

Mapping of protein structural ensembles by chemical shifts

Kumaran Baskaran · Konrad Brunner ·
Claudia E. Munte · Hans Robert Kalbitzer

Received: 9 July 2010 / Accepted: 19 July 2010 / Published online: 1 August 2010
© Springer Science+Business Media B.V. 2010

Abstract Applying the chemical shift prediction programs SHIFTX and SHIFTS to a data base of protein structures with known chemical shifts we show that the averaged chemical shifts predicted from the structural ensembles explain better the experimental data than the lowest energy structures. This is in agreement with the fact that proteins in solution occur in multiple conformational states in fast exchange on the chemical shift time scale. However, in contrast to the real conditions in solution at ambient temperatures, the standard NMR structural calculation methods as well chemical shift prediction methods are optimized to predict the lowest energy ground state structure that is only weakly populated at physiological temperatures. An analysis of the data shows that a chemical shift prediction can be used as measure to define the minimum size of the structural bundle required for a faithful description of the structural ensemble.

Keywords Chemical shift · Solution structure · Structural ensemble

Introduction

Since the advent of protein structure determination it has been a long time debate if X-ray crystallography is clearly

superior to NMR spectroscopy because X-ray structures are better defined than NMR structures. This is true when focusing to the precision of the coordinates when the available crystals diffract sufficiently well. It is often not realized that the two methods cannot calculate directly the three-dimensional structures of proteins from the experimental data but use iterative search algorithms to find a solution that is consistent with the data. The main differences are in the language of NMR spectroscopy that in general in X-ray crystallography the number and precision of the structural restraints is superior and that the back calculation of NMR spectra from structural models is not as straight forward as the back calculation of diffraction patterns.

However, for proteins in solution the question may be ill-posed since the minimum energy state in the crystal lattice (or better the weighted average of the structural ensemble in the crystal) may not correspond well enough to the average of the structural ensemble in solution. There are a number of examples for that in literature; a typical example is HPr from *S. faecalis* where we could show that the dominant active center structure in solution (Hahmann et al. 1998) clearly differs from the crystal structure (Jia et al. 1994). In many of these cases, probably more than one conformational state (defined by different local energy minima) exists in solution but only one of them is selected by the crystallization conditions. In addition, software packages used in crystallography tend to suppress alternative conformations even when they are present in the single crystals. A prominent case is the Ras protein complex with Mg^{2+} . GppNHp that exists in solution in two almost equally populated conformational states (Geyer et al. 1996; Spoerner et al. 2001). The published crystal structure (Pai et al. 1990) shows only a single, well-resolved structure but solid-state NMR on the same crystals

K. Baskaran · K. Brunner · C. E. Munte · H. R. Kalbitzer
Department of Biophysics and Physical Biochemistry,
University of Regensburg, 93040 Regensburg,
Federal Republic of Germany

H. R. Kalbitzer (✉)
Institut für Biophysik und Physikalische Biochemie,
Lehrstuhl Biophysik, Universität Regensburg,
Postfach, 93040 Regensburg, Federal Republic of Germany
e-mail: hans-roboter.kalbitzer@biologie.uni-regensburg.de

proves that the two conformational states coexist also in the single crystals (Stumber et al. 2002; Iuga et al. 2004).

Even when only one global energy minimum exists that is identical in crystal and solution, the thermally populated states create a conformational ensemble that in general is different in solid state and solution, since the details of the local energy surface will most probably be different. A full structural description of a protein would require the knowledge of the whole ensemble of structures not only a minimum structure because the properties of the protein may depend on a subset of structures which are similar but not identical to the minimum energy structure.

In crystallography, the B-factor is used to describe the thermally induced conformers as the atoms move from their average positions. In solution, classically relaxation time measurements give information about atomic motions in the local conformational space. Chemical shifts could also provide information on the local conformational equilibrium, since they are strongly structure dependent. Close to a single energy minimum they are usually population averaged since the exchange between neighboring states should be fast on the NMR time scale.

The main problem is that protein chemical shifts cannot be predicted with high accuracy, in spite of the many groups that have worked on this problem over the years. The first attempts to calculate protein chemical shifts started already in 1977 (Perkins et al. 1977). In the mean time, a number of programs are available that are able to predict chemical shifts from a structural data base or calculate chemical shifts from a physical model. Popular examples are SHIFTS (Xu and Case 2001) and SHIFTX (Neal et al. 2003) but also other programs exist such as PROSHIFT (Meiler 2003), PSRI (Wang 2004), SPARTA (Shen and Bax 2007), and 4DSpot (Lehtivarjo et al. 2009).

SHIFTS is a program for predicting ^{15}N , $^{13}\text{C}^\alpha$, $^{13}\text{C}^\beta$, and $^{13}\text{C}'$ chemical shifts from protein structures. It was developed based on an additive model of chemical shift contributions, corresponding to various conformational effects found in a database of density functional theory (DFT) calculations on more than 2,000 peptides. Some empirical extensions were used for covering additional regions in the conformation space for different residue types. When experimental shifts are available, an optional refinement process for side-chain orientation can also be carried out, which may help identify problems in either the structure or the shift assignments themselves. SHIFTX can be used to predict all backbone and some of side chain ^1H , ^{13}C and ^{15}N protein chemical shifts using only its PDB file as input. SHIFTX uses a unique semi-empirical approach to calculate protein chemical shifts.

In this paper we will focus on the ensemble properties of protein structures and their connections to chemical shifts calculated from these ensembles.

Materials and methods

NMR spectroscopy and structures

The sequential assignments of the NMR signals of the set of proteins (Table 1) were taken from the BMRB data base, the corresponding NMR structures from the PDB-data base. Sequential assignments for wildtype HPr(wt) from *S. aureus* were taken from Maurer et al. (2004), for the mutant HPr(H15A) from Munte et al. (manuscript in preparation).

Molecular dynamics calculations

Structure calculations on HPr were performed using the molecular dynamics program CNS v.1.2. (Crystallography and NMR System for crystallographic and NMR structure determination) (Brunger et al. 1998). The restraints (Table 2) were employed in a simulated annealing protocol (Brunger 2007) using extended-strands as starting structures. The high number of experimental restraints required a threefold reduction of the time step (default value 15 fs) for the integration of the equation of motion. In the first stage simulated annealing was performed in the torsional angle space starting at 50,000 K and cooling down in 1,000 temperature steps to 0 K. 3,000 time steps were calculated at each temperature. The second annealing stage was

Table 1 Test data set of proteins

Protein ^a	BMRB ID ^b	PDB ID ^c
Bet v 1-L (159)	4,417 (pH 7.0, 298 K)	1B6F (23)
P14a (135)	4,301 (pH 5.5, 303 K)	1CFE (20)
CA RSV (262)	4,384 (pH 6.0, 303 K)	1D1D (20)
Pathogenesis-protein (159)	4,671 (pH 7.0, 298 K)	1E09 (22)
Dynein light chain 8 (89)	4,911 (pH 7.0, 298 K)	1F96 (20)
Phosphoglycerate mutase (211)	4,648 (pH 6.4, 310 K)	1FZT (21)
HTL V-1 capsid protein(134)	4,649 (pH 6.0, 302 K)	1G03 (20)
β 2-GP1 domain V (86)	4,981 (pH 6.0, 298 K)	1G4F (20)
ERp29 C-domain(120)	4,920 (pH 4.9, 308 K)	1G7D (20)
Erp29 N-domain(137)	4,919 (pH 4.9, 308 K)	1G7E (20)
CDC4P (141)	4,851 (pH 6.5, 303 K)	1GGW (26)
Vam3p N-terminal (123)	4,945 (pH 6.0, 302 K)	1HS7 (20)
Mouse doppel (132)	4,938 (pH 5.2, 299 K)	1I17 (20)
Core binding factor (143)	4,092 (pH 6.6, 293 K)	2JHB (20)
Rabphilin_3_C2B (140)	4,360 (pH 6.1, 304 K)	3RPB (20)
P55 (166)	4,321 (pH 6.5, 300 K)	5GCN(24)

The data set comprises a subset of NMR structures with published chemical shifts used for an analysis by Wang and Jardetzky (2002)

^a Values in brackets, number of residues

^b Values in brackets, experimental conditions

^c Values in brackets, number of structures deposited

Table 2 Number of experimental restraints used for the calculation of the structural sets of wildtype HPr and the mutant HPr(H15A)

	HPr(wt)	HPr(H15A)
NOE restraints	1,219	1,248
Dihedral	130	130
<i>J</i> -coupling	78	69
H-bond	51	53

performed with Cartesian dynamics with a starting temperature of 1,000 K and a final temperature of 0 K using 1,000 temperature steps. Again, 3,000 time steps were calculated at each temperature. In the final stage, 2,000 time steps of energy minimization were performed. As force field the default settings were used. The obtained structures were accepted based on the NOE violations. Those structures having more than 5% NOE violations after the final energy minimization were rejected. Once 2,000 structures were calculated using the above simulated annealing protocol, they were refined in explicit water using the TIP3P model (Jorgensen et al. 1983) according to a protocol given by Linge et al. (2003). It consists of three stages, a heating stage from 100 to 500 K, a refinement stage at 500 K, and a cooling stage from 500 to 25 K followed by a short conjugate gradient minimization. The force field OPLSX (Jorgensen and Tirado-Rives 1988) proposed in this protocol was used. After the water refinement the population distribution is fitted with a Gaussian distribution and those structures whose energy is $>5\sigma$ is removed and refined again with different initial seeds until their energies were $<5\sigma$.

Programs and structure validation

The program PROCHECK_NMR (Laskowski et al. 1996) was employed to check the stereochemical quality by calculating Ramachandran plots. The program MOLMOL (Koradi et al. 1996) was used to display the structures and to calculate the RMSD-values. The combined chemical shift based error ε (eq. 4–8) were calculated with the chemical shift and atom specific weighting factors published by Schumann et al. (2007).

Theoretical considerations

Structural ensembles and chemical shifts

In solution, a protein is described by a multistate energetic profile, at a given time t it is described by a space ensemble $S^V = \{s_1, s_2, \dots, s_N\}$ with N the number of molecules in the solution. In a typical NMR experiment (0.5 mL of a 1 mM

solution) N equals to 3.01×10^{20} . In addition, for each individual molecule in solution a time ensemble S^T is defined as all structural states visited in a time interval Δt . The NMR spectrum obtained in a typical repetitive n D-NMR experiment represents a non-uniform spatial and temporal average of these states. However, the averaging mechanism depends on different NMR properties (e.g., chemical shift and *J*-coupling), that vary from atom to atom in the same molecule, and may also depend on the path on the energy landscape.

In a time interval Δt the ensemble can be divided in subsets where the exchange between different states is fast on the NMR time-scale for a given atom i . For these subsets S_k of N_k molecules the chemical shift δ_i of an atom i corresponds to the population average $\langle \delta_i(s_j) \rangle$ of the shifts $\delta_i(s_j)$ with

$$\langle \delta_i \rangle = \frac{1}{N_k} \sum_{j=1}^{N_k} \delta_i(s_j) \quad (1)$$

The fast exchange condition can be defined by

$$\frac{1}{\tau(s_j, s_k)} \gg |\omega_i(s_j) - \omega_i(s_k)| \quad (2)$$

with $\tau(s_j, s_k)$ the exchange correlation time for the transition between states s_j and s_k , and $\omega_i(s_j)$ and $\omega_i(s_k)$ the resonance frequencies of nucleus i in states s_j and s_k , respectively. In its simplest form the fast exchange condition must apply for all pairs of states s_j and s_k .

Since in the repetitive NMR experiment time averaging over the possible M^* structural sub states of the total ensemble M where mutual fast exchange conditions exist is also performed, eq. 1 is better written as

$$\langle \delta_i^* \rangle = \sum_{j=1}^{M^*} p(s_j) \delta_i(s_j) = \frac{1}{Z} \sum_{j=1}^{M^*} \delta_i(s_j) e^{-\frac{G(s_j)}{RT}} \quad (3)$$

Here, δ_i^* is the chemical shift of atom i in the sub states in mutual fast exchange, $p(s_j)$ the probability to find state s_j , Z the state sum over all possible structural states of M^* , and $G(s_j)$ the corresponding free enthalpies. Note that the fast exchange condition has not to be fulfilled for all atoms of the macromolecule simultaneously since the selection of structural states enclosed in M^* depends on the differences of the resonance frequencies of the states.

For the sake of simplicity we will restrict in the following to an ensemble where for all (or essentially all) structures fast exchange conditions apply ($M^* = M$). Let us denote the experimentally measured chemical shift of atom i as δ_i^e , the predicted average chemical shift of the same atom in a structure s_j in the ensemble as $\delta_i^p(s_j)$. Then the mean of the predicted chemical shifts δ_i^p of the ensemble of N structures is given by

$$\langle \delta_i^p \rangle = \frac{1}{N} \sum_{j=1}^N \delta_i^p(s_j) \quad (4)$$

Using the Hamming distance the error ε_i in the back calculation for a single atom i can be defined as

$$\varepsilon_i = |\langle \delta_i^p \rangle - \delta_i^e| \quad (5)$$

Alternatively, using the Euclidean distance it would be defined as

$$\varepsilon_i = \frac{1}{N} \sqrt{\sum_{j=1}^N (\delta_i^p(s_j) - \delta_i^e)^2} \quad (6)$$

The mean error ε for a subset of n atoms (e.g., all backbone atoms H^N, N, C^α, C in the protein or all atoms of a given amino acid) of structural ensemble is defined as

$$\varepsilon = \frac{1}{\sum_{i=1}^n w_i} \sum_{i=1}^n w_i \varepsilon_i \quad (7)$$

or alternatively

$$\varepsilon = \sqrt{\frac{1}{\sum_{i=1}^n w_i^2} \sum_{i=1}^n (w_i \varepsilon_i)^2} \quad (8)$$

where w_i are weighting factors e.g., the atom and amino acid specific weighting factors as defined by Schumann et al. (2007). The second moment σ^2 of the errors ε_i of a selection of n atoms is then

$$\sigma^2 = \langle w_i^2 \varepsilon_i^2 \rangle - \langle w_i \varepsilon_i \rangle^2 = \frac{1}{\sum_{i=1}^n w_i^2} \sum_{i=1}^n (w_i \varepsilon_i - \varepsilon)^2 \quad (9)$$

The expectation value of ε should go to zero if (1) the experimental data are error free, (2) the experimental ensemble and the ensemble used for the prediction of the chemical shifts are identical and (3) the chemical shift calculation is perfect. In practice, all three conditions are not fulfilled. The experimental data have errors that are caused by assignment errors and the limited precision of chemical shift measurements. The experimental ensemble is not known but has to be replaced by an ensemble obtained from the structure calculation, usually in NMR spectroscopy a restrained molecular dynamics simulation followed by an energy minimization. In general, the number N_p of structures is also much smaller than the experimental ensemble with N of the order of 10^{20} . Up to now the classical chemical shift calculations are far from perfect although they are getting better with time.

The error ε can then be written as a function of the experimental error in the experimental determination of chemical shifts $\Delta\delta^e$, the differences between the experimental ensemble and the calculated ensemble ΔS and the error of the chemical shift calculation method $\Delta\delta^s$ as

$$\varepsilon(\Delta\delta^e, \Delta S, \Delta\delta^s) = \varepsilon(0, 0, 0) + \frac{\partial \varepsilon(0, 0, 0)}{\partial \Delta\delta^e} \Delta\delta^e + \frac{\partial \varepsilon(0, 0, 0)}{\partial \Delta S} \Delta S + \frac{\partial \varepsilon(0, 0, 0)}{\partial \Delta\delta^s} \Delta\delta^s \quad (10)$$

Although the prediction error $\Delta\delta^s$ not only depends on the simulation method C used and the atom types T included in calculation but also on specific structural properties of the protein under consideration, for a given method C it can be approximated by a constant $\Delta\delta^s(C, T)$ and the corresponding derivative by 1 when enough atoms are involved in the calculation. When we neglect the small experimental error $\Delta\delta^e$ and realize that in the above definition $\varepsilon(0, 0, 0)$ must vanish, this means that eq. 10 simplifies to

$$\varepsilon = \frac{\partial \varepsilon(0, 0, 0)}{\partial \Delta S} \Delta S + \Delta\delta^s(C, T) \quad (11)$$

At first glance, the error ΔS of the structural ensemble depends on two factors, the correctness of the calculated structures and of the structural ensemble obtained. Even if the two conditions are sufficiently well fulfilled, it is a practical problem to select a minimum number of structures that can represent the experimental structural ensemble from the point of averaged chemical shifts. Therefore, ΔS also is a function of the usually arbitrarily chosen number N of structures that are either ordered according to their energies with the lowest energy assigned to structure s_1 or according to the probability.

Results

Prediction of chemical shifts in a test data set

Wang and Jardetzky (2002) prepared a data set of proteins where high resolution structures and heteronuclear chemical shift data were available for the development of a new method of secondary structure prediction. Here, we use a subset of 16 NMR structures for an analysis of the chemical shift predictions (Table 1) where structural bundles are deposited in PDB database. For obtaining an estimate of $\Delta\delta^s$ the chemical shifts were calculated with the program SHIFTX and SHIFTS for these structures and compared with the experimental data. The mean ε and the second moment σ^2 was calculated using eq. 7 and 9. The deviation of the predicted combined chemical shifts $\Delta\delta_{\text{comb}}$ of the backbone atoms from the experimental chemical shifts are shown in Fig. 1. The chemical shifts were calculated from the first structure in the data base (usually the lowest energy structure) and the total ensemble. In general the performance of SHIFTX is slightly better for all structures studied than that of SHIFTS, for the single structure as well as for the ensemble. The weighted mean over all 16

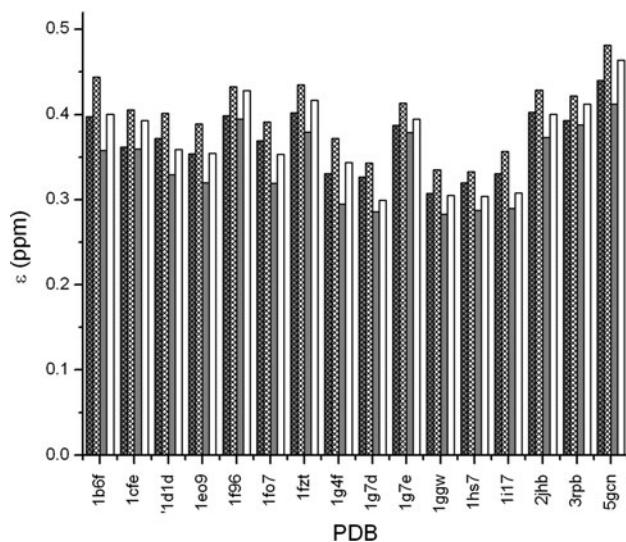


Fig. 1 Precision of chemical shift predictions. The weighted chemical shift deviations ε (eq. 7) are plotted for the structures listed in Table 1. The amino acid and atom specific weights used to calculate the combined chemical shifts $\Delta\delta_{\text{comb}}$ are defined by Schumann et al. (2007). The chemical shifts were calculated from the lowest energy structure with SHIFTX (white bars) and SHIFTS (dotted white bars) or from the available structural ensemble with SHIFTX (gray bars) and SHIFTS (dotted gray bars). Only those atoms were considered that could be predicted by both methods

proteins drops from 0.63 to 0.58 ppm for SHIFTX, and from 0.66 to 0.61 ppm for SHIFTS. The non-weighted average for all atoms (Tables 3, 4) drops from 0.66 to 0.61 ppm for SHIFTS and from 0.63 to 0.58 ppm for SHIFTX. The same trend, that SHIFTX gives a more correct prediction than SHIFTS, is observed for the majority of the predicted chemical shifts of groups of atoms (Tables 3, 4). Especially the predicted chemical shifts of the backbone resonances are more precise for SHIFTX.

Effect of ensemble size and quality

Ensemble size

In line with the fact that the experimentally observed chemical shifts are ensemble averages, the chemical shift predictions calculated as averages over the structural ensemble, the prediction of the chemical shifts from the ensembles is always more correct. This is clearly seen for the weighted shifts depicted in Fig. 1 where the chemical shift prediction is more accurate for all proteins when ensembles are used. Also for the ensembles the predictions by SHIFTX are again always better than those of SHIFTS. For individual types of atoms the ensemble prediction is always more precise than that obtained from the lowest energy structure independent of the prediction method used (Tables 3, 4). However, when applying the t-test to the data, the

improvement is very significant when all atoms or all backbone or all side chain atoms are used (Tables 3, 4) but it is not always significant with $p < 0.05$, essentially because sometimes the number of events in the data base is too small.

The data base used contains only relatively small structural ensembles, usually of the order of 20 structures. A complete description of a real structural ensemble would most probably require a much larger number of structures. Therefore, we calculated large ensembles of 2,000 structures each for wildtype HPr and a mutant HPr(H15A) from *S. aureus* (Maurer et al. 2004; Munte et al. to be published). The number of experimental restraints used for the simulated annealing (SA) is given in Table 2. Structures that obviously did not converge and therefore showed large violations of the experimental restraints (see “Materials and methods”) were removed before analysis. The remaining structures were ordered according to their total energies (not including the pseudo energies from experimental restraints) and the weighted cumulative shift difference $\varepsilon(N)$ (with s_1 the lowest energy structure) was plotted for the backbone atoms as well as for all atoms (more precisely for all atoms with assigned chemical shifts). Figure 2 shows clearly that $\varepsilon(N)$ first decreases substantially for the two proteins and shows an asymptotical behavior. The shape of the function is virtually independent on the prediction method used; however, the magnitude of the effect strongly depends on the atoms selected: the lowest values are obtained for the side chain atoms, the highest values for the main chain atoms, and intermediate values for the weighted average of all atoms (constant C in Table 5). The data can rather well be fitted by a lognormal distribution with an additional offset.

Refinement of the obtained structures in explicit water in general leads to an improved quality of the structures and possibly also to a change of $\varepsilon(N)$. Therefore, all the 2,000 structures were subjected to a water refinement and ε was recalculated. When the refined structures are given as input the prediction error decreases significantly at the same ensemble size (Fig. 3). However, for large ensembles the asymptotic value differs only slightly by a few percent (Table 5).

The largest differences are observed for small ensemble sizes (Fig. 3). Here, water refinement leads to much smaller initial values for the prediction errors. The data can sufficiently well be fitted by the lognormal distribution; however at very small sample sizes a substructure is clearly observable. Without water refinement about 18 structures are necessary for coming close to the asymptotic value, after water refinement only 10 structures are required.

For the structures contained in the experimental data base (Table 1) the same analysis leads to analogous results when we assume that the structures are ordered according to their energies (a fact not known). In general, the ensemble

Table 3 Average performance of chemical shift prediction for specific atoms using SHIFTS

Atom type	ε /ppm lowest energy structures	ε /ppm structural ensembles	Number of atoms	$\Delta\varepsilon/\varepsilon^a$	p value ^b
H ^N	0.53	0.50	3,378	-0.06	<0.07
N	3.50	3.45	2,954	-0.01	<0.22
H ^{α}	0.31	0.28	3,814	-0.10	<0.001
C ^{α}	1.22	1.06	3,344	-0.13	<0.001
C	1.61	1.48	2,171	-0.08	<0.03
H ^{β} (methylene)	0.25	0.23	4,766	-0.08	<0.002
H ^{β} (methyl)	0.18	0.16	222	-0.11	<0.227
C ^{β} (methylene)	1.29	1.10	1,954	-0.15	<0.001
C ^{β} (methyl)	1.19	1.03	188	-0.13	<0.15
H ^{γ} (methylene)	0.25	0.21	2,464	-0.16	<0.001
H ^{γ} (methyl)	0.23	0.21	572	-0.09	<0.052
H ^{δ} (methylene)	0.21	0.20	934	-0.05	<0.127
H ^{δ} (methyl)	0.19	0.17	542	-0.11	<0.08
H ^{δ} (aromatic)	0.23	0.21	372	-0.09	<0.121
H ^{δ} (amide)	0.59	0.56	266	-0.05	<0.294
H ^{ϵ} (methylene)	0.17	0.15	454	-0.88	<0.138
H ^{ϵ} (methyl)	0.28	0.21	56	-0.25	<0.20
H ^{ϵ} (aromatic)	0.31	0.30	370	-0.03	<0.348
H ^{η} (aromatic)	0.42	0.39	32	-0.07	<0.5
H ^{ζ} (aromatic)	0.24	0.23	154	-0.04	<0.30
Average error					
All atoms	0.77	0.70	29,259	-0.09	<0.009
Weighted average ^c					
All atoms	0.38	0.36	29,259	-0.05	<0.009
All backbone atoms	0.77	0.70	15,661	-0.09	<0.02
All side chain atoms	0.38	0.36	13,598	-0.05	<0.001

The average chemical shift differences ε_i were calculated using the Hamming distance (eq. 7) and a weighting factor $w_i = 1$, that is ε was calculated for groups of n atoms as average of the M proteins with N_k ensemble members listed in Table 1. The error ε is defined as $\varepsilon = \langle \varepsilon_i \rangle = \frac{1}{M} \sum_{k=1}^M \frac{1}{N_k} \sum_{j=1}^{N_k} \frac{1}{n} \sum_{i=1}^n |(\delta_i^p(s_j) - \delta_i^e)|$. For stereospecifically not assigned atoms such as methylene protons the chemical shifts of the corresponding protons were averaged before calculating the difference

^a $(\varepsilon(\text{ensemble}) - \varepsilon(\text{lowest energy}))/\varepsilon(\text{lowest energy})$

^b Probability p that the two distributions have the same means are based on t -statistics. Values <0.05 are highlighted in bold letters

^c Weighted average of all atoms as defined in Fig. 1. Only atoms that can be treated by SHIFTS as well as SHIFTX were considered

gives better performance of the chemical shift prediction by SHIFTX (Fig. 4) and SHIFTS (data not shown). In most cases a minimum value seems to be reached when about 18 structures are used for the calculations.

Energy distributions and their impact on the chemical shift prediction

The ensembles obtained for wildtype HPr and its mutant by simulated annealing followed by refinement in explicit water should not be determined mainly by the experimental pseudo energies but by the physical model itself. Figure 5 shows the energies with and without inclusion of the

pseudo energies resulting from the restraint violation. The energies were ordered according to their magnitude for the 2,000 structures. It is obvious that the restraint violations only contribute little to the energy. The probability distributions of the energies are represented in Fig. 5c and d and can be fitted in a good approximation by a Gaussian.

The quality of the chemical shift prediction of smaller sets of structures may depend on the total energy of the structures under consideration. This was tested for wildtype and mutant HPr in two different ways: either (1) the total ensemble (ordered according to the total energy) was divided in classes of energies that were multiples of the standard deviation σ starting from the mean value or

Table 4 Average performance of chemical shift prediction for specific atoms using SHIFTX

Atom type	ε /ppm lowest energy structures	ε /ppm structural ensembles	Number of atoms	$\Delta\varepsilon/\varepsilon^a$	p value ^b
H ^N	0.52	0.48	3,350	−0.08	<0.006
N	2.53	2.38	3,278	−0.02	<0.01
H ^{α}	0.28	0.24	3,218	−0.14	<0.0001
C ^{α}	1.04	0.95	3,500	−0.08	<0.001
C	1.28	1.22	2,266	−0.05	<0.121
H ^{β} (methylene)	0.23	0.21	4,718	−0.09	<0.008
H ^{β} (methyl)	0.20	0.18	222	−0.10	<0.271
C ^{β} (methylene)	1.13	1.01	2,095	−0.12	<0.007
C ^{β} (methyl)	0.99	0.84	188	−0.15	<0.187
H ^{γ} (methylene)	0.23	0.21	2,452	−0.09	<0.026
H ^{γ} (methyl)	0.22	0.20	572	−0.09	<0.05
H ^{δ} (methylene)	0.21	0.19	930	−0.10	<0.05
H ^{δ} (methyl)	0.26	0.22	734	−0.15	<0.004
H ^{δ} (amide)	0.63	0.55	254	−0.13	<0.171
H ^{ϵ} (methylene)	0.15	0.14	450	−0.07	<0.22
H ^{ϵ} (methyl)	0.39	0.38	54	−0.03	<0.32
H ^{ϵ} (amide)	0.42	0.40	320	−0.05	<0.27
H ^{δ} (aromatic)	1.23	1.22	72	−0.01	<0.45
H ^{ϵ} (aromatic)	1.04	1.03	62	−0.01	<0.47
Average error					
All atoms	0.63	0.58	28,735	−0.08	<0.001
Weighted average ^c					
All atoms	0.37	0.34	28,735	−0.08	<0.001
All backbone atoms	0.70	0.64	15,612	−0.09	<0.007
All side chain atoms	0.35	0.33	13,123	−0.06	<0.001

The average chemical shift differences ε_i were calculated using the Hamming distance (eq. 7) and a weighting factor $w_i = 1$, that is ε was calculated for groups of n atoms as average of the M proteins with N_k ensemble members listed in Table 1. The error ε is defined as $\varepsilon = \langle \varepsilon_i \rangle = \frac{1}{M} \sum_{k=1}^M \frac{1}{N_k} \sum_{j=1}^{N_k} \frac{1}{n} \sum_{i=1}^n |(\delta_i^p(s_j) - \delta_i^e)|$. For stereospecifically not assigned atoms such as methylene protons the chemical shifts of the corresponding protons were averaged before calculating the difference

^a $(\varepsilon(\text{ensemble}) - \varepsilon(\text{lowest energy}))/\varepsilon(\text{lowest energy})$

^b Probability p that the two distributions have the same means are based on t -statistics. Values <0.05 are highlighted in bold letters

^c Weighted average of all atoms as defined in Fig. 1. Only atoms that can be treated by SHIFTS as well as SHIFTX were considered

(2) classes containing the same number of 20 structures (Fig. 6) ordered according to their energies. In the first case, a second order polynomial was required to fit the data; in the second case a first order polynomial was sufficient.

Discussion

Multiconformational ensembles and conformational averaging of chemical shifts

From first principles of thermodynamics it is clear that protein structures in solution form a large ensemble of

multiple conformational states. A complete description of a protein would require the knowledge of all coexisting structures; even the knowledge of energetically unfavorable states that are only weakly populated may be important since functional excited states or folding intermediates may be contained in the higher energy part of the energy landscape (Kalbitzer et al. 2009). A complete representation of all states is practically not possible because of the extremely large number of states. Even when one restricts to the ground state ensemble, only the representation of a limited selection of structures is feasible. However, it is not clear, which structures should be selected as being representative for the total ensemble and how many structures must be included for a faithful representation of the ensemble.

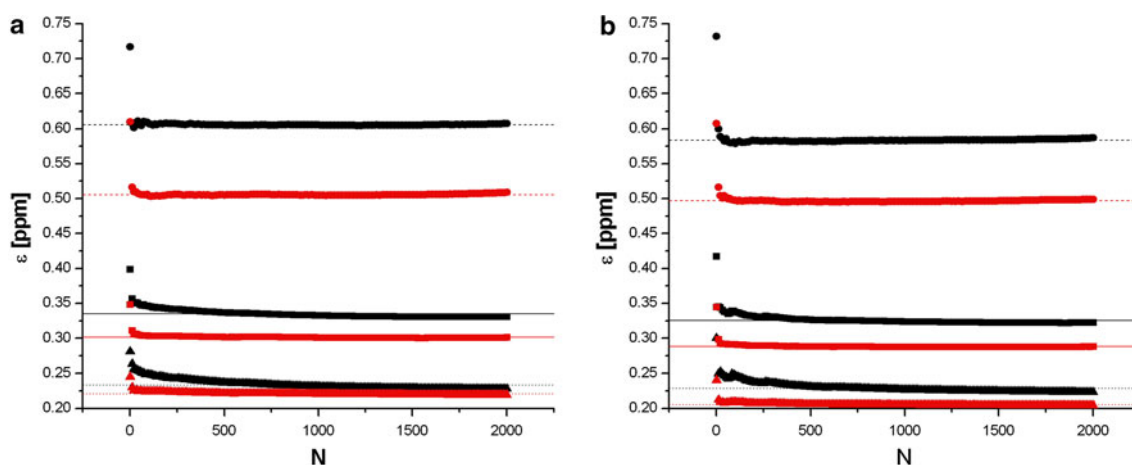


Fig. 2 Dependence of the chemical shift error ε on the size of the structural ensemble before water refinement. The structures used were obtained by restrained simulated annealing (see “Materials and methods”). The weighted mean error ε of the back bone atoms H^N , H^α , N, C^α , C (circle), side chain atoms (triangle) and all atoms (square) were plotted as a function of the size N of the ensemble using the Hamming distance (eq. 5). **a** HPr(wt) type using SHIFTS (black)

and HPr(wt) using SHIFTX (gray); **b** HPr(H15A) using SHIFTS (black), and HPr(H15A) using SHIFTX (gray). The data were fitted with a lognormal distribution $\varepsilon = \frac{1}{N\sigma\sqrt{2\pi}} e^{-\frac{(\ln N)^2}{2\sigma^2}} + C$ and the value of C for back bone atoms (dashed line), side chain atoms (dotted line), and all atoms (solid line) is represented by a straight line

Table 5 Minimum ensemble size and error offset

Ensemble	SHIFTS		SHIFTX	
	σ	C/ppm	σ	C/ppm
HPr(wt) SA				
All	0.026	0.33	0.019	0.30
Backbone	0.044	0.61	0.041	0.51
Sidechain	0.019	0.23	0.010	0.22
HPr(wt) WREF				
All	0.015	0.32	0.008	0.29
Backbone	0.036	0.58	0.013	0.49
Sidechain	0.008	0.23	0.006	0.21
HPr(H15A) SA				
All	0.037	0.32	0.022	0.29
Backbone	0.059	0.58	0.044	0.50
Sidechain	0.028	0.23	0.014	0.21
HPr(H15A) WREF				
All	0.011	0.32	0.005	0.28
Backbone	0.014	0.56	0.010	0.48
Sidechain	0.009	0.22	0.004	0.20

The mean error ε of the back bone atoms H^N , H^α , N, C^α , C, side chain atom and all atoms was calculated as a function of the size N of the ensemble using the Hamming distance (eq. 5) for wildtype HPr and mutant HPr(H15A) with SHIFTS and SHIFTX. The data (Figs. 2, 3) were fitted with a lognormal distribution with a constant offset C :

$$\varepsilon = \frac{1}{N\sigma\sqrt{2\pi}} e^{-\frac{(\ln N)^2}{2\sigma^2}} + C$$

Indeed, the definition of a faithful representation of the ensemble depends on the properties of the ensemble that should be represented. In biochemistry, it would often focus

on the explanation of functional properties. In the present context, the following general questions are important: (1) Can the quality of the chemical shift prediction be increased by using ensembles of structures in agreement with theory, (2) is the improvement of the prediction by using ensembles independent of the method used, (3) what is the minimum size of the ensemble required for optimum chemical shift prediction, and (4) can the chemical shift prediction be used to define the representative ensemble?

Quality of the chemical shift prediction

For testing and quantifying the quality of the chemical shift prediction we used a set of NMR structures (Table 1) previously designed by Wang and Jardetzky (2002). In general, from the two programs tested here SHIFTX performs somewhat better than SHIFTS, the weighted average error of all atoms of 0.340 ppm calculated with SHIFTX for the lowest energy structure is about 8.2% lower than 0.368 ppm calculated with SHIFTS (Tables 3, 4). The standard deviations of the errors ε are for all atom groups rather high (almost as large as the mean itself), indicating that either the structural quality varies much or that the parameterization is not optimal for all conditions found in the structures. From the data itself this cannot be decided but probably variations in the structural quality as well as the computational methods used for the structure calculation may be the dominant factor for these variations. The prediction error varies for the different atom types. For SHIFTX we found mean errors of 0.52, 0.28, 2.53, 1.04, and 1.28 ppm for the H^N , H^α , N, C^α , and C atoms,

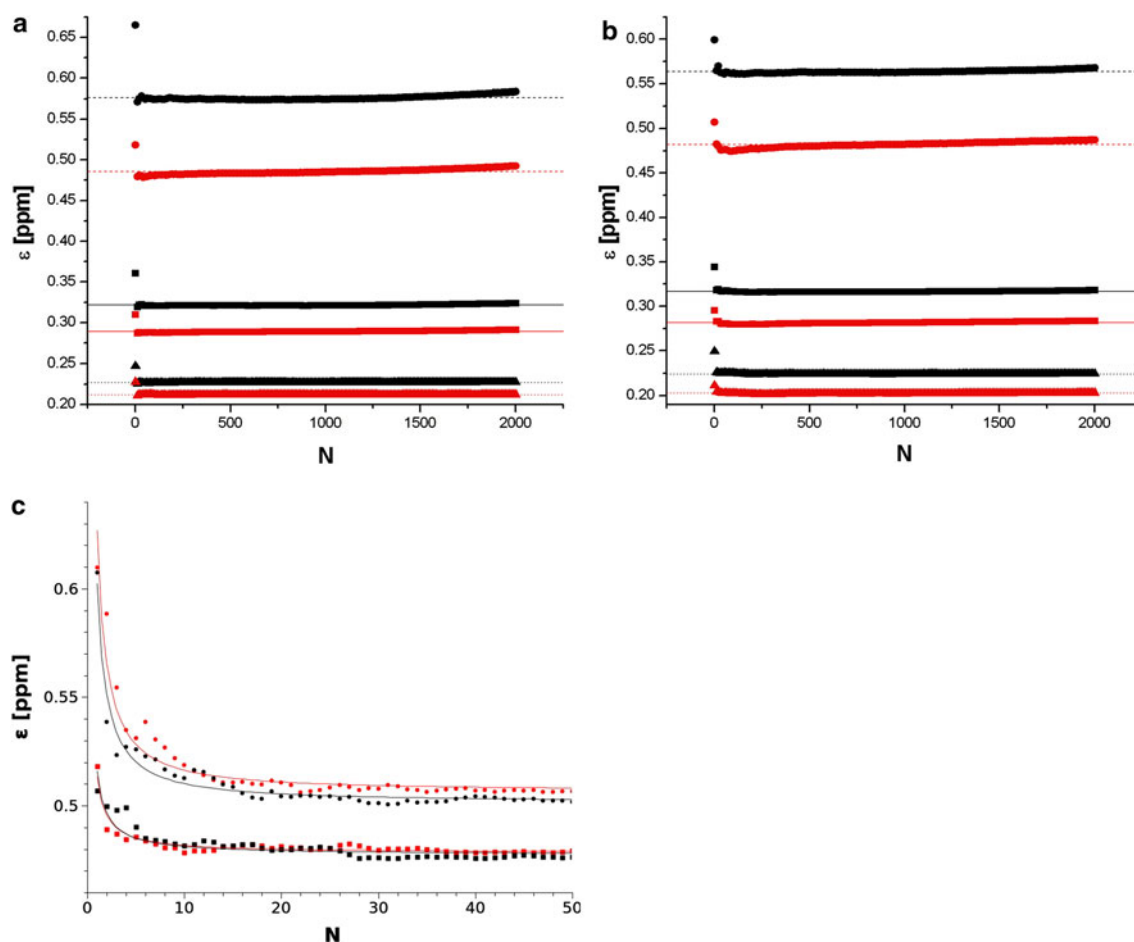


Fig. 3 Dependence of the chemical shift error ε on the ensemble size after water refinement. The structures used were obtained from the ensemble shown in Fig. 2 by refinement in explicit water (see “Materials and methods”). The weighted mean error ε of the back bone atoms H^N , H^α , N, C^α , C (circle), side chain atoms (triangle) and all atoms (square) were plotted as a function of the size N of the ensemble using the Hamming distance (eq. 5). **a** HPr(wt) using SHIFTS (black) and HPr(wt) using SHIFTX (gray); **b** HPr(H15A) using SHIFTS (black) and HPr(H15A) using SHIFTX (gray). The data were fitted as described in Fig. 2. The value of the parameter C

for back bone atoms (dashed line), side chain atoms (dotted line) and all atoms (solid line) is represented by a straight line. **c** The plot represents expansions of Figs. 2a, b and 3a, b. The mean weighted error $\varepsilon(N)$ of the backbone atoms are plotted as function of the size of the ensemble. Only the first 50 structures are shown. HPr(wt) before (circle) and after (square) water refinement, HPr(H15A) before (diamond) and after (triangle) water refinement. Solid line shows the corresponding lognormal fit. Only the data calculated with SHIFTX are depicted

respectively. Similar results were published most recently by Lehtivarjo et al. (2009) for a smaller data base of protein structures with 0.55 and 0.37 ppm for the H^N and H^α resonances. Similar results were also published earlier by Arun and Langmead (2004). For the side chain atoms the prediction error is usually much smaller (Tables 3, 4), one factor is the smaller chemical shift variations found here experimentally. However, this cannot be the only reason since the backbone prediction error is also larger than the side chain prediction error when it is calculated with the amino acid type and atom type specific weighting factors (Schumann et al. 2007) that correct for the chemical shift distribution of the atoms under consideration. The chemical shift prediction by SHIFTS and SHIFTX (and by all methods published so far) is still more than one order of

magnitude too inaccurate when it should be used for a direct assignment of resonances: here a precision of the order of the typical line width would be required that is about 0.01 ppm for proton and about 0.1 ppm for nitrogen resonances.

Prediction of chemical shifts from the ensemble or the lowest energy structure

In accordance with the fact that chemical shifts represent ensemble averages the use of ensembles generally improves the chemical shift prediction for all structures of the data base (Fig. 1) and for most of the atoms taken into account (Tables 3, 4). The weighted mean error of all atoms decreases by 8.8% when SHIFTX is used and 8.3%

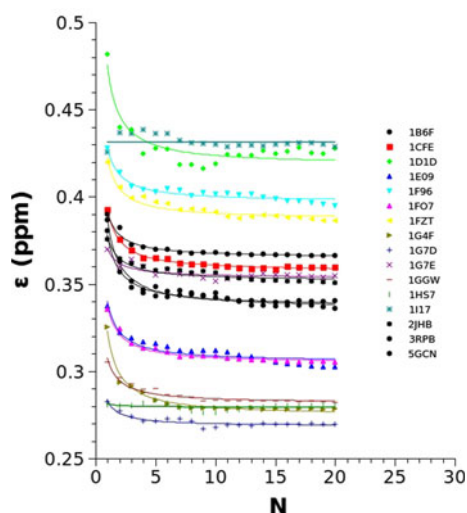


Fig. 4 Chemical shift error as a function of the ensemble size for the structural data base. The curves shown are fit curves as defined in Fig. 2 of the structures contained in the experimental data base (Table 1). The chemical shift predictions were performed with SHIFTX

when SHIFTS is used. The statistical significance of the observed decrease of the individual prediction errors is not equal for all groups of atoms considered, however atoms and groups of atoms with a large number of observations mostly have p values < 0.05 , although exceptions are existing (Tables 3, 4). It also is to be expected that an improvement by averaging over ensembles is largely independent on the prediction method used as it is shown here for SHIFTS and SHIFTX. In fact, for 4Dspot a similar result has been reported recently.

Minimum size of the ensemble required for chemical shift prediction

Usually 10–20 NMR structures are stored in the data base. Our result indicates that this is sufficient as far as the chemical shift prediction is concerned. The extensive simulation of the HPr structures shows that an asymptotic value is reached before water refinement when about 18 structures are averaged (Figs. 4, 5), after water refinement when about 10 structures are averaged (Fig. 4). Under this aspect the traditional way to deposit NMR structures can be considered as sufficient. When during the calculation of the structures those structures are removed that show large violations of the experimental restraints and thus have not converged properly, the error dependence of the chemical shift prediction on the size of the ensemble can be sufficiently well described by a lognormal distribution with a constant offset. Whereas the description with a lognormal distribution is purely empirical, the asymptotic behavior to a constant value can be expected from the chemical shift

averaging (eq. 3). However, when the structures with larger pseudo energies (badly converged structure calculations) are included, a continuous increase of the prediction error with the number N can be observed (data not shown).

Dependence of the prediction error on the energy distribution

From a general point of view it is surprising that a very small number of low energy structures can lead to a virtual optimum ensemble when the chemical shift prediction is concerned, although they clearly are not representative for the experimental ensemble but by definition only represent weakly populated states. The obtained energy distributions are shown in Fig. 5c and d for HPr that can be approximated well by a Gaussian. Structures having energy values less than -2σ relative to the mean energy are considered as lowest energy structures, energy values between $-\sigma$ and $+\sigma$ are considered as most probable structures and structures above $+2\sigma$ are considered as high energy structures. When the prediction error is plotted as a function of the deviation from the mean a minimum prediction error is detected close to the most probable ensemble at the mean energy (Fig. 6a) as to be expected from theory. However, the effect is rather small. In the intervals between $[\langle E \rangle - \sigma, \langle E \rangle]$ and $[\langle E \rangle, \langle E \rangle + \sigma]$ the number of calculated structures is much higher than in the other intervals. This may cause a bias on the data evaluation favoring the chemical shift prediction from the larger ensemble of structures. Therefore, the structures were sorted according to their energies and sets of identical size (20 structures) were taken for the chemical shift prediction. Here, a minimum of the error cannot be detected anymore but the prediction error is almost constant and can be well approximated by a straight line with a very small positive slope (Fig. 6b).

Prediction error

The experimentally observed prediction error $\Delta\delta^S(C, T)$ (eq. 11) is still rather large for all prediction methods C and for all atoms T considered and is especially large for the backbone atoms. It is much larger than the effects resulting from the ensemble averaging itself, according to the analysis of our structural data basis the ensemble effect ΔS is of the order of 10% of the $\Delta\delta^S(C, T)$ (Table 3, 4). Therefore, it is not surprising that we can only observe small effects from the selection of the structural ensemble (Fig. 6). In fact, using the same ensemble size, the quality of the ensemble prediction slightly decreases with the mean energy of the structural set. This bias may be caused by the parameterization procedure and the calculation methods itself that are optimized to obtain the lowest energy NMR structures but not a full structural ensemble in thermal

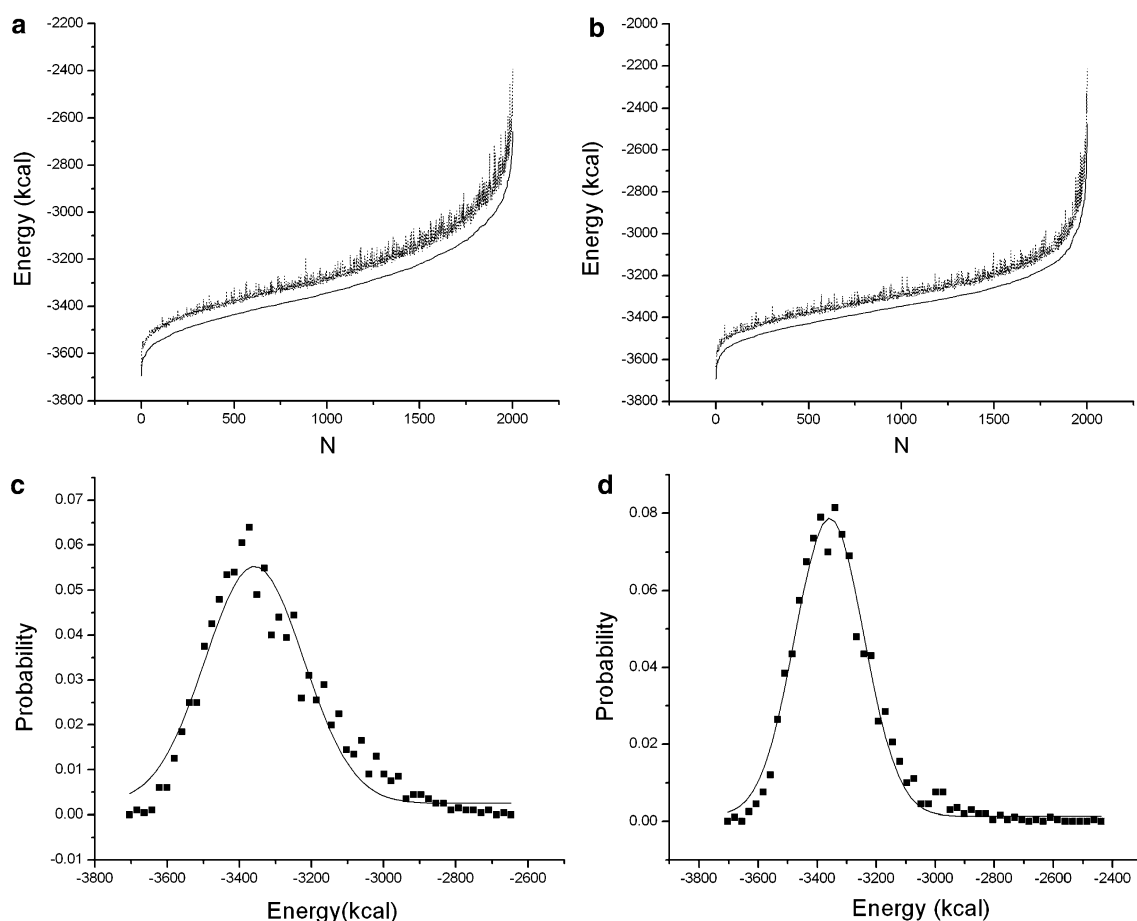


Fig. 5 Population of energy states. The structures of wildtype HPr (**a**) and mutant HPr (**b**) were ordered according to their total energy ($N = 1, \dots, 2000$). The total energy (*solid line*) and the sum of the total energy and the violation energy (*dotted line*) were plotted as a function of N . **c, d** The probability of each energy state was plotted as function of total energy (all energies excluding the experimental pseudo energies) and fitted with Gaussian function: **c** HPr(wt), mean

energy $\langle E \rangle = -3358.5$ kcal/mol, $\sigma = 274.0$ kcal/mol; and **d** HPr (H15A), $\langle E \rangle = -3358.5$ kcal/mol, $\sigma = 233.5$ kcal/mol. All structures were refined in explicit water, structures which have more than 5% NOE violation were removed before water refinement and structures having energies greater than 3σ were refined again with different random seed

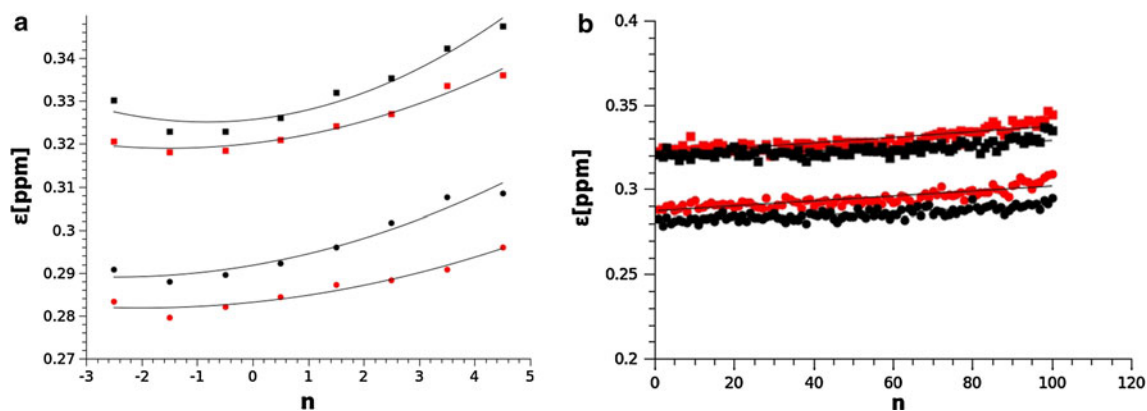


Fig. 6 Dependence of the prediction error on the energy of the ensemble. The prediction error ϵ for the backbone atoms is plotted as a function of the total energies E . The probability distribution of the energy E was divided **a** in classes E_n representing multiples of the standard deviation σ that is $\epsilon(E_n) = \langle \epsilon(E) \rangle$ with $E \in$

$(\langle E \rangle + (n-1)\sigma, \langle E \rangle + n\sigma)$ or **b** in classes $E_n (n = 0, \dots)$ containing 20 structures ordered according increasing values of $E - E_{\min}$. Squares, prediction with SHIFTS; circles, prediction with SHIFTX; black, HPr(wt); gray, HPr(H15A). The data were fitted with a polynomial of first or second order

equilibrium. In addition, the parameters of chemical shift predictions are usually optimized with respect to crystal structures that clearly do not represent the solution ensemble measured experimentally and thus may also introduce a bias in the prediction methods towards lowest energy structures and cause a contribution to the prediction error of chemical shifts $\Delta\delta^S$. In a most recent paper, that appeared during the preparation of this manuscript, Lehtivarjo et al. (2009) show that indeed better results can be obtained when MD-ensembles are used for parameterization.

Conclusion

There are a number of good reasons to use an ensemble of structures instead of a single energy minimized structure for chemical shift predictions as well as for functional analyses. The minimum energy structure calculated by restraint molecular dynamics, simulated annealing, and water refinement are actually only representing a subset of structures in the conformation space. The method is not aimed to describe the structural ensemble at a given temperature in thermal equilibrium. However, in the presence of experimental restraints the method creates ensembles that NMR spectroscopists assume to somewhat represent the conformational space at the experimental conditions where the NMR experiments were performed. Experimentally, chemical shift prediction can be improved significantly when a structural ensemble is used. An ensemble size of about 20 structures is sufficient, further increase of the size seems not to lead to better results. The conclusion primarily holds for the two tested, most popular prediction programs SHIFTX and SHIFTS but for theoretical reason most probably also applies for other prediction programs.

The calculation methods used in NMR spectroscopy are aimed to find the lowest energy structure with respect to the experimental restraints, the obtained ensemble is mainly determined by the reduction of the accessible conformational space, not by the MD-potentials. Nevertheless, properties of the true ensemble under the experimental conditions given are reflected in these restraints and lead to calculated ensembles that have some similarities to true thermodynamic ensembles. An interesting question not dealt with in this paper would be, under what conditions a perfect MD simulation of protein in explicit water at a given temperature would provide even better chemical shift predictions or if the inherent prediction error $\Delta\delta^S$ not optimized for ensembles would prevent such an effect.

Acknowledgments This work was supported by the Bundesministerium für Bildung und Forschung (BMBF), the Deutsche Forschungsgemeinschaft (DFG), the priority program 760 “Medical Chemistry: molecular ligand–receptor interaction” and the European Union (Extend-NMR, SPINE2).

References

- Arun K, Langmead CJ (2004) Large-scale testing of chemical shift prediction algorithms and improved machine learning-based approaches to shift prediction. Computational Systems Bioinformatics Conference (CSB'04) 712–713
- Brunger AT (2007) Version 1.2 of the crystallography and NMR system. Nat Protoc 2:2728–2733
- Brunger AT, Adams PD, Clore GM, DeLano WL, Gros P, Grosse-Kunstleve RW, Jiang JS, Kuszewski J, Nilges M, Pannu NS, Read RJ, Rice LM, Simonson T, Warren GL (1998) Crystallography & NMR System: a new software suite for macromolecular structure determination. Acta Crystallogr D Biol Crystallogr 54:905–921
- Geyer M, Schweins T, Herrmann C, Prisner T, Wittinghofer A, Kalbitzer HR (1996) Conformational transition in p21ras and its complexes with effector protein Raf-RBD and the GTPase activating protein GAP. Biochemistry 35:10308–10320
- Hahmann M, Maurer T, Lorenz M, Glaser W, Hengstenberg W, Kalbitzer HR (1998) Structural studies of the Histidine-Containing Phosphocarrier Protein (HPr) from *Enterococcus faecalis*. Eur J Biochem 252:51–58
- Iuga A, Spoerner M, Kalbitzer HR, Brunner E (2004) Solid-state ^{31}P NMR spectroscopy of microcrystals of the Ras protein and its effector loop mutants: comparison of solution and crystal structures. J Mol Biol 342:1033–1040
- Jia Z, Vandonselaar M, Hengstenberg W, Quail JW, Delbaere LTJ (1994) The 1.6 Å structure of the histidine containing phosphocarrier protein HPr from *Streptococcus faecalis*. J Mol Biol 236:1341–1355
- Jorgensen WL, Tirado-Rives J (1988) The OPLS force field for proteins. Energy minimizations for crystals of cyclic peptides and crambin. J Am Chem Soc 110:1657–1666
- Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML (1983) Comparison of simple potential functions for simulating liquid water. J Chem Phys 79:926–935
- Kalbitzer HR, Spoerner M, Ganser P, Hosza C, Kremer W (2009) Fundamental link between folding states and functional states of proteins. J Am Chem Soc 131:16714–16719
- Koradi R, Billeter M, Wuthrich K (1996) MOLMOL: a program for display and analysis of macromolecular structures. J Mol Graph 14:51–55
- Laskowski RA, Rullmann JA, MacArthur MW, Kaptein R, Thornton JM (1996) AQUA and PROCHECK-NMR: programs for checking the quality of protein structures solved by NMR. J Biomol NMR 8:477–486
- Lehtivarjo J, Hassinen T, Korhonen S-P, Peräkylä M, Laatikainen R (2009) 4D prediction of protein ^1H chemical shifts. J Biomol NMR 45:413–426
- Linge JP, Williams MA, Spronk CAEM, Bonvin AMJJ, Nilges M (2003) Refinement of protein structures in explicit solvent. Proteins Struct Funct Genet 50:496–506
- Maurer T, Meier S, Kachel N, Munte CE, Hasenbein S, Koch B, Hengstenberg W, Kalbitzer HR (2004) High-resolution structure of the histidine-containing phosphocarrier protein (HPr) from *Staphylococcus aureus* and characterization of its interaction with the bifunctional HPr kinase/phosphorylase. J Bacteriol 186:5906–5918
- Meiler J (2003) PROSHIFT: protein chemical shift prediction using artificial neural networks. J Biomol NMR 26:25–37
- Neal S, Nip AM, Zhang H, Wishart DS (2003) Rapid and accurate calculation of protein ^1H , ^{13}C and ^{15}N chemical shifts. J Biomol NMR 2:215–240
- Pai EF, Krengel U, Petsko GA, Goody RS, Kabsch W, Wittinghofer A (1990) Refined crystal structure of the triphosphate conformation

- of H-Ras p21 at 1.35 Å resolution: implications for the mechanism of GTP hydrolysis. *EMBO J* 9:2351–2359
- Perkins SJ, Johnson LN, Philipps DC, Dwek RA (1977) Conformational changes, dynamics and assignments in ^1H NMR studies of proteins using ring current calculations. Hen egg white lysozyme. *FEBS Lett* 82:17–22
- Schumann FH, Riepl H, Maurer T, Gronwald W, Neidig K-P, Kalbitzer HR (2007) Combined chemical shift changes and amino acid specific chemical shift mapping of protein-protein interactions. *J Biomol NMR* 39:275–289
- Shen Y, Bax A (2007) Protein backbone chemical shifts predicted from searching a database for torsion angle and sequence homology. *J Biomol NMR* 38:289–302
- Spoerner M, Herrmann C, Vetter IR, Kalbitzer HR, Wittinghofer A (2001) Dynamic properties of the Ras switch I region and its importance for binding to effectors. *Proc Natl Acad Sci* 98:4944–4949
- Stumber M, Geyer M, Graf R, Kalbitzer HR, Scheffzek K, Haeblerlen U (2002) Observation of slow dynamic exchange processes in Ras protein crystals by ^{31}P solid state NMR spectroscopy. *J Mol Biol* 323:899–907
- Wang Y (2004) Secondary structural effects on protein NMR chemical shifts. *J Biomol NMR* 30:233–244
- Wang Y, Jardetzky O (2002) Probability-based protein secondary structure identification using combined NMR chemical-shift data. *Protein Sci* 11:852–861
- Xu XP, Case DA (2001) Automated prediction of ^{15}N , $^{13}\text{C}\alpha$, $^{13}\text{C}\beta$ and $^{13}\text{C}'$ chemical shifts in proteins using a density functional database. *J Biomol NMR* 21:321–333