Journal of Biomolecular NMR (2006) 35: 39–51
DOI 10.1007/s10858-006-9003-3

Article

Chemically accurate protein structures: Validation of protein NMR structures by comparison of measured and predicted pK_a values

N. Powers[a,b] & Jan H. Jensen[a,*]
[a]*Department of Chemistry, University of Iowa, Iowa City, USA;* [b]*Department of Physics and Astronomy, University of Iowa, Iowa City, IA, 52242, USA*

Received 25 January 2005; Accepted 8 March 2006

Key words: pK_a, structure validation, ubiquitin

Abstract

A new method is presented for evaluating the quality of protein structures obtained by NMR. This method exploits the dependence between measurable chemical properties of a protein, namely pK_a values of acidic residues, and protein structure. The accurate and fast empirical computational method employed by the PROPKA program (http://www.propka.chem.uiowa.edu) allows the user to test the ability of a given structure to reproduce known pK_a values, which in turn can be used as a criterion for the selection of more accurate structures. We demonstrate the feasibility of this novel idea for a series of proteins for which both NMR and X-ray structures, as well as pK_a values of all ionizable residues, have been determined. For the 17 NMR ensembles used in this study, this criterion is shown effective in the elimination of a large number of NMR structure ensemble members.

Introduction

The efficient and accurate determination of protein structures is a key goal of proteomics research. Currently, the two primary methods used to determine protein structure are X-ray crystallography and NMR spectroscopy.

Both of these techniques have unique experimental limitations. X-ray techniques suffer from the time consuming component of crystal formation (Durbin & Feher, 1996; Kragh-Hansen et al., 1998; Wiencek, 1999), ligand effects (Garavito and Ferguson-Miller, 2001), and structural effects resulting from crystallization itself, such as crystal packing (Nielsen and Vriend, 2001; Georgescu et al., 2002; Nielsen and McCammon, 2003a, b) and dehydration (Nagendra, 1998; Tarek & Tobias, 2002), as well as the loss of heterogeneity of structure resulting from a loss of protein mobility in crystallized form (Montelione et al., 2000; DePristo et al., 2004). These limitations have lead to the great interest in obtaining protein structures free in solution using NMR spectroscopy.

NMR structure generation has its own limitations, the most acknowledged of which is protein size, although multi-dimensional NMR techniques and stronger magnetic fields have led to solved protein structures over 60 kD (Clore and Gronenborn, 1998a, b). Inherently, however, the NMR method is hindered by the limited number of constraints obtainable from an experiment, usually NOE distance constraints (Nabuurs et al., 2003). Generally, the number of constraints is mathematically insufficient to completely determine a unique structure. Of the many structures that satisfy the imposed constraints, a sub-set is selected for publication, usually based on an assessment by some form of error function (Clore and Gronenborn, 1998a, b). Typically, 5–50 structures are deposited in the Protein Data Bank

*To whom correspondence should be addressed.
E-mail: Jan-Jensen@uiowa.edu

for a given protein. *A priori* it is not clear which structures are most accurate and the development of algorithms to determine the best structures is an active area of research. Often, in lieu of a better means of structural assessment, a single, averaged structure is generated to serve as a representative of the ensemble, usually by averaging the atomic positions of the ensemble members and subsequently restraining this structure to satisfy typical bond lengths and angles. Alternatively, the structure with the lowest energy is often selected.

Current refinement and validations schemes exist such as applications of molecular dynamics and simulated annealing (Kuriyan et al., 1989; Brunger et al., 1997; Brunger and Adams 2002; Spronk et al., 2002) and validation schemes based on the identification of anomalous bond lengths and angles, stearic clashes, and other straightforward restraint violations. Many useful approaches have been developed such as MOLPROBITY (Davis et al., 2004), PROCHECK (Laskowski et al., 1996), WHATCHECK (Hooft, 1996), and others (Doreleijers et al., 1999; Linge, 2003) to help identify erroneous structures and eliminate NMR ensemble candidates. Apart from the recognition of gross errors in structure however, the power of these checks to refine structures or recognize correct structure is limited. It is difficult to ascertain when ''refinements'' may lead to larger deviations from a more accurate structure, particularly when the initial structure is ill determined by experimental data, or when the protein actually possesses novel structure that would be unfavored by traditional force fields.

Nonetheless, validation remains critical in the case of NMR structures. Unlike the ability to increase resolution and precision in X-ray experiments, there exists no analogous quantifiable validation approach in terms of precision available for NMR structures. Attempts to create such schemes have been slow to develop, due to the fact that well-ordered parts of the structure have lower errors, and greater numbers of constraints, whereas the more mobile, flexible regions, particularly on the surface of the protein, tend to be more poorly defined, and lack a large number of constraints (Laskowski et al., 1998; Snyder et al., 2005).

Perhaps the best criterion for the quality of a structure is whether it allows the biological function of the protein to be quantitatively predicted. For example, the ''chemical accuracy'' of an enzyme structure could be quantified by using the structure to compute the overall rate and comparing it to the experimental value. However, at present, the computational prediction of a single rate constant of an enzymatic reaction, for example using QM/MM methods (Hall et al., 2000; Zhang et al., 2002) is far from routine and can require years of effort.

We argue that the pK_a values of the ionizable residues of the protein represent an attractive alternative to the rate constant. The pK_a values are a measure of the chemical reactivity of ionizable residues and determine the pH-dependence of the rate for many enzymes. The pK_a value of a given residue depends on the local protein geometry surrounding it (Li et al., 2004). Thus, the measurement of pK_a values can be used to aid in the determination of local structure. The protein pK_a values can be predicted based on the protein structure in a matter of seconds using our PROPKA approach (Li et al., 2005) at http://www.propka.chem.uiowa.edu. Thus, ''chemically inaccurate'' structures can be easily eliminated by measuring the pK_a values of some or all of the ionizable groups in the protein (a relatively easy task compared to measuring the NMR constraints) and comparing them to the PROPKA predictions for each NMR structure. This results in a refinement protocol that improves the quality of protein structures based on knowledge of additional independent experimental measurements, namely, the pK_a values of ionizable residues.

The use of pK_a values to validate structures is attractive in that (a) it is experimentally verifiable and (b) it is correlated with the *chemical* properties of the protein in solution. Such direct comparisons between predictions based on structure and experimental data should generally be more precise than validations of structure inferred from less experimental data. The use of pK_a values can also be used to determine how X-ray structures may differ from those in solution.

Dillet et al. (1998) were among the first to investigate the quality of NMR strucutures using pK_a values in a seminal study of *E. coli* thiroredoxin. Since this study appeared, there has been a significant increase in the number of proteins for which both pK_a values and NMR structures are available. Thus, we are now in a position to investigate this general approach for more proteins.

This paper first presents two ways of using pK_a values for structure validation, both of which are scoring methods devised to eliminate and validate members of NMR ensembles. The accuracy of the resulting selected structures is then evaluated by comparison to available X-ray structures. An alternate approach of averaging to find pK_a values will also be briefly addressed.

Methods

17 different NMR protein ensembles from the protein data bank with known experimental pK_a values were considered. Of these 17 structures, 11 also had similar available structures obtained by X-ray (Table 1). The PROPKA program was used to predict pK_a values for all aspartate and glutamate residues within these structures. These values were then compared to those obtained by experiment. The corresponding NMR ensembles were then superimposed on their respective X-ray structures to determine correlations between the differences in NMR and X-ray structures and the error in predicted pK_a values.

pK_a prediction

The PROPKA program (Li et al., 2005) (http://www.propka.chem.uiowa.edu) was used to predict the pK_a values for all acidic residues for all the structures used in this study (Table 1). PROPKA utilizes a very fast empirical method to predict pK_a values and is successful at predicting unusual pK_a values. The program uses three factors to determine pK_a perturbations: desolvation, hydrogen bonding, and charge-charge interactions. The program's predictions are particularly accurate in the case of Asp and Glu residues, thus the choice of these residues for this study. These residues are also relatively common, important for intraprotein, protein-solvent and protein-ligand interactions

Table 1. Summary of the NMR and X-ray structures used in this study.

Protein	No. Xpt. Res. In structure	NMR	⟨NMR⟩	X-ray	X-ray Resolution (Å)
Hirudin[1]	6	1HIC[2]			
Insulin[3]	5	1MHI[4]			
Epidermal growth factor[5]	6	1EGF[6]			
Turkey ovomucoid third domain[7]	5	1TUR[8]		1PPF[9]	1.80
Bull seminal Inhibitor IIA[10]	4	1BUS[11]	2BUS[11]		
Cardiotoxin AV[12]	3	1CVO[13]		1KXI[14]	2.19
Calbindin D_{9K}[15]	8	2BCB[16], 1CLB[17]	2BCA[16]	4ICB[18]	1.60
Ubiquitin[19]	11	1D3Z[20]		1UBQ[21]	1.80
Subunit c of H^+-transporting F_1F_0ATP synthase[22]	5	1A91[23]			
Cryptogein[24]	3	1BEG[25]		1BEO[26]	2.20
Thioredoxin (oxidized)[27]	17	1TRU[28]	1TRS[28]	1ERU[29]	2.10
Thioredoxin (reduced)[27]	17	1TRV[28]	1TRW[28]	1ERT[29]	1.70
Barnase[30]	12	1BNR[31]		1A2P[32]	1.50
Bovine pancreatic Ribonuclease A[33]	10	2AAS[34]		1RNZ[35]	1.90
Lysozyme (Hen)[36]	9	1E8L[37]		1LYS[38]	1.72
alpha-sarcin[39]	17	1DE3[40]			
Ribonuclease HI[41]	19	1RCH[42]		2RN2[43]	1.48

References in the first column describe the pK_a measurements for each protein, and the second column lists the number of pK_a values determined. The full references for these authors are presented in the bibliography1. Szyperski T. (1994), 2. Szyperski T. (1992), 3. Sorensen M.D., 4. Jorgensen A.M., 5. Kohda D., 6. Montelione GT, 7. Schaller W, 8. Krezel AM, 9. Bode W, 10. Ebina S, 11. Williamson MP, 12. Chiang CM, 13. Singhal AK, 14. Sun YJ, 15. Kesvatera T, 16. Kordel J, 17. Skelton NJ, 18. Svensson LA, 19. Sundd M, 20. Cornilescu GM, 21. Vijay-Kumar S, 22. Assadi-Porter FM, 23. Girvin ME, 24. Gooley PR, 25. Fefeu S, 26. Boissy G, 27. Qin J (1996), 28. Qin J (1994), 29. Weichsel A, 30. Oliveberg M, 31. Bycroft M, 32. Mauguen Y, 33. Rico M, 34. Santoro J, 35. Fedorov AA, 36. Bartik K, 37. Schwalbe H, 38. Harata K, 39. Perez-Canadillas JM (1998), 40. Perez-Canadillas JM (2000), 41. Oda Y, 42. Fujiwara M, 43. Katayanagi K, 34. Santoro J.

(Warshel, 1981), and play key roles in protein solubility, folding, stability, binding ability and catalytic activity. Ionizable groups with unusually low or high pK_a values tend to occur at protein active sites and ligand binding sites (Kortemme et al., 1996; Forsyth et al., 2002). Thus the identification of unusual pK_a values may facilitate the identification of properly structured enzyme active sites (Elcock, 2001; Ondrechen et al., 2001), as well as establish their functional mechanisms (Harris, 2002; Nielsen & McCammon, 2003a, b).

Structural validation

The PROPKA method has been shown to predict pK_a values of Asp and Glu residues with an RMSD from experiment of 0.89 pH units by validation against more than 200 experimental values (Li et al., 2005). The accuracy for other ionizable groups appears comparable but has not been sufficiently validated due to a paucity of experimental data. Thus, a reasonable and straightforward use of PROPKA for structural validation is to eliminate structures for which one or more predicted pK_a values of Asp and Glu residues deviate from experiment by more than 1 pH unit. For an NMR ensemble of Q structures of a protein for which the Asp and Glu pK_a values have been experimentally measured, the procedure is as follows:

(1) Predict the pK_a values of all Glu and Asp residues in each ensemble structure using PROPKA.
(2) Calculate the absolute error between predicted and experimental values for each residue in each structure,

$$\Delta \mathrm{p}K_a \equiv |\mathrm{p}K_{a(\mathrm{Pred})} - \mathrm{p}K_{a(\mathrm{Exp})}| \quad (1)$$

If an experimental pK_a value consists of an upper or lower limit, then ΔpK_a is taken to be zero if the predicted pK_a value is above or below the limit, respectively.

(3) Retain only the (M) structures for which the maximum error is less than one,

$$\max[(\Delta \mathrm{p}K_a)_s] \leq 1 \quad (2)$$

As we demonstrate below, this criterion often leads to the elimination of all structures ($M = 0$). For these cases we

(4) Retain only the (N) structures for which

$$\max[(\Delta \mathrm{pK_a})] + \langle(\Delta \mathrm{pK_a})\rangle \equiv \varepsilon \leq 2 \quad (3)$$

where $\langle(\Delta \mathrm{pK_a})\rangle$ is the average of the absolute pK_a errors for the given structure. This criterion allows for the selection of structures that may have a large error for a single residue, but low error overall.

Results and Discussion

General assessment

We have applied the above criterion to 17 proteins for which both NMR structures and pK_a values of Asp and Glu residues have been determined, and the results are summarized in Table 2. In all cases a significant percentage of ensemble structures are eliminated using the first criterion [Equation (2), i.e. $M < Q$]. For example, 7 of the 20 structures of hirudin in the ensemble are eliminated, while the remaining 13 structures all are consistent with the experimental pK_a values (within the accuracy of the PROPKA approach).

In 10 cases none of the structures satisfy the first criterion ($M = 0$), but for 4 of these cases (bull seminal inhibitor, ubiquitin, RNase A, and RNase H1) some structures satisfy the second criterion [Equation (3), $N > 0$]. This leaves six cases [B9(Asp) insulin, subunit c of ATP synthase, oxidized and reduced thioredoxin, barnase, and lysozyme] for which none of the structures in the NMR ensemble are judged chemically accurate by either criterion outlined above. In three of these cases (insulin, barnase, and lyzosyme) the minimum values of ε are relatively close to 2 (2.3, 2.2 and 2.1), and one option is to eliminate all structures but the one with the lowest ε. However, these structures lead to a maximum pK_a error of 1.8, 1.5, and 1.5, respectively, and, as will be discussed next, in some cases this error can be reduced.

For example, for barnase, the structure with the lowest ε (structure number 18 in the NMR ensemble) has a maximum pK_a error of 1.5 pH units for Glu60. In this structure Glu60 is completely solvent exposed and has no hydrophobic, hydrogen bonding, or charge-charge interactions with the rest of the protein (Figure 1) and the

Table 2, Summary of the number of structures in each ensemble (Q) and the number of structures deemed chemically accurate using criterion 1 (M) and 2 (N) outlined in Section "Methods".

Protein	Residues	Q	M	N	NMR Min. ε	X-ray Min. ε	ε′	$\varepsilon_{\langle NMR \rangle}$
Hirudin	51	20	13	19	0.8		0.2	
Insulin	51	20	0	0	2.3		2.0	
Epidermal growth factor	53	16	12	15	0.8		0.1	
Turkey ovomucoid third domain	56	12	3	12	1.0	1.1	0.8	
Bull seminal inhibitor IIA	57	5	0	1	1.8		0.2	0.5
Cardiotoxin AV	62	2	1	2	1.7	1.6[a]	1.3	
Calbindin D_{9K}	76	65[a]	24	45	0.7	2.0	0.1	2.0
Ubiquitin	76	10	0	7	1.7	1.7	1.6	
Subunit c of H^+-transporting F_1F_0ATP synthase	79	10	0	0	4.3		4.2	
Cryptogein	98	18	1	2	1.2	1.8	1.1	
Thioredoxin (oxidized)	105	40	0	0	3.3	2.2	3.1	3.7
Thioredoxin (reduced)	105	40	0	0	4.6	3.0	4.3	5.0
Barnase	108	20	0	0	2.2	5.0[b]	1.3	
Bovine pancreatic Ribonuclease A	124	32	0	2	1.9	3.3	1.2	
Lysozyme (Hen)	129	50	0	0	2.1	1.4[a]	1.5	
alpha-sarcin	150	20	4	6	1.0		0.4	
Ribonuclease HI	155	8	0	2	2.0	1.9	1.4	

The lowest value of ε [cf. Equation (3)] found in of all ensemble structures found is given as "NMR Min.". Lower ε values correspond to the more chemically accurate structures. ε′ denotes an ε computed using only the best *local* structures of all the ensemble members to assemble the protein. Columns marked X-ray refer to values computed with the X-ray geometry and the $\varepsilon_{\langle NMR \rangle}$ refers to a prediction based on an average NMR structure when available.
[a]Based on two structures (X-ray)/ensembles (NMR).
[b]Based on three structures.

predicted pK_a is thus the unperturbed value of 4.5, 1.5 pH units higher than the experimental value of 3.0. The lowest ΔpK_a predicted for Glu60 is 1.1 pH units (using structure number 8) where the pK_a value is lowered by a hydrogen bond to Gln104 (Figure 1), suggesting that this hydrogen bond exists in solution. Thus, the structure with the smallest ε value (number 18) can be improved by introducing a hydrogen bond between Glu60 and Gln104 (perhaps as a an additional distance constraint in the structure refinement process or simply by changing select side-chain dihedral angles in structure 18). The ε for this new structure can be estimated to be 1.9 by recomputing ε using a ΔpK_a of 1.1 for Glu60 instead of 1.5. This new structure would thus satisfy the second criterion (i.e. $N = 1$). The lowest possible ε value based on the available ensemble structures can be estimated to be 1.3 based on the minimum ΔpK_a for each residue (ε′ in Table 2).

Similarly, structures with ε less than or equal to 2.0 can be constructed for B9(Asp) insulin and lysozyme. However, in the case of insulin, the

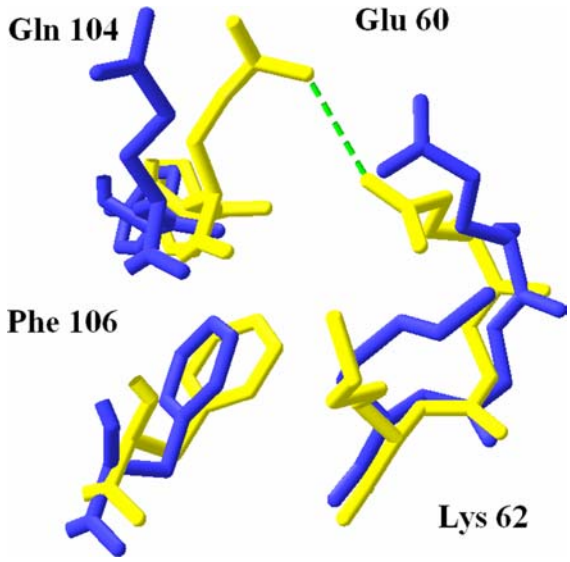

Figure 1. Overlay of structure 8 (yellow) and structure 18 (blue) in the NMR ensemble of barnase in the region surrounding Glu60. The pK_a value of Glu60 indicates that the hydrogen bond to Gln104 exists in solution and should be introduced in structure 18.

lowest pK_a value predicted for Glu13 on chain B is 3.8 while the experimental value is 2.2. This led us to consider the case of insulin more carefully, and conclude that the error is most likely due to the tendency of B9(Asp) insulin to dimerize in solution. The NMR structures were obtained at pH 1.8–1.9 where B9(Asp) insulin forms a symmetric dimer (Jorgensen et al., 1992). However, due to the symmetry, only one set of signals per residue is observed, and only the coordinates for the monomer were deposited in the PDB. Furthermore, B9(Asp) insulin remains a dimer in the 1.7–3.7 pH range used to obtain the pK_a values (for the monomer) (Jorgensen et al., 1992). Interestingly, Glu13 has been shown to be part of the binding interface (Ottensmeyer et al., 2000) and the pK_a of Glu13 is likely different in the dimer and the monomer. Unfortunately, the B9(Asp) insulin dimer structure is not available to verify this hypothesis.

In the case of subunit c of ATP synthase or oxidized and reduced thioredoxin, the ε' values are significantly larger than 2.0 (4.2, 3.1, and 4.3, respectively). Interestingly, the large value of ε' is due to the error found for a single residue of primary functional importance.

In the case of subunit c of F_0 ATP synthase the pK_a error from the internal H^+ transporting Asp61 residue (Dmitriev et al., 1999) contributes the bulk of ε', with a minimum error from all ensemble structures of 2.7 pK_a units. The experimental value of the Asp61 pK_a is 7.0, which is in contrast with the average predicted value of 4.1 from the monomer. Like insulin, the subunit c monomers are also known to aggregate, forming the c-ring in the F_0 complex of ATP synthase (Boyer, 1997; Weber & Senior, 2003), which could result in a significant pK_a shift. In this case a theoretical model of the oligomer is available (Dmitriev et al., 1999) composed of 12 F_0 c sub-units brought together to form a ring under a molecular dynamics simulation. Utilizing this structure to make pK_a predictions resulted in Asp61 pK_a values ranging from 5.2 to 6.4 (mainly due to desolvation). Using this theoretical oligomer to predict pK_a, the minimum ΔpK_a error due to this residue was reduced to from 2.7 to 0.9.

In the case of both oxidized and reduced thioredoxin the smallest ΔpK_a error (2.7 and 3.9 pH units, respectively) is observed for Asp26, is a conserved residue that plays an integral part in the function of thioredoxin (Gleason, 1992; Dyson et al., 1997). The values were computed using the respective experimental pK_a values of 8.1 and 9.9 of Qin et al. (1996) However, the titration curve of Asp26 in reduced thioredoxin shows evidence of coupled titration with other ionizable residues, and Chiver's et al. (1997) have argued that the pK_a of Asp26 depends strongly on whether nearby Cys32 is protonated. PROPKA predicts that Asp26 titrates before Cys32 and Chiver's et al. (1997) have determined the corresponding pKa value to be 7.5 for reduced thioredoxin from *E. coli*. Using 7.5 as a representative value for the experimental pK_a values of Asp26 in reduced thioredoxin leads to 4 structures with ε values less than 2.0 ($N = 4$).

In contrast, the titration curve of Asp26 in oxidized human thioredoxin shows no evidence of coupling and the pK_a value is clearly 8.1. One possible cause of the large ΔpK_a can be found by looking at the X-ray structure of oxidized thioredoxin. Using the X-ray structure, PROPKA predicts a pK_a value of 7.4, which is in good agreement with experiment. The high pK_a is caused by desolvation and a charge–charge interaction with Glu56. The latter interaction is missing in all NMR structures of the oxidized NMR structures due to a significantly larger Asp26-Glu56 separation. Interestingly, a similar interaction is present in the X-ray structure and the 4 NMR structures of reduced thioredoxin that satisfy criterion 2. This indicates that there is an Asp26-Glu56 interaction in the solution structure of both oxidized and reduced thioredoxin. It is not clear why the interaction is absent in the NMR structure of oxidized thioredoxin, but it may be due to the observed loss of two NOE contacts involving Glu56 on going from the reduced to the oxidized form of thioredoxin (Qin et al., 1994).

In summary, of the 17 NMR ensembles considered in this study, 7 contained at least one structure that lead to predicted pK_a values that were all within 1 pH unit of experiment, i.e. that was judged chemically accurate by criterion 1 [Equation (2)]. An additional 5 ensembles (including reduced thioredoxin) had at least one structure that was chemically accurate by the second, less stringent, criterion 2 [Equation (3)]. A similarly accurate structure can be easily constructed, based on the ensemble structures, for 2 (barnase and lysozyme) of the remaining 5 proteins as described above. For another 2 proteins [B9(Asp) insulin and subunit c of

F_0 ATP synthase] the measured pK_a are most likely affected by protein aggregation. Thus, only the NMR ensemble of oxidized thioredoxin did not yield a chemically accurate structure as judged by either criterion. However, the inaccuracy is local and affects only one pK_a value (Asp26). Furthermore, it may be possible that additional sampling would yield more accurate structures.

Comparison of selected NMR and X-ray geometries

X-ray structures are available for 11 of the 17 proteins considered (Tables 1 and 2) and we offer a comparison for a few representative cases. While the quality of NMR structures are often benchmarked by comparison to X-ray structures (usually by computing and RMSD of atomic positions) we note that 4 of the 11 X-ray structures result in ε values that are significantly larger than some of the NMR ensemble structures (calbindin D9K, cryptogein, barnase, and RNase A). Thus, the NMR structure with the lowest RMSD relative to the X-ray structure is not necessarily the most chemically accurate structure. Furthermore, due to the relatively complex dependence of pK_a values on protein structure there is no direct correlation between the RMSD and Δ pK_a.

In general, the relationships between NMR/X-ray RMSD and NMR ensemble ΔpK_a fall into four categories: (I) Large relative changes in pK_a with large relative changes in RMSD. (II) Large relative changes in pK_a with small relative changes in RMSD. (III) Small relative changes in pK_a with large changes in RMSD. (IV) Small relative changes in pK_a with small changes in RMSD. These general trends are depicted for bovine ribonuclease A and ribonuclease H1 in Figure 2, while representatives of each category are presented in Figure 3. In Figure 3 the RMSD between X-ray and NMR structures is computed by a least-squares alignment minimization (calculated using SwissPDB (Guex & Peitsch, 1997)) of all atoms of all amino acids within 6 Å of C_γ for aspartate, and C_δ for glutamate.

Residues in category I are exemplified by Glu131 in RNase H1 as shown in Figure 3A. Here the NMR structure that deviates most from the X-ray structure (in the protein region surrounding Glu131) is structure 4, resulting in the largest ΔpK_a (1.41) for Glu121, and largest RMSD (1.7 Å). Figure 4 shows a comparison of this structure with the X-ray structure (for which the ΔpK_a is 0.2). For comparison, structure 5, one of the two structures retained (of 5 and 8) using the second criterion ($N=2$, $M=0$), is also shown. The NMR

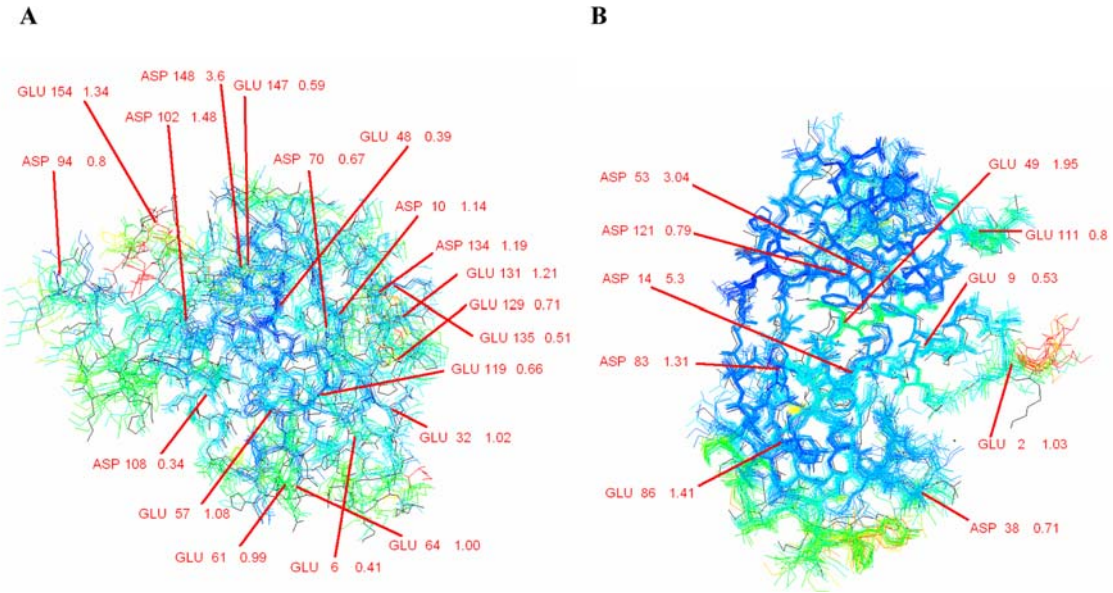

Figure 2. Overlays of the X-ray and NMR ensemble structures of (A) RNase A and (B) RNase H1. The NMR ensemble structures are colored according to their RMSD relative to the X-ray structure (black). Blue and red correspond to a low and high RMSD, respectively. The range of pK_a values predicted using the ensemble structures are also indicated for all Asp and Glu residues.

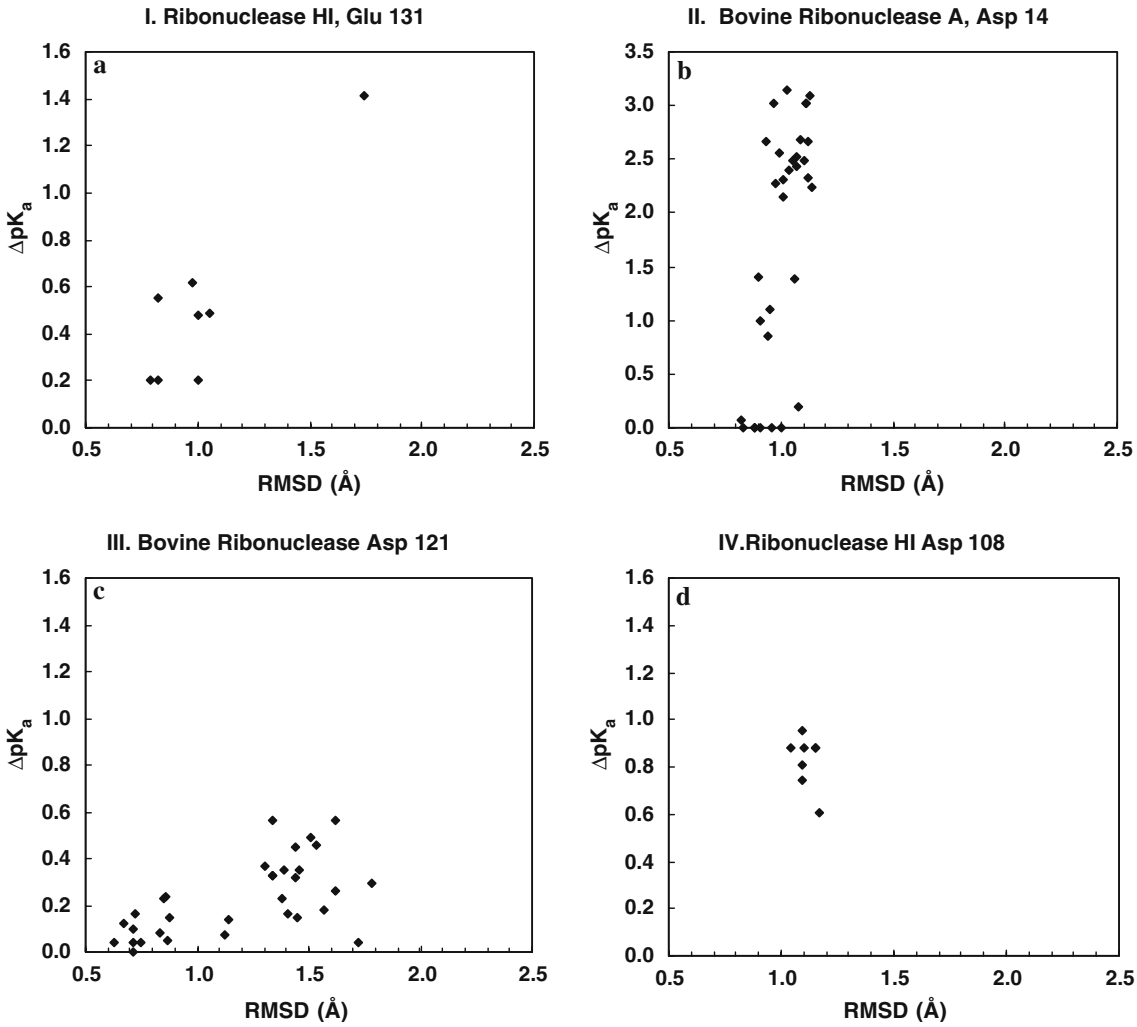

Figure 3. Four representative examples of the relationship between structural RMSD from the X-ray structure and the error in the predicted pK_a values relative to experiment.

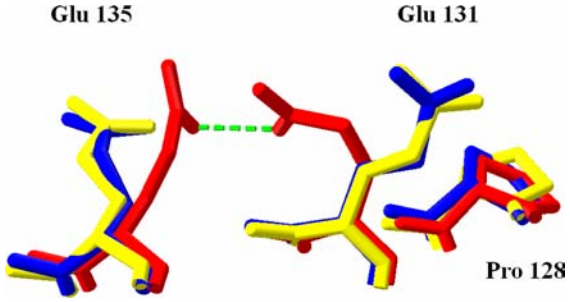

Figure 4. Overlay of structure 4 (red), structure 5 (blue) in the NMR ensemble of RNase A, and the X-ray structure (yellow) in the region surrounding Glu131. Structure 4 has the largest ΔpK_a as well as the largest RMSD relative to the X-ray structure (cf. Figure 3A).

structure 4 has Glu131 much closer to Glu135 than is found in the X-ray structure, leading to a hydrogen bond and a predicted pK_a value that is significantly higher than the experimental value. Applying either criterion outlined above eliminates this structure. The two selected structures, 5 and 8, both lack the interaction with Glu 135, as in the X-ray structure, leading to local ΔpK_a values of 0.2 and 0.6, and local RMSD values of 1.0 for both.

Residues in category II are exemplified by Asp14 in RNase A as shown in Figure 3B. Asp14 is a residue critical for protein stability, (Kim & Raines, 1993; Kalyan S. Chakrabarti, 2004) forming several strong inter-chain hydrogen bonds

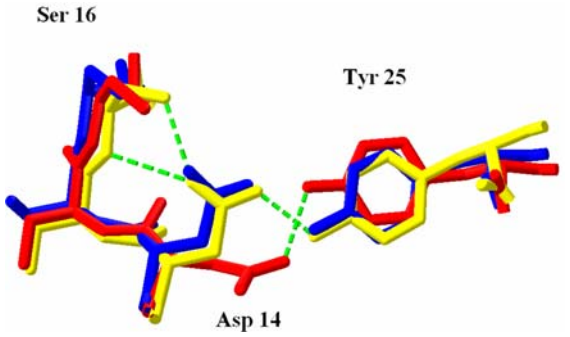

Figure 5. Overlay of structure 27 (red), structure 5 (blue) in the NMR ensemble of RNase H1, and the X-ray structure (yellow) in the region surrounding Asp14. Structure 27 has the largest ΔpK_a but an RMSD relative to X-ray very similar to structure 5 (cf. Figure 3B)

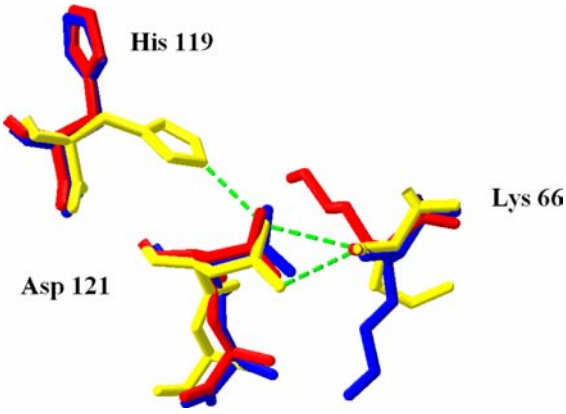

Figure 6. Overlay of structure 1 (red), structure 5 (blue) in the NMR ensemble of RNase H1, and the X-ray structure (yellow) in the region surrounding Asp121. Structures 1 and 5 have very similar pK_a values yet differ significantly from each other and from the X-ray structure (cf. Figure 3C).

that result in an extreme dependence of pK_a on local structure. Figure 5 shows a comparison of the NMR structure with the largest ΔpK_a, structure 27 (for which the ΔpK_a is 3.15), with the X-ray structure (for which the ΔpK_a is 0.4). The NMR structure 27 has only the Tyr25 hydrogen bond in common with X-ray and lacks the Ser16 and Ser16 backbone hydrogen bonds found in the X-ray structure. This leads to a predicted pK_a value that is significantly higher than the experimental value. Applying either criterion discussed above eliminates this structure. One structure is retained using the second criterion ($N=1$, $M=0$), structure 5, and this structure has the hydrogen bond observed in the X-ray structure (Figure 5), leading to a ΔpK_a value of 0.2 and a local RMSD value of 1.1. Residues in this category are often associated with binding or catalytic mechanisms and it is troubling to note that most of the NMR ensemble structures are chemically inaccurate in these regions. Elcock (2001) has shown that such critical residues have unusually high energies and this may tend to disfavor the correct conformation during the simulated annealing process.

Residues in category III are exemplified by the trend of Asp121 in RNase A as shown in Figure 3C. Here there is a weak dependence of pK_a on geometry, typical of many surface residues such as Asp121. In such cases, several varied conformations may exist with nearly identical pK_a values, as the primary interaction is between the residue and the solvent. Fortunately in these cases, a definite structure is often of less interest as well, as regions on the protein surface cannot be expected to have as definite a structure as internal regions, suggesting that these varied conformations may all be equally good. The single NMR structure retained using the second criterion, structure 5, $\Delta pK_a=0.2$, RMSD = 1.6 lacks any strong interactions with the rest of the protein with the exception of two possible backbone hydrogen bonds to Lys66, which are also observed in the X-ray structure (Figure 6). These are also observed for the structure with greatest ΔpK_a for this residue, structure 1 (with a ΔpK_a of 0.56 and local RMSD of 1.62 Å). The predicted pK_a's of all these varied structures are very similar however, suggesting that aside from the Lys66 hydrogen bonds, changes in the local structure, including an additional hydrogen bond from His119, do not significantly change the pK_a of this residue. Interestingly, Asp121 is important for catalysis (Trautwein et al., 1991), which requires that the residue to remain ionized for enzymatic function, and may suggest why the pK_a of this residue seems particularly immune to conformational change.

Residues in category IV are exemplified by Asp108 in RNase H1 as shown in Figure 3D. This is a residue within a region that has been accurately determined, with an average of 15 restraints per residue in the region of residues 101–115 (Fujiwara et al., 2000). The resulting structures are all very similar overall, and as a result, there is little discrepancy in the resulting pK_a predictions. There are also no strong interactions in this region of the protein to produce large pK_a shifts. The

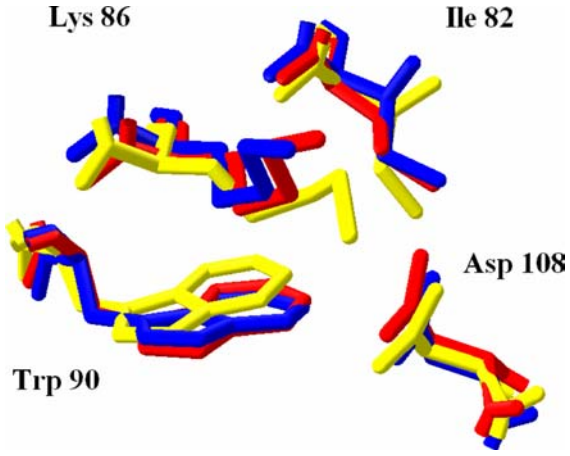

Figure 7. Overlay of structure 8 (red), structure 5 (blue) in the NMR ensemble of RNase A, and the X-ray structure (yellow) in the region surrounding Glu131. All structures are very similar and lead to very similar ΔpK_a values (cf. Figure 3D).

NMR structures are also similar to the X-ray in this region of the protein. Figure 7 shows a comparison of the X-ray structure and both of the two structures, 5 (with RMSD = 1.0 Å and $\Delta pK_a = 0.9$) and 8 (with RMSD = 1.1 and $\Delta pK_a = 0.95$), retained using the second criterion. In this case, structure 8 has the highest ΔpK_a for Asp108.

In summary, the NMR structures selected using criterion 2 lead to pK_a predictions that are within 1.0 pH units from experiment and have local RMSD values of 1.1–1.6 Å from the X-ray structure for the residues of RNase H1 and RNase A in Figure 3.

Average pK_a values

One interpretation of the different NMR structures in an ensemble is that they reflect the dynamical motion of the protein in solution. Therefore any property of the protein (such as a pK_a value) should reflect an average value computed using all the NMR ensemble structures. Such an average pK_a value should be more representative of the experimental value than any of the pK_a values computed with a single structure.

To investigate this issue further, we have computed the ensemble average pK_a values for each residue and compared it to the corresponding experimental pK_a value. The RMSD from experiment for all residues and maximum deviation from experiment observed for each protein is listed in Table 3. While the RMSD is generally low, maximum pK_a errors of >1 are observed for 12 of the 17 proteins. Since these the errors are significantly larger than the 1 pH unit criterion used in this study, the average pK_a values are not sufficiently accurate to the determine the chemical accuracy of protein structures as defined here. For example, using the average pK_a values instead of the

Table 3. RMS and maximum deviations from experiment computed using an ensemble averaged pK_a values for each protein.

Protein	RMSD pK_a deviation	Maximum pK_a deviation
Hirudin	0.14	0.49
Insulin	0.43	2
Epidermal growth factor	0.14	0.5
Turkey ovomucoid third domain	0.26	0.23
Bull seminal inhibitor IIA	0.5	0.5
Cardiotoxin AV	1.45	3.63
Calbindin D(9K)	0.15	0.65
Ubiquitin*	0.16	1.5
Subunit c of H^+-transporting F1F0ATP synthase	0.81	2.9
Cryptogein	0.55	1.16
Thioredoxin (oxidized)	0.22	2.9
Thioredoxin (reduced)	0.29	4.3
Barnase	0.31	2.7
Bovine pancreatic ribonuclease A	0.27	1.5
Lysozyme (Hen)	0.31	1.6
alpha-sarcin	0.15	1.9
Ribonuclease HI	0.15	1.32

See text [III-C] for further discussion.

experimentally determined values for RNase A leads to $M=16$ and $N=22$, compared to $M=0$ and $N=2$ obtained using experimental pK_a values. The discrepancy is mainly due to the fact that most structures result in relatively unshifted average pK_a values for Asp14 and Glu2 (3.5 and 4.5, respectively) while the experimental values are significantly shifted (<2.0 and 2.8, respectively). For comparison, using the X-ray structure leads to predicted pK_a values of −0.6 and 2.7, respectively.

Validation of X-ray geometries

The criteria outlined in this paper can also be used to validate X-ray structures. For example, two different X-ray structures were obtained for hen egg-white lysozyme by Harata (1994) and are present in the PDB file 1LYS. The two structures differ from each other in the flexible molecular regions of residues 46–49, 65–73, and 100–104. The author notes, ''The local structure of these regions differs between the monoclinic, triclinic and tetragonal crystals. This suggests that the conformational difference at the molecular surface reflects the crystal packing.'' A comparison of the predicted pK_a values for Asp 48 from the two structures can be used to reveal which of the two conformations is similar to that in solution. The pK_a of Asp 48 predicted for the first structure is 3.6, for the second it is 1.7. The experimental value given by Bartik et al. (1994) is <2.5. The subsequent assumption that the second structure most likely resembles lysozyme in solution was then substantiated by lysozyme NMR structural data (1E8L). It was found that the average RMSD between the NMR ensemble structures and the X-ray structures was 3.1 Å for the first X-ray structure, and 1.7 Å for the second.

Summary

A method is presented for evaluating the ''chemical accuracy'' of protein ensemble structures determined by NMR by (1) measuring the pK_a values of the ionizable residues in the protein and (2) comparing them to the values predicted for each ensemble structure using the PROPKA method. Two more and less stringent criteria for judging the chemical accuracy are proposed (Section ''Methods'') and applied to ensemble structures for 17 proteins for which the pK_a values have been determined. In 14 of the cases at least one chemically accurate structure can be identified or constructed, while evidence of protein aggregation is found in 2 other cases. Thus, a chemically accurate structure could not be found in only in one case (oxidized thioredoxin).

Comparison of the chemically accurate NMR structures to available X-ray structures show a local RMSD of roughly 1 Å for the residues where there is a strong dependence of the pK_a value with protein structure. These residues are typically in regions of the protein critical to structure or function and only a few NMR ensemble structures are chemically accurate in these regions.

There are, of course, many ways in which the general approach presented here can be modified or extended. Different cutoffs, pK_a prediction methods, or even entirely different criteria for chemical accuracy can be applied, and they can be applied to other types of residues (such as His or Cys or to only a subset of residues [perhaps only in the active site as done previously by Dillet et al. (1998)]. We hope the current work will provide the impetus for such further study.

Acknowledgements

This work was supported by a grant from the National Science Foundation (MCB 209941). We thank Prof. Lawrence McIntosh for helpful suggestions on the manuscript.

References

Assadi-Porter, F.M. and Fillingame, R.H. (1995) *Biochemistry*, **34**(49), 16186–16193.
Bartik, K., Redfield, C. and Dobson, C.M. (1994) *Biophys. J.*, **66**(4), 1180–1184.
Bode, W., Wei, A.Z., Huber, R., Meyer, E., Travis, J. and Neumann, S. (1986) *EMBO J.*, **5**(10), 2453–2458.
Boissy, G., de LaFortelle, E., Kahn, R., Huet, J.C., Bricogne, G., Pernollet, J.C. and Brunie, S. (1996) *Structure*, **4**(12), 1429–1439.
Boyer, P.D. (1997) *Ann. Rev. Biochem.*, **66**, 717–749.
Brunger, A.T. and Adams, P.D. (2002) *Acc. Chem. Res.*, **35**(6), 404–412.
Brunger, A.T., Adams, P.D. and Rice, L.M. (1997) *Structure*, **5**(3), 325–336.
Bycroft, M., Ludvigsen, S., Fersht, A.R. and Poulsen, F.M. (1991) *Biochemistry*, **30**(35), 8697–8701.

Chiang, C.M., Chang, S.L., Lin, H.J. and Wu, W.G. (1996) *Biochemistry*, **35**(28), 9177–9186.
Chivers, P.T., Prehoda, K.E., Volkman, B.F., Kim, B.M., Markley, J.L. and Raines, R.T. (1997) *Biochemistry*, **36**(48), 14985–14991.
Clore, G.M. and Gronenborn, A.M. (1998a) *Trend. Biotechnol.*, **16**(1), 22–34.
Clore, G.M. and Gronenborn, A.M. (1998b) *PNAS*, **95**(11), 5891–5898.
Cornilescu, G., Marquardt, J.L., Ottiger, M. and Bax, A. (1998) *J. Am. Chem. Soc.*, **v120**(27), 6836–6837(Communication).
Davis, I.W., Murray, L.W., Richardson, J.S. and Richardson, D.C. (2004) *Nucleic Acids Res.*, **32**(Web Server issue), W615–W619.
DePristo, M.A., de Bakker, P.I. and Blundell, T.L. (2004) *Structure (Camb.)*, **12**(5), 831–838.
Dillet, V., Dyson, H.J. and Bashford, D. (1998) *Biochemistry*, **37**(28), 10298–10306.
Dmitriev, O.Y., Jones, P.C. and Fillingame, R.H. (1999) *Proc. Natl. Acad. Sci. USA*, **96**(14), 7785–7790.
Doreleijers, J.F., Vriend, G., Raves, M.L. and Kaptein, R. (1999) *Proteins*, **37**(3), 404–416.
Durbin, S.D. and Feher, G. (1996) *Ann. Rev. Phys. Chem.*, **47**, 171–204.
Dyson, H.J., Jeng, M.-F., Tennant, L.L., Slaby, I., Lindell, M., Cui, D.-S., Kuprin, S. and Holmgren, A. (1997) *Biochemistry*, **36**(9), 2622–2636.
Ebina, S. and Wuthrich, K. (1984) *J. Mol. Biol.*, **179**(2), 283–288.
Elcock, A.H. (2001) *J. Mol. Biol.*, **312**(4), 885–896.
Fedorov, A.A., Joseph-McCarthy, D., Fedorov, E., Sirakova, D., Graf, I. and Almo, S.C. (1996) *Biochemistry*, **35**(50), 15962–15979.
Fefeu, S., Bouaziz, S., Huet, J.C., Pernollet, J.C. and Guittet, E. (1997) *Protein Sci.*, **6**(11), 2279–2284.
Forsyth, W.R., Antosiewicz, J.M. and Robertson, A.D. (2002) *Proteins*, **48**(2), 388–403.
Fujiwara, M., Kato, T., Yamazaki, T., Yamasaki, K. and Nagayam, K. (2000) *Biol. Pharm. Bull.*, **23**(10), 1147–1152.
Garavito, R.M. and Ferguson-Miller, S. (2001) *J. Biol. Chem.*, **276**(35), 32403–32406.
Georgescu, R.E., Alexov, E.G. and Gunner, M.R. (2002) *Biophys. J.*, **83**(4), 1731–1748.
Girvin, M.E., Rastogi, V.K., Abildgaard, F., Markley, J.L. and Fillingame, R.H. (1998) *Biochemistry*, **37**(25), 8817–8824.
Gleason, F.K. (1992) *Protein Sci.*, **1**(5), 609–616.
Gooley, P.R., Keniry, M.A., Dimitrov, R.A., Marsh, D.E., Keizer, D.W., Gayler, K.R. and Grant, B.R. (1998) *J. Biomol. NMR*, **12**(4), 523–534.
Guex, N. and Peitsch, M.C. (1997) *Electrophoresis*, **18**(15), 2714–2723.
Hall, R.J., Hindle, S.A., Burton, N.A. and Hillier, I.H. (2000) *J. Comput. Chem.*, **21**(16), 1433–1441.
Harata, K. (1994) *Acta Crystallogr. D Biol. Crystallogr.*, **v50**, 250–257.
Harris, T.K. and Turner, G.J. (2002) *IUBMB LIFE*, **53**, 85–98.
Hooft, R.W., V.G., , Sander, C. and Abola, E.E. (1996) *Nature*, **381**, 272 .
Jorgensen, A.M., Kristensen, S.M., Led, J.J. and Balschmidt, P. (1992) *J. Mol. Biol.*, **227**(4), 1146–1163.
Kalyan S. Chakrabarti, B.S.S.S.V. (2004) *Chemistry & Biodiversity*, **1**(5), 802–818.
Katayanagi, K., Miyagawa, M., Matsushima, M., Ishikawa, M., Kanaya, S., Nakamura, H., Ikehara, M., Matsuzaki, T. and Morikawa, K. (1992) *J. Mol. Biol.*, **223**(4), 1029–1052.
Kesvatera, T., Jonsson, B., Thulin, E. and Linse, S. (1996) *J. Mol. Biol.*, **259**(4), 828–839.
Kim, J. and Raines, R.T. (1993) *Protein Sci.*, **2**(3), 348–356.
Kohda, D., Sawada, T. and Inagaki, F. (1991) *Biochemistry*, **30**(20), 4896–4900.
Kordel, J., Skelton, N.J., Akke, M. and Chazin, W.J. (1993) *J. Mol. Biol.*, **231**(3), 711–734.
Kortemme, T., Darby, N.J. and Creighton, T.E. (1996) *Biochemistry*, **35**(46), 14503–14511.
Kragh-Hansen, U., le Maire, M. and Moller, J.V. (1998) *Biophys. J.*, **75**(6), 2932–2946.
Krezel, A.M., Darba, P., Robertson, A.D., Fejzo, J., Macura, S. and Markley, J.L. (1994) *J. Mol. Biol.*, **242**(3), 203 .
Kuriyan, J., Brunger, A.T., Karplus, M. and Hendrickson, W.A. (1989) *Acta Crystallogr. A*, **45**(Pt 6), 396–409.
Laskowski, R.A., MacArthur, M.W. and Thornton, J.M. (1998) *Curr. Opin. Struct. Biol.*, **8**(5), 631–639.
Laskowski, R.A., Rullmann, J.A.C., MacArthur, M.W., Kaptein, R. and Thornton, J.M. (1996) *J. Biomol. NMR*, **8**(4), 477(Historical Archive).
Li, H., Robertson, A.D. and Jensen, J.H. (2004) *Proteins*, **55**(3), 689–704.
Li, H., Robertson, A.D. and Jensen, J.H. (2005). *Struct. Funct. Bioinform.* 61(4), 704–721.
Linge, J. P., Williams, M. A., Spronk, C. A. E. M., Bonvin, A. M. J. J. and Nilges M. (2003). *Proteins Struct. Funct. Genet.* **50**(3): 496–506.
Mauguen, Y., Hartley, R.W., Dodson, E.J., Dodson, G.G., Bricogne, G., Chothia, C. and Jack, A. (1982) *Nature*, **297**(5862), 162–164.
Montelione, G.T., Wuthrich, K., Burgess, A.W., Nice, E.C., Wagner, G., Gibson, K.D. and Scheraga, H.A. (1992) *Biochemistry*, **31**(1), 236–249.
Montelione, G.T., Zheng, D., Huang, Y.J., Gunsalus, K.C. and Szyperski, T. (2000) *Nat. Struct. Mol. Biol.*, **Nov**, 982 .
Nabuurs, S.B., Spronk, C.A., Krieger, E., Maassen, H., Vriend, G. and Vuister, G.W. (2003) *J. Am. Chem. Soc.*, **125**(39), 12026–12034.
Nagendra, H.G., Sukumar, N. and Vijayan, M. (1998) *Proteins Struct. Funct. Genet.*, **32**(2), 229–240.
Nielsen, J.E. and McCammon, J.A. (2003a) *Protein Sci.*, **12**(9), 1894–1901.
Nielsen, J.E. and McCammon, J.A. (2003b) *Protein Sci.*, **12**(2), 313–326.
Nielsen, J.E. and Vriend, G. (2001) *Proteins*, **43**(4), 403–412.
Oda, Y., Yamazaki, T., Nagayama, K., Kanaya, S., Kuroda, Y. and Nakamura, H. (1994) *Biochemistry*, **33**(17), 5275–5284.
Oliveberg, M., Arcus, V.L. and Fersht, A.R. (1995) *Biochemistry*, **34**(29), 9424–9433.
Ondrechen, M.J., Clifton, J.G. and Ringe, D. (2001) *Proc. Natl. Acad. Sci. USA*, **98**(22), 12473–12488.
Ottensmeyer, F.P., Beniac, D.R., Luo, R.Z. and Yip, C.C. (2000) *Biochemistry*, **39**(40), 12103–12112.
Perez-Canadillas, J.M., Campos-Olivas, R., Lacadena, J., del Pozo, A.M., Gavilanes, J.G., Santoro, J., Rico, M. and Bruix, M. (1998) *Biochemistry*, **37**(45), 15865–15876.
Perez-Canadillas, J.M., Santoro, J., Campos-Olivas, R., Lacadena, J., Martinez del Pozo, A., Gavilanes, J.G., Rico, M. and Bruix, M. (2000) *J. Mol. Biol.*, **299**(4), 1061–1073.
Qin, J., Clore, G.M. and Gronenborn, A.M. (1994) *Structure*, **2**(6), 503–522.
Qin, J., Clore, G.M. and Gronenborn, A.M. (1996) *Biochemistry*, **35**(1), 7–13.

Rico, M., Santoro, J., Gonzalez, C., Bruix, M. and Neira, J. L. (1990). *Structure, Mechanism and Function of Ribonucleases Proceedings of the 2nd International Meeting*.

Santoro, J., Gonzalez, C., Bruix, M., Neira, J.L., Nieto, J.L., Herranz, J. and Rico, M. (1993) *J. Mol. Biol.*, **229**(3), 22–34.

Schaller, W. and Robertson, A.D. (1995) *Biochemistry*, **34**(14), 4714–4723.

Schwalbe, H., Grimshaw, S.B., Spencer, A., Buck, M., Boyd, J., Dobson, C.M., Redfield, C. and Smith, L.J. (2001) *Protein Sci.*, **10**(4), 677–688.

Singhal, A.K., Chien, K.Y., Wu, W.G. and Rule, G.S. (1993) *Biochemistry*, **32**(31), 8036–8044.

Skelton, N.J., Kordel, J. and Chazin, W.J. (1995) *J. Mol. Biol.*, **249**(2), 441–462.

Snyder, D.A., Bhattacharya, A., Huang, Y.J. and Montelione, G.T. (2005) *Proteins*, **59**(4), 655–661.

Sorensen, M.D. and Led, J.J. (1994) *Biochemistry*, **33**(46), 13727–13733.

Spronk, C.A.E.M., Linge, J.P., Hilbers, C.W. and Vuister, G.W. (2002) *J. Biomol. NMR*, **22**(3), 281 .

Sun, Y.J., Wu, W.G., Chiang, C.M., Hsin, A.Y. and Hsiao, C.D. (1997) *Biochemistry*, **36**(9), 2403–2413.

Sundd, M., Iverson, N., Ibarra-Molero, B., Sanchez-Ruiz, J.M. and Robertson, A.D. (2002) *Biochemistry*, **41**(24), 7586–7596.

Svensson, L.A., Thulin, E. and Forsen, S. (1992) *J. Mol. Biol.*, **223**(3), 601–606.

Szyperski, T., Antuch, W., Schick, M., Betz, A., Stone, S.R. and Wuthrich, K. (1994) *Biochemistry*, **33**(31), 9303–9310.

Szyperski, T., Guntert, P., Stone, S.R. and Wuthrich, K. (1992) *J. Mol. Biol.*, **228**(4), 1193–1205.

Tarek, M. and Tobias, D.J. (2002) *Phys. Rev. Lett.*, **88**(13), 138101.

Trautwein, K., Holliger, P., Stackhouse, J. and Benner, S.A. (1991) *FEBS Lett.*, **281**(1–2), 275–277.

Vijay-Kumar, S., Bugg, C.E. and Cook, W.J. (1987) *J. Mol. Biol.*, **194**(3), 531–544.

Warshel, A. (1981) *Acc. Chem. Res.*, **14**, 284–290.

Weber, J. and Senior, A.E. (2003) *FEBS Lett.*, **545**(1), 61–70.

Weichsel, A., Gasdaska, J.R., Powis, G. and Montfort, W.R. (1996) *Structure*, **4**(6), 735–751.

Wiencek, J.M. (1999) *Ann. Rev. Biomed. Eng.*, **1**, 505–34.

Williamson, M.P., Havel, T.F. and Wuthrich, K. (1985) *J. Mol. Biol.*, **182**(2), 295–315.

Zhang, Y., Kua, J. and McCammon, J.A. (2002) *J. Am. Chem. Soc.*, **124**(35), 10572–10577.