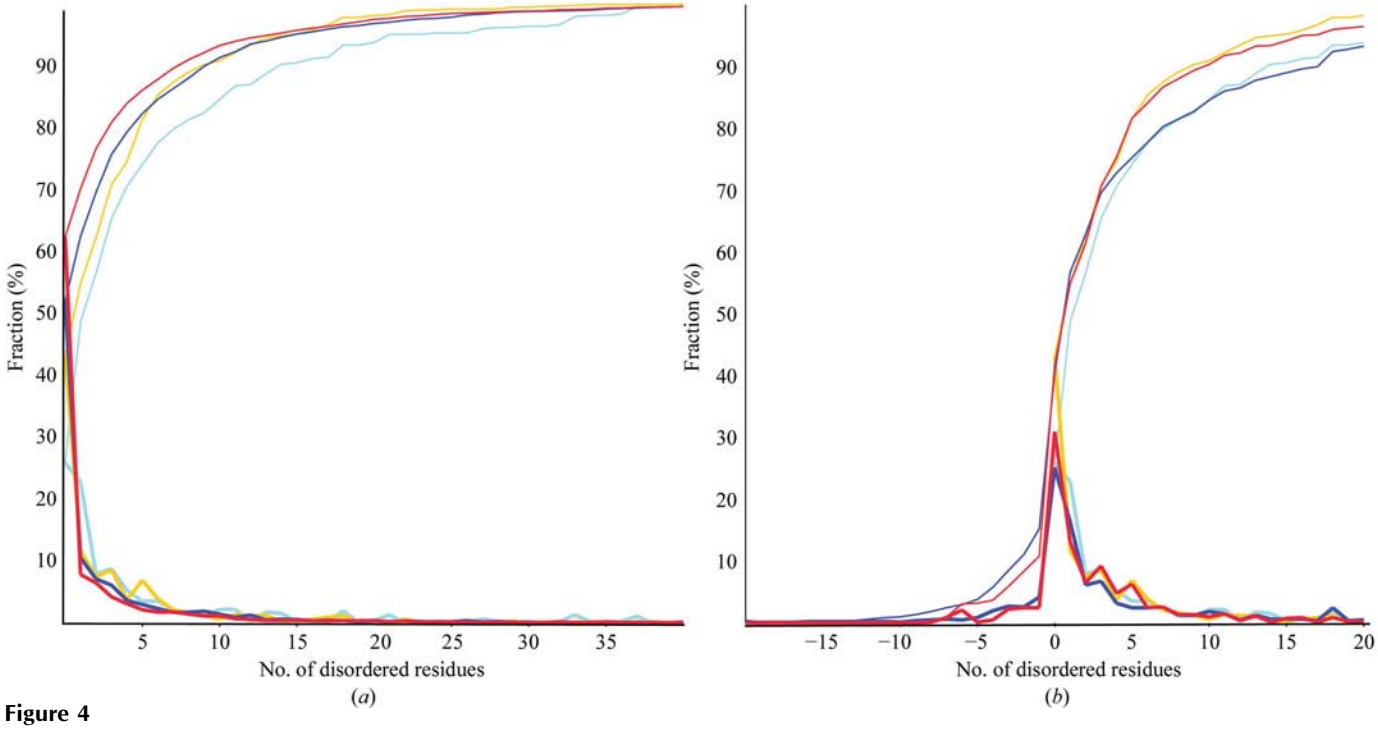

Figure 4
Disorder at the termini of untagged and tagged structures. The fraction with each number of disordered residues is shown by the thicker curves below; the cumulative fraction is given by the thinner curves above. (*a*) All untagged structure data are shown in blue for the N-terminus and in red for the C-terminus. Data for all tagged structures considering only the ends opposite to the tag are shown in cyan for the N-terminus and in yellow for the C-terminus. (*b*) Data for all tagged structures considering only the ends opposite to the tag are shown in cyan for the N-terminus and in yellow for the C-terminus as in (*a*). Data for the tagged ends, after subtracting the length of the tag, are shown in blue for the N-terminus and in red for the C-terminus. Values of less than zero mean that ordered tag residues exist.

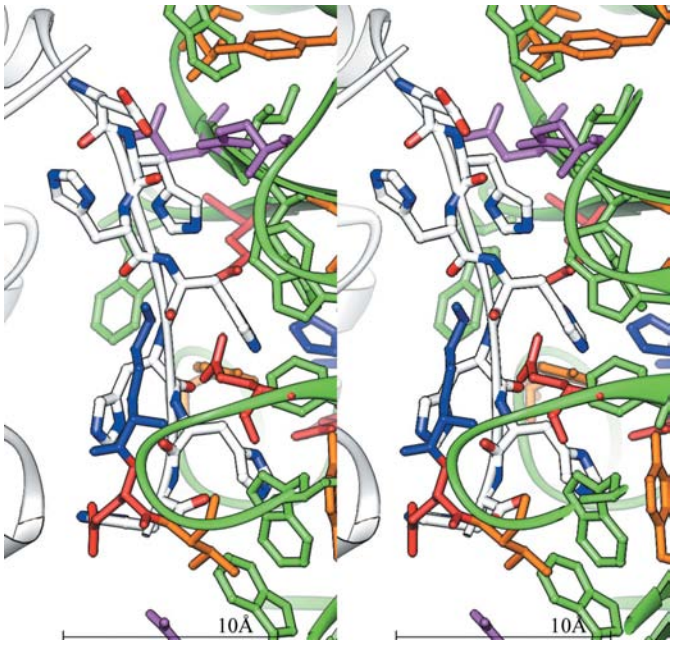

Figure 5
Crystal packing of a typical His tag. The 2bv9 structure of Fig. 3 is shown in white as a ribbon with the tag atoms. A symmetry-related molecule is shown as a green ribbon with residues within 6 Å of the His-tag atoms colored by residue type: hydrophobic, green; acidic, red; basic, blue; alcohol, orange; amide, magenta.

cannot be addressed from structural data alone. Mining of the new PepcDB (protein-expression, purification and crystallization database) assembled from the structural genomics projects at the PDB may prove very useful.

There are only a few reports on the adverse effect of a His tag on structure–function relationships. There is a report where a His tag influences folding and actually changes the disulfide-bonding pattern (Klose *et al.*, 2004). Only one pair of the untagged/tagged structures examined here has multiple disulfide bonds (1up3 and 1uoz) and no structural change was seen. In a cautionary report, Chant *et al.* (2005) used fluorescence spectroscopy to show that the attachment of a tag causes a conformational change at a DNA-binding site.

In some cases, a His tag may be useful or even required for crystallization. Tajika *et al.* (2004) report the structure, 1v30, of a 116-residue protein in which a long C-terminal helix which includes the His tag (LEHHHHHH) protrudes outside the molecule and packs with another molecule about the crystallographic twofold axis. They note that they were unable to obtain crystals of the wild-type sequence under the same crystallization conditions, sodium citrate pH 9.5.

The crystallization conditions observed for the well ordered His-tag structures run the gamut of conditions observed in our database. There are roughly as many using organics as there are using salts as the primary precipitant. The pH values range from 5.5 to 9.5, with a peak around 7.5. We were unable to discern a trend.

The employment of the tags in crystal packing is likely to be what allows the His tags observed in this survey to be well resolved. Although the sample is very small and biased towards very small proteins, the improved refinement statistics are intriguing. Simulations were performed to randomly select subsets of untagged completely ordered structures. The average and standard deviation of the distributions of the number of residues in each subset was similar to the observed distribution of the well ordered His-tag structures. The resolution and R factors were consistently better for the His-tagged structures, but the differences are not statistically significant. Given that His tags have little effect on protein structure, it may be possible to design crystallization tags to enhance structure determination.

Thanks to Craig Smith, Debasish Chattopadhyay, Champy Deivanayagam, Roy Hantgan and Naomi Fineberg for useful discussions and to the referees for helpful comments.

References

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). *J. Mol. Biol.* **215**, 403–410.
Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000). *Nucleic Acids Res.* **28**, 235–242.
Carson, M. (1997). *Methods Enzymol.* **277**, 493–505.
Chant, A., Kraemer-Pecore, C. M., Watkin, R. & Kneale, G. G. (2005). *Protein Expr. Purif.* **41**, 61–67.
Derewenda, Z. S. (2004). *Methods*, **34**, 354–363.
Doray, B., Chen, C. D. & Kemper, B. (2001). *Arch. Biochem. Biophys.* **393**, 143–153.
Emsley, P. & Cowtan, K. (2004). *Acta Cryst.* D**60**, 2126–2132.
Jędrzejczak, R., Dauter, Z., Dauter, M., Piatek, R., Zalewska, B., Mroz, M., Bury, K., Nowicki, B. & Kur, J. (2006). *Acta Cryst.* D**62**, 157–164.
Klose, J., Wendt, N., Kubald, S., Krause, E., Fechner, K., Beyermann, M., Bienert, M., Rudolph, R. & Rothemund, S. (2004). *Protein Sci.* **13**, 2470–2475.
Kuhlman, B., O'Neill, J. W., Kim, D. E., Zhang, K. Y. J. & Baker, D. (2001). *Proc. Natl Acad. Sci. USA*, **98**, 10687–10691.
Li, X., Romero, P., Rani, M., Dunker, A. K. & Obradovic, Z. (1999). *Genome Inform.* **10**, 30–40.
McDonald, H., Pruett, P., Deivanayagam, C., Protassevitch, I., Carson, M., DeLucas, L., Brouillette, W. & Brouillette, C. (2007). Submitted.
Mast, N., Andersson, U., Nakayama, K., Bjorkhem, I. & Pikuleva, I. A. (2004). *Arch. Biochem. Biophys.* **428**, 99–108.
Peat, T. S., Christopher, J. A. & Newman, J. (2005). *Acta Cryst.* D**61**, 1662–1669.
Smith, M. C., Furman, T. C., Ingolia, T. D. & Pidgeon, C. (1988). *J. Biol. Chem.* **263**, 7211–7215.
Tajika, Y., Sakai, N., Tamura, T., Yao, M., Watanabe, N. & Tanaka, I. (2004). *Proteins*, **57**, 862–865.
Taylor, E. J., Goyal, A., Guerreiro, C. I. P. D., Prates, J. A. M., Money, V. A., Ferry, N., Morland, C., Planas, A., Macdonald, J. A., Stick, R. V., Gilbert, H. J., Fontes, C. M. G. A. & Davies, G. J. (2005). *J. Biol. Chem.* **280**, 32761–32767.
Waugh, D. S. (2005). *Trends Biotechnol.* **23**, 316–320.
Woestenenk, E. A., Hammarstrom, M., van den Berg, S., Hard, T. & Berglund, H. (2004). *J. Struct. Funct. Genomics*, **5**, 217–229.