J Biomol NMR (2007) 38:269Ð287
DOI 10.1007/s10858-007-9161-y

Electronic Supplementary Material The online version of this article (doi: 10.1007/s10858-007-9161-y) contains supplementary material, which is available to authorized users.

E. Czinki A. G. Csa´sza´r (&)
Laboratory of Molecular Spectroscopy, Institute of Chemistry, Eo¨tvo¨s University, P. O. Box 32, Budapest 112, H-1518, Hungary
e-mail: csaszar@chem.elte.hu

^{13}C shifts and 30% of ^{15}N shifts are incorrectly referenced. As a consequence, Wishart and co-workers have developed RefDB, (Zhang et al. 2003) a secondary database of reference-corrected protein chemical shifts, based on primary data from the BMRB. In this paper chemical shifts were obtained exclusively from the RefDB database.

Since most of the results of this paper concern structure–ICS relations of proteins, it is important to point out why these relations are worth exploring, especially in view of the well-known fact that although ICSs incorporate a wealth of structural information, they have been playing almost no role in NMR structure refinement procedures (Szilágyi 1995). The usually quoted reason (Redfield and Dobson 1990) for not using chemical shifts in structural studies of proteins is that ICSs do not depend on well-defined pair-wise interactions, unlike the nuclear Overhauser effect (NOE) or the J-coupling constants. Therefore, determination of the 3-dimensional (3D) structure of proteins by NMR spectroscopy has been based on distance-type constraints from NOE(SY), Nuclear Overhauser effect spectroscopy experiments (Neuhaus and Williamson 1989). The disadvantage of using these types of constraints is that their number might not be sufficient to fix the structure of the protein, resulting in ambiguous results even for less flexible regions. Augmenting these studies by further information, perhaps derived from residue-dependent ICS(ϕ,ψ) surfaces, should thus be highly beneficial for the everyday practice of determining dependable protein structures.

Over the years, the number of entries of peptides and proteins has grown considerably both in the PDB and the BMRB databases. Nevertheless, reasonably accurate structural and NMR isotropic chemical shift data that are related to one another, i.e., the number of entries which can be used to better understand structure–chemical shift relations, is still relatively small (Beger and Bolton 1997; Wang 2004). Furthermore, most of the experimental data are constrained to small regions of the Ramachandran surface and there is considerable scatter both in the measured structural and ICS data. Consequently, the complicated dependence of ICSs on structural parameters makes the use and interpretation of experimental chemical shifts for analysis of the conformational space of proteins difficult if not impossible. Although rules have been constructed trying to describe structural contributions to chemical shifts by statistically investigating the related databases, (Spera and Bax 1991; Wishart et al. 1991, 1992; Gronenborn and Clore 1994; Dalgarno et al. 1983; Szilágyi and Jardetzky 1989; Ösapay and Case 1991; Ösapay and Case 1994) it has become clear that to obtain more or less complete and useful ICS(ϕ,ψ) surfaces one needs to utilize both experimental and computational information. The joint use of these two complementary sources of information offers the only real possibility to construct meaningful ICS(ϕ,ψ) surfaces, which can then be described as empirical surfaces.

Development of computers made the ab initio computation of structures and chemical shieldings easier, opening a new route to the analysis of the various contributions and an improved possibility to investigate structural effects, especially those not easily accessible to experiments. These ab initio studies, Sun et al. 2002; Czinki and Csaszar 2004 and references therein, involving relatively small peptide models, could reproduce and extend several findings of the analyses of the experimental data available for proteins. In the case of peptides the $^{13}C^{a}$, $^{1}H^{a}$, $^{13}C^{b}$, and to some extent the $^{13}C'$ chemical shifts are mainly affected by the local structure, in particular the ϕ and ψ dihedral angles, with much smaller contributions from other effects. The chemical shift of backbone ^{15}N is strongly influenced by the sequence, in particular the nature of the neighbouring residues. The shifts of carbonyl $^{13}C'$ of the ith residue is considerably affected by residue $i+1$ (Braun et al. 1994; Yao et al. 1997). Unlike on these nuclei, there is no significant effect of sequence on $^{13}C^{a}$ (and $^{13}C^{b}$) shifts (Yao et al. 1997; Iwadate et al. 1999). The only residue that had any possible effect is proline, but the reason behind this is the conformational effect of proline—it encourages β-sheet or poly-proline-II-like conformer types. Previous studies also suggested that aromatic ring-current effects are generally small for $^{13}C^{a}$ as well as for $^{13}C^{b}$ (Iwadate et al. 1999). The importance of sequence and H-bonding on carbonyl, and especially on ^{15}N shifts, and the small effect of these factors on $^{13}C^{a}$ and $^{13}C^{b}$ chemical shifts was shown in an extensive study by Xu and Case (Xu and Case 2001, 2002) on 1335 peptide models. These studies also proved that there is good correlation between secondary structures and chemical shifts for the α-helix and β-strand regions. Finally, the two most important works related to this study. Sun et al. (2002) determined ICS(ϕ,ψ) surfaces for all 20 naturally occurring amino acids, while the authors of this article recently reported a thorough investigation of the ICS(ϕ,ψ) surfaces of the peptide model For-Ala-NH_2 (Czinki and Csaszar 2004). To summarize previous experimental and ab initio studies, they indicated that there are certain nuclei, such as $^{13}C^{a}$ and $^{1}H^{a}$, where the local effect seems to dominate the ICSs, resulting in meaningful ICS(ϕ,ψ) surfaces, and other nuclei, where the local effect does not dominate and thus there is no hope in deriving useful ICS(ϕ,ψ) surfaces.

There have been several goals pursued resulting in this paper. (1) Establishing links between entries in the PDB and BMRB/RefDB databases is an important exercise which should be repeated time after time. Not forgetting about earlier attempts, especially that of Wang (2004) and Wishart et al. (2003), links have been determined between the PDB and RefDB databases as part of this study. These links are summarized in Table 1. (2) As emphasized above, attempts to determine residue-specific ICS(ϕ,ψ) surfaces for major nuclei of all naturally occurring α-amino acids serves several purposes. Therefore, we seeked to combine experimental and ab initio information in order to establish empirical

Table 1 The list of proteins employed in this study as experimental data. The PDB code (+ chain ID), with the resolution of the structure determination, and the corresponding BMRB accession code are provided

PDB code + chain ID	BMRB code	Resolution (Å)
5PTI0	46	1.0
4ICB0	390	1.6
1EY0A	497	1.6
1IOB0	1061	2.0
2OVO0	1375	1.5
1HOE0	1642	2.0
1RBV0	1657	1.8
2TRXA	1812	1.68
1FGLA	2208	1.8
1CM2A	2371	1.80
1HCB0	4022	1.6
3RN30	4031	1.45
1RSY0	4039	1.9
1SNC0	4053	1.65
1A6K0	4061	1.1
1HFC0	4064	1.56
1CT2I	4068	1.65
2FKE0	4077	1.72
1VAPB	4078	1.6
1FIL0	4082	2.0
1L3KA	4084	1.10
2BOPA	4087	1.7
1IARA	4094	2.30
1LFO0	4098	2.3
1MHO0	4099	2.0
1CEX0	4101	1.0
1A7TB	4102	1.85
1EMVA	4115	1.70
1U9B0	4132	2.0
4AKEB	4152	2.2
1DSZB	4153	1.70
1EPFC	4162	1.85
1CBS0	4186	1.8
1EZ3A	4198	1.90
3DFR0	4262	1.7
1MXEB	4270	1.70
1GZZB	4278	2.3
1EMVB	4293	1.70
1MJC0	4296	2.0
1B79B	4297	2.30
1HKA0	4299	1.50
1SMTA	4306	2.2
1AIL0	4317	1.9
3SSI0	4331	2.3
1EKGA	4342	1.80
1EMVB	4352	1.70

Table 1 continued

PDB code + chain ID	BMRB code	Resolution (Å)
1B72B	4357	2.35
1HY7B	4364	1.50
1ONC0	4371	1.7
1EDHB	4380	2.0
1DFUP	4395	1.80
1F2LD	4397	2.00
1AVSB	4401	1.75
1CLVI	4490	2.00
1B72B	4572	2.35
1MOLB	4633	1.7
1FD3D	4642	1.35
830CB	4679	1.6
1XNC0	4704	1.6
1I4MA	4736	2.00
1LOPA	4765	1.8
1DDWA	4766	1.70
1OSPO	4773	1.95
1PTF0	4774	1.6
1GO4C	4775	2.05
1BYLA	4785	2.30
1HFZC	4811	2.3
1HAVA	4836	2.0
1HC9B	4838	1.8
1G6SA	4848	1.50
1CM9B	4852	2.10
1G6TA	4854	1.6
2A0B0	4857	1.57
1CT2I	4865	1.65
3TGII	4868	1.8
1A1VA	4885	2.2
1FMK0	4888	1.5
1J8RA	4897	1.80
1JOCA	4898	2.20
1A3K0	4909	2.1
1CM9B	4914	2.10
1RZL0	4917	1.6
1CIQA	4926	2.2
1BWOB	4932	2.10
1QJ8A	4936	1.9
1QIOA	4943	1.20
1JF8A	4944	1.12
1IIBB	4955	1.8
1BJAA	4957	2.19
1EYHA	4959	1.56
1BNJB	4964	2.1
2CI2I	4974	2.0
11BGB	4980	1.90
1CMXC	4983	2.25

Table 1 continued

PDB code + chain ID	BMRB code	Resolution (Å)
1F80E	4989	2.30
1DTLA	4994	2.15
1F9FC	4998	1.90
1F80E	5004	2.30
1PUC0	5009	1.95
1HC9A	5025	1.8
1GNUA	5058	1.75
1KJTA	5064	2.00
1NCX0	5071	1.8
1I71A	5075	1.45
1B560	5083	2.05
1KX9B	5094	1.65
1F94A	5097	0.97
1G4YB	5102	1.6
1IWUA	5124	1.40
1IP2A	5125	1.80
1KJTA	5128	2.00
109M0	5158	1.83
1GCPB	5179	2.10
1M5EA	5182	1.46
2VPFH	5186	1.93
1EK8A	5190	2.30
1MHO0	5206	2.0
1IG5A	5207	1.50
1BPB0	5208	2.3
1JN3A	5209	2.35
1D4TA	5211	1.1
1D4TA	5212	1.1
1KZNA	5218	2.30
1I1JB	5220	1.39
2END0	5244	1.45
1DCDA	5249	2.00
1GO4C	5299	2.05
1FDQB	5320	2.10
1EZGB	5323	1.40
1REC0	5332	1.9
1B72B	5349	2.35
1H4HC	5352	1.9
1AM10	5355	2.0
1HULA	5373	2.4
1MHO0	5377	2.0
1EJMF	5381	1.85
1DTLA	5386	2.15
3DFR0	5396	1.7
1AO3A	5456	2.2
1D8CA	5471	2.00
1G7FA	5474	1.80
1PCS0	5475	2.15

Table 1 continued

PDB code + chain ID	BMRB code	Resolution (Å)
1EB0A	5484	1.85
1H7MA	5485	1.96
1RGEB	5492	1.15
1FHNA	5508	1.75
1PPFI	5519	1.8
1C7FB	5540	2.00
1MHO0	5544	2.0
1F4PA	5571	1.30
1L0SC	5572	2.30
1L0SB	5573	2.30
1CRB0	5579	2.1
1ICFJ	5583	2.00
1C09A	5600	1.6
1BRFA	5601	0.95
1MN8D	5623	1.00
1P42B	5627	2.00
1MFLA	5631	1.88
1LP1B	5656	2.30
1KWIA	5688	2.19
1BMBA	5693	1.80
1EJHC	5712	2.2
4AKEB	5720	2.2
1SEMB	5729	2.0
1AVSB	5738	1.75
1RX20	5740	1.8
1RX40	5741	2.2
1E4VA	5746	1.85
1N0SA	5756	2.00
1NAT0	5899	2.45
1AZPA	5908	1.6
1BF4A	5909	1.6
1BF4A	5910	1.6
1TIPA	5935	2.2

ICS(u,w) surfaces. (3) This study should help understanding whethera-alanine, the favourite residue of modelers, could be considered as a good structural model for most naturally occurringa-amino acids. Not unrelated to this, one wonders about the range of validity of ICS computations on small peptide models based on alanine. We must recall here that the alanine residue has been used as a building block for computational models (Sun et al. 2002; Czinki and Csaszar 2004) as it offers several advantages: (a) ease of model building; (b) no side-chain effects are introduced; (c) simple handling of neighboring effect(s); and (d) even larger models are amenable to relatively high-quality computational techniques. (4) Following our previous ab initio studies, (Czinki and Csaszar 2004; Perczel and Csaszar 2000, 2001, 2002;

Perczel et al. 2003) we wanted to investigate whether empirical ICS(u,w) surfaces could be used to support the utility of ICS–ICS correlation plots in the elucidation of the conformational characteristics of proteins.

Computational details

Electronic structure computations

For exploring the dependence of chemical shifts on backbone dihedral angles u and w, we performed electronic structure computations on three simple alanine-containing peptide models, For–(Ala)$_n$–NH$_2$, with n = 1, 3, and 5. In the rest of the paper the peptide models For–(L-Ala)$_n$–NH$_2$ with n = 1, 3, and 5, will be called mono-, tri-, and penta-peptide models, respectively. For obtaining ICS(u,w) surfaces, constrained geometry optimizations have been performed with u (w) fixed in a systematic manner but relaxing all other structural parameters.

In case of the smallest model, For–Ala–NH$_2$, a dense grid of 10° was used for the u and w angles in the whole Ramachandran space, [–180°, +180°]. In case of the larger models, the ICSs of the nuclei of only the middle residue as a function of its own backbone dihedral angles were investigated. For the larger For–(Ala)$_3$–NH$_2$ and For–(Ala)$_5$–NH$_2$ models, two choices have been made. For the first set (Model 1), the backbone dihedral angles of the edge residues were kept fixed at (–60°, –50°), corresponding approximately to an a-helix arrangement, and the middle dihedral angles were rotated by 30° in the whole Ramachandran space, [–180°, +180°]. During the second set of calculations (Model 2), the backbone dihedral angles of the edge residues were fixed at values characteristic of the given region, and then the u and w angles of the middle residue were changed by 10° within the range of angles characteristic of that secondary structure type. The following characteristic dihedral angle values have been chosen for these computations for fixing all residues but the middle one: a-helix: (u,w) = (–60°, –40°), b-strand: (u,w) = (–120°, 120°), polyproline II: (u,w) = (–75°, 145°), and left-handed a-helix: (u,w) = (60°, 30°).

Since the first description of conformations of peptides by the two backbone torsional angles, (Ramachandran and Sasikekharan 1968) several revisions of the Ramachandran map and the secondary structure regions have been published (Kleywegt and Jones 1996; Vlasov et al. 2001; Hovmoller et al. 2002; Ho et al. 2003). A few remarks are in order. (1) There are almost always amino acid residues in peptides and proteins which fall into a certain region based on their u and w angles but are actually not part of a sequence characterized by that secondary structure type. A thorough analysis of a large set of PDB entries revealed that ÔÔnearly one third of all amino acids in random coil have torsional angles in the a-helical regionÕÕ (Hovmoller et al. 2002). (2) Random coil residues can be found in the b-strand region, as well. Moreover, the b-strand region should be divided according to the structure of the sheet, e.g., parallel, anti-parallel, etc. (3) Each amino acid has slightly different Ramachandran plots. Consequently, selecting regions representing all amino acids must involve certain compromises. One must take into account the average values available for the different residues, the regions should be large enough to have a statistically significant number of grid points, while at the same time it is advantageous to avoid inclusion of too many ÔforeignÕ points at the edges of the region. After careful consideration, the following regions have been selected for the computations: a-helix: $-90° \le u \le -40°$ and $-60° \le w \le -10°$; b-strand: $-140° \le u \le -100°$ and $110° \le w \le 150°$; polyproline II: $-90° \le u \le -60°$ and $120° \le w \le 160°$; left-handed a-helix (aD): $40° \le u \le 80°$ and $10° \le w \le 60°$.

The constrained geometry optimizations were performed at the density-functional theory DFT(B3-LYP)/6–31+G* level, using the program system Gaussian98 (Frisch et al. 1998) and Gaussian03 (Frisch et al. 2004). Since the importance of employing relatively large basis sets for NMR chemical shielding calculations is well established, (Helgaker et al. 1999) we used a triple-zeta plus double polarization (TZ2P) (Schafer et al. 1994) basis set that is especially suited for NMR shielding calculations. We computed the NMR shielding tensors for all nuclei in our models using the B3-LYP (Becke3–Lee–Yang–Parr) (Becke 1988; Lee et al. 1988) DFT functional with the Gauge-Including Atomic Orbital (GIAO) method, (Ditchfield 1974; Wolinski et al. 1990) as implemented in Gaussian98 and Gaussian03.

Given the large number of reference structures and chemical shift data, automation of data handling becomes mandatory. Therefore, Python (http://www.python.org) and (http://www.perl.org) scripts were developed for input generation, data handling, output testing, and data extraction. The large amount of data extracted was entered into a MySQL (http://www.mysql.com) database with the aid of Python-MySQL and Perl-MySQL interfaces (http://www.cpan.org/modules/).

Handling the PDB and RefDB entries

A Perl script was developed for the automatic handling of the PDB and RefDB ßat Þles. The script allowed automatic extraction of the desired information, performed the necessary computations, like computing the backbone u and w angles from the Cartesian coordinates, and executed the linking of the two datasets. The matching pairs, the links

between the two databases, found are given in Table 1. The database assembled contains 175 proteins and 16,341 residues and includes data for all common amino acids though in a decidedly uneven manner.

Finding the corresponding PDB entry for each RefDB protein is not a trivial task (Zhang et al. 2003; Wang 2004; redpoll.pharmacy.ualberta.ca/RefDB/bmrpdb.html). The starting pairs of this study were those obtained by Wishart et al. (2003; redpoll.pharmacy.ualberta.ca/RefDB/bmrpdb.html) An important restraint imposed here was that only those proteins were considered whose X-ray structure determination claimed a resolution better than 2.5 Å. Inspection, involving description of the proteins, their biological and chemical environments, and their properties, of questionable or multiple pairs in the original pairs was performed resulting in changes of a few entries. A particular problem found was that the numbering of the sequence of a protein is sometimes different in the two databases. Ideally, in these cases an entry exists in the BMRB/RefDB files, entitled _Residue_author_seq_code, mapping one numbering scheme into the other. Since in several cases such entries did not exist in the BMRB/RefDB for cases where the two numbering schemes were actually different, we implemented into our Perl script a procedure mapping the two numbering schemes into one another.

Surface fitting

For comparing the experimental ICS dataset with the computed one, construction of an interpolating surface is necessary. In our preceding study, (Czinki and Csaszar 2004) after probing many mathematical functions, we have shown the adequacy of a 10th-order cosine expansion for describing the complex, structured $ICS(\varphi,\psi)$ surfaces of the Ala-containing models. A natural variant of the surface study of Czinki and Csaszar 2004 is the comparison of the complex computed surface with experimental $ICS(\varphi,\psi)$ data for the most important secondary structure regions. We constructed precise models of each secondary structure type investigated (Model 2). A detailed discussion of the Model 2 surfaces is given below. It is important to emphasize that the Model 1 surfaces give the same results though exhibit minor deviations from their Model 2 counterparts.

After performing the ab initio chemical shielding computations for the peptide models specific to each secondary structure region, one has chemical shift values, for the middle residue in the cases of For-(Ala)$_3$-NH$_2$ and For-(L-Ala)$_3$-NH$_2$, at all the computed points. Within any of the four regions considered, the $ICS(\varphi,\psi)$ surfaces are not as complex as the chemical shift surfaces corresponding to the whole Ramachandran map developed earlier (Czinki and Csaszar 2004). For these smaller regions, a simple fitting function is appropriate to describe the dependence of the ICSs on the φ and ψ backbone dihedral angles. Consequently, the Fourier series

$$\begin{aligned}ICS(\varphi,\psi) = a &+ \sum_{n=1}^{2}\left[b_n\cos\left(n\frac{\varphi+\pi}{2}\right) + c_n\cos\left(n\frac{\psi+\pi}{2}\right)\right] \\ &+ \sum_{n=1}^{2}\left[d_n\sin\left(n\frac{\varphi+\pi}{2}\right) + e_n\cos\left(n\frac{\psi+\pi}{2}\right)\right] \\ &+ f\cos\left(\frac{\varphi+\pi}{2}\right)\cos\left(\frac{\psi+\pi}{2}\right) + g\cos\left(\frac{\varphi+\pi}{2}\right)\sin\left(\frac{\psi+\pi}{2}\right) \\ &+ h\sin\left(\frac{\varphi+\pi}{2}\right)\cos\left(\frac{\psi+\pi}{2}\right) + i\sin\left(\frac{\varphi+\pi}{2}\right)\sin\left(\frac{\psi+\pi}{2}\right)\end{aligned}$$

was used for fitting. As to the statistics of the fit, the $ICS(\varphi,\psi)$ surface of the $^{13}C^{\alpha}$ nucleus, for example, is obtained with r^2 fit values better than 0.998 for all regions of the tri- and penta-peptide models. Figure 1 shows the fitted surfaces of the $^{\alpha}C$ and H^{α} nuclei in case of α-helix and β-strand regions. Table 1 of the Supplementary Material contains the fitting coefficients of the above function for the four secondary structure regions of the model For-(Ala)$_3$-NH$_2$.

Filtering of the experimental ICS data

After the availability of the fitted theoretical surfaces for each of the four regions considered, one is ready to compare them to experimental data available in that region. It is important to emphasize that the uncertainties of the experimental structural and ICS data are mostly unknown. As previous studies clearly show, (Spera and Bax 1991; Beger and Bolton 1997; Wang 2004; Dalgarno et al. 1983; Szilagyi and Jardetzky 1989; Ösapay and Case 1991, 1994; Sun et al. 2002; Czinki and Csaszar 2004; Iwadate et al. 1999; Czinki et al. 2003; Perczel and Csaszar 2000, 2001, 2002; Perczel et al. 2003) the backbone torsional effect is the main one determining the chemical shifts of the nuclei C^{α} and H^{α}. However, there exist further effects, whose importance needs to be investigated in "problematic" cases. Previous studies, (Spera and Bax 1991; Beger and Bolton 1997; Iwadate et al. 1999) whose aim was to investigate the clear trend between ICS data and φ and ψ angles, excluded some problematic residues. Having data from hundreds of proteins, the visual inspection of the problematic residues is not feasible. Therefore, an automated filtering procedure was needed, and the computed ICS surfaces served as guides in this respect (vide infra). A further issue concerns the absolute values of the shifts. As with other quantum chemical results for spectroscopic parameters obtained from lower levels of electronic structure

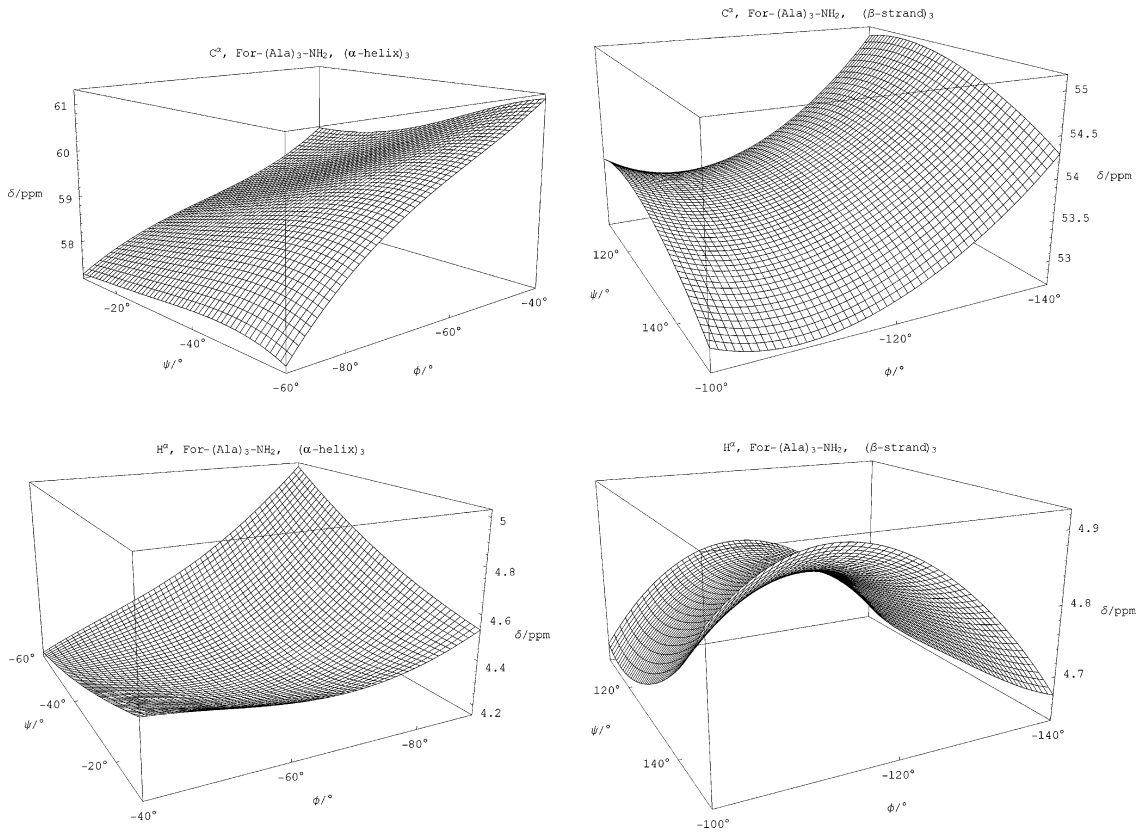

Fig. 1 Fitted, computed surfaces of the For–(Ala)$_3$–NH$_2$ peptide model in the α-helix and β-strand regions for the ^{a}C and H^{a} nuclei

theory, the strength of theory is in predicting trends. In other words, differences between computed shifts of conformations should be considerably more accurate than the absolute values. The need for shifting the computed ICS values appeared in other studies, as well (Sun et al. 2002; Le et al. 1995). For filtering the experimental points and adjusting the computed ICS values to experimental ones, the following simple protocol was followed for each amino acid residue:

(1) Determine the signed difference between each experimental point and the ab initio surface and compute the signed average.
(2) Adjust the computed surface, in form of a simple shift by a constant value given by the signed average.
(3) Compute the differences d_i between the experimental points and the shifted surface, and determine the mean absolute deviation (MAD) of these differences,

$$\mathrm{MAD} = \frac{\sum_{n} |d_i - d_{ave}|}{n}$$

(4) Exclude those experimental points which differ from the shifted computed surface by more than 2 MAD.
(5) Recompute the d_i differences between the reduced set of experimental points and the directly computed surface, and take the signed average of these new differences.
(6) Shift the computed surface by this new average.

We report some statistics in Tables 5–6 concerning the final adjusted (empirical) surfaces and the number of data points excluded. The shift from the original computed surfaces to the empirical surfaces generated by the above procedure is given in the same table.

Results and discussion

Although this study basically concentrates on the ICS variations within four regions of the Ramachandran surface, it is worth investigating whether the ICS of the $^{13}C^{a}$ nucleus is suitable in itself for discriminating between secondary structure types. In line with the first observations of the secondary chemical shifts, (Spera and Bax 1991; Wishart et al. 1991; Dalgarno et al. 1983; Szilágyi and Jardetzky 1989) the ICSs of $^{13}C^{a}$ differ significantly between the α-helix and β-strand regions. As to the computed

Table 2 Statistics concerning the experimental $^{13}C^{a}$ chemical shifts as compared to the computed, fitted surfaces of mono- (M), tri- (T) and penta-peptide (P) models of the α-helix region[a]

Aaa	No. of Exp points	No. of filtered expt. Points			Average of Abs[diffs] (ppm)			MAD of Abs[diffs] (ppm)			Max of Abs[diffs] (ppm)			Shift (ppm)		
		M	T	P	M	T	P	M	T	P	M	T	P	M	T	P
Trp	85	84	84	84	1.19	1.20	1.18	0.67	0.74	0.74	2.97	3.05	3.16	2.13	0.81	0.29
Leu	563	541	536	538	0.56	0.52	0.55	0.35	0.31	0.32	1.91	1.79	1.86	−0.38	−1.67	−2.16
Met	160		159	159		1.25	1.26		0.71	0.71		3.32	3.42		−1.17	−1.69
Phe	230		222	222		1.05	1.08		0.64	0.66		3.55	3.74		1.60	1.08
Tyr	139	132	134	136	0.90	0.93	1.03	0.63	0.69	0.74	3.06	3.03	3.49	2.93	1.84	1.27
Cys	62	62	62	62	2.74	2.77	2.82	1.39	1.31	1.27	6.42	6.91	6.93	2.93	1.61	1.12
Glu	621	599	593	596	0.71	0.63	0.68	0.45	0.40	0.41	2.34	2.17	2.32	1.19	−0.14	−0.63
Gln	259		249	252		0.63	0.69		0.40	0.43		2.13	2.22		−0.68	−1.20
Lys	470	456	457	458	0.80	0.80	0.83	0.48	0.45	0.48	2.73	2.66	2.84	1.09	−0.25	−0.77
Arg	333	323	322	322	0.76	0.75	0.77	0.50	0.47	0.48	2.46	2.54	2.58	1.11	−0.20	−0.76
His	116	113	114	114	1.07	1.09	1.12	0.66	0.68	0.69	3.41	3.38	3.33	0.86	−0.44	−0.90
Asp	354	340	335	336	0.67	0.60	0.64	0.42	0.38	0.42	2.46	2.38	2.38	−0.76	−2.14	−2.69
Asn	182	175	175	176	0.64	0.61	0.69	0.38	0.40	0.42	1.91	2.06	2.17	−2.10	−3.48	−4.00
Val	348	334	334	333	0.72	0.71	0.75	0.44	0.44	0.45	2.58	2.69	2.68	8.12	6.92	6.46
Ile	314		309	309		1.09	1.12		0.67	0.69		3.34	3.41		5.24	4.77
Thr	235		227	227		1.11	1.09		0.68	0.71		3.84	3.78		6.59	6.12
Pro	200		196	197		0.64	0.69		0.38	0.42		1.95	2.10		5.90	5.15
Ala	650	623	623	626	0.50	0.47	0.51	0.31	0.28	0.31	1.71	1.65	1.70	−3.16	−4.49	−45.07
Ser	302		289	289		0.77	0.77		0.51	0.51		2.69	2.88		2.07	1.53
Gly	194	189	191	192	0.62	0.60	0.66	0.42	0.40	0.45	2.86	2.40	2.87	−11.0	−12.3	−12.9

[a] The amino acids are grouped according to a scheme advocated by Iwadate et al. (1999). The amino acids His and Cys were excluded from Iwadate's analysis. They were placed to a group based on their shift values

ICSs of the middle residue in the For-(Ala)$_3$-NH$_2$ model, the averaged $\delta^{13}C^{a}$ for the α-helix and β-strand regions is 59.1 and 53.8 ppm, respectively. The range of computed ICS values in the α-helix region is 57.2–61.1 ppm, whilst the range is 52.8–55.1 ppm in the β-strand region. The corresponding standard deviations are only 1.1 ppm and 0.6 ppm. Since these points are on a surface, the standard deviations only indicate how much the surface changes in a given region. Therefore, the greater standard deviation of the helical region is not in contrast with the usual observation that in case of experimental information the ICSs in the β-strand region have a greater spread (Spera and Bax 1991). The important fact to notice is that the ICSs characterizing the α-helix and β-strand regions are clearly separated. On the other hand, the average ICS values of the poly-proline II and left-handed α-helix regions, 55.0 and 59.8 ppm, respectively, reflect the usual problem that these regions, using only a single ICS parameter, cannot be clearly separated from the other two regions.

A more detailed discussion of the computed and the related empirical ICS surfaces is preceded by a brief analysis of model size effect and a discussion concerning the general validity of the use of Ala as a model residue in studies of the general structural properties of peptides and proteins.

All ab initio ICS(φ,ψ) surfaces for the α-helix and β-strand regions and all empirical ICS(φ,ψ) surfaces with the experimental points for all four regions investigated are deposited in the Supplementary Material. The Supplementary Material also contains tables of all so-called outlying residues (vide infra).

Mono- versus tri- versus penta-peptide ICS(φ,ψ) surfaces

As shown earlier, (Czinki and Csizmadia 2004) the ICS(φ,ψ) NMR surfaces of the nuclei of the simplest For-Ala-NH$_2$ model, covering the full Ramachandran map, are essentially the same as those corresponding to the larger peptide models. Comparing the surfaces of the tri- and penta-peptide models of this study corresponding to secondary structure regions (Model 2), basically confirms these earlier results.

The structure of the surfaces corresponding to the mono-, tri-, and penta-peptide models varies slightly. For example, in case of the helical region of $^{13}C^{a}$ the mono-peptide

Table 3 Statistics concerning the experimental $^{13}C^{a}$ chemical shifts as compared to the computed, fitted surfaces of mono- (M), tri- (T) and penta-peptide (P) models of the β-strand region[a]

AA	No. of Exp points	No. of filtered exp. points			Average of Abs[diffs] (ppm)			MAD of Abs[diffs] (ppm)			Max of Abs[diffs] (ppm)			Shift (ppm)		
		M	T	P	M	T	P	M	T	P	M	T	P	M	T	P
Trp	45	43	43	43	1.04	1.06	1.07	0.51	0.53	0.53	3.03	2.86	2.86	6.24	2.47	2.38
Leu	209	202	200	200	0.63	0.62	0.63	0.38	0.35	0.36	2.16	2.00	1.94	3.80	−0.07	−0.16
Met	49		48	48		0.73	0.73		0.39	0.38		2.19	2.14		0.45	0.35
Phe	109		103	103		0.79	0.80		0.43	0.44		2.23	2.25		2.76	2.66
Tyr	118	112	111	111	1.22	1.14	1.15	0.78	0.70	0.70	4.12	3.91	3.92	6.69	2.93	2.83
Cys	36	35	35	35	1.51	1.48	1.48	0.85	0.83	0.84	4.43	4.21	4.21	5.90	1.95	1.86
Glu	118	115	114	114	0.83	0.71	0.71	0.55	0.48	0.49	2.45	2.51	2.53	4.90	1.01	0.91
Gln	73		70	70		0.74	0.75		0.41	0.41		2.06	2.11		0.54	0.44
Lys	137	132	132	132	0.79	0.73	0.74	0.46	0.44	0.45	2.34	2.50	2.56	4.94	1.06	0.97
Arg	103	99	99	99	0.90	0.84	0.84	0.55	0.53	0.53	2.82	2.52	2.52	4.59	0.71	0.61
His	42	42	42	42	1.27	1.26	1.27	0.69	0.68	0.68	3.64	3.54	3.55	4.57	0.73	0.64
Asp	42		42	42		0.73	0.74		0.49	0.50		2.08	2.04		−0.66	−0.76
Asn	34		31	31		0.58	0.58		0.31	0.31		2.12	2.04		−1.30	−1.40
Val	339	332	328	328	1.03	0.91	0.92	0.63	0.57	0.58	3.42	3.12	3.14	11.0	7.07	6.97
Ile	285	280	278	278	0.96	0.87	0.88	0.57	0.52	0.52	3.06	2.90	2.92	10.1	6.16	6.06
Thr	206		201	201		0.90	0.91		0.55	0.55		2.89	2.89		7.62	7.52
Pro	0															
Ala	67	64	64	64	0.63	0.61	0.62	0.40	0.33	0.33	1.93	2.19	2.20	0.57	−3.22	−3.32
Ser	70	67	68	68	0.78	0.79	0.80	0.49	0.49	0.49	2.54	2.56	2.58	6.74	3.08	2.97
Gly	25	24	24	24	0.78	0.62	0.63	0.46	0.40	0.41	2.20	2.13	2.15	−5.51	−9.25	−9.35

[a] See footnote "a" to Table 2

surface is a bit more structured than the tri- and penta-peptide surfaces, but only at the less interesting edges of the surface, where no experimental points are found. The β-strand region of the $^{13}C^{a}$ surface of the mono-peptide model is a bit less structured, lacks a slight twist, in its counterpart tri- and penta-peptides, with the latter two being essentially the same. None of these slight variations have any consequence on the quality of the data deduced in this study.

The same conclusions apply when one investigates the ICS surface of $^{1}H^{a}$. In case of the β-strand, the tri- and penta-peptide surfaces are the same, deviating very little from the mono-peptide model surface. Besides these qualitative similarities, the statistical measures of the fit of experimental points to these surfaces are almost identical for the different models, especially close is the agreement for the tri- and penta-peptide models. In case of the α-helical $^{1}H^{a}$ ICS(φ,ψ) surface, one again sees only a small change in the structure of the surfaces when moving from the mono- to the penta-peptide models.

These observations strengthen our previous findings (Czinki and Csizmadia 2004; Czinki et al. 2003; Perczel and Csizmadia 2000, 2001, 2002; Perczel et al. 2003) that, considering only the middle residue, there is no significant difference in the utility of the tri- and penta-peptide models. The absolute ICS values may vary slightly when going from the mono- to the penta-peptide models, but this has no consequence for the present study. Even the absolute values do not change in any significant manner when moving from the tri- to the penta-peptide models. Overall, it is clear that the ICS(φ,ψ) surfaces of all peptide models investigated are essentially the same; therefore, in what follows the tri-peptide surfaces will be used for analysis or when general statements are made.

Is Ala a good structural model for other residues?

For all four Ramachandran regions of the models, not just the experimental $^{13}C^{a}$ ICS values corresponding to the Ala residue fit well the computed surfaces of the tri-peptide model but basically all other residues, as well. After the adjustment protocol described above, the average of the absolute remaining errors of the experimental ICS points provides a measure suggesting how well the surface corresponding to the Ala model describes the other residues. Based on the data presented in Table 2–6, we can

Table 4 Statistics concerning the experimental[a] chemical shifts as compared to the computed, fitted surfaces of mono- (M), tri- (T) and penta-peptide (P) models of the α-helix region[a]

Aaa	No. of Exp points	No. of filtered exp. points			Average of Abs[diffs] (ppm)			MAD of Abs[diffs] (ppm)			Max of Abs[diffs] (ppm)			Shift (ppm)		
		M	T	P	M	T	P	M	T	P	M	T	P	M	T	P
Trp	63	61	61	62	0.286	0.286	0.301	0.158	0.155	0.175	0.788	0.724	0.907	0.049	0.064	−0.057
Leu	455	438	440	441	0.184	0.189	0.195	0.114	0.119	0.122	0.635	0.675	0.662	−0.312	−0.290	−0.404
Met	133		130	130		0.229	0.231		0.142	0.139		0.726	0.737		−0.184	−0.303
Phe	178		178	178		0.345	0.349		0.208	0.210		1.020	1.020		−0.239	−0.345
Tyr	106	100	102	102	0.218	0.235	0.243	0.137	0.155	0.166	0.752	0.822	0.834	−0.201	−0.220	−0.324
Cys	74	69	69	69	0.328	0.338	0.340	0.169	0.178	0.184	0.923	0.897	0.835	−0.051	−0.028	−0.134
Glu	503	480	477	477	0.134	0.132	0.136	0.084	0.083	0.085	0.491	0.517	0.513	−0.305	−0.283	−0.391
Gln	216		207	206		0.181	0.180		0.114	0.112		0.606	0.549		−0.270	−0.373
Lys	404	381	382	383	0.157	0.161	0.174	0.101	0.102	0.106	0.599	0.611	0.646	−0.312	−0.289	−0.395
Arg	270	259	258	259	0.186	0.184	0.193	0.130	0.128	0.134	0.639	0.643	0.653	−0.335	−0.314	−0.415
His	79	75	74	74	0.216	0.223	0.219	0.163	0.153	0.144	0.775	0.757	0.699	0.029	0.026	−0.085
Asp	276	264	264	267	0.132	0.131	0.147	0.083	0.082	0.088	0.441	0.405	0.498	0.068	0.101	0.008
Asn	147	144	143	143	0.149	0.152	0.153	0.080	0.081	0.084	0.464	0.434	0.429	0.118	0.155	0.049
Val	295	290	287	288	0.240	0.240	0.248	0.156	0.151	0.149	0.772	0.785	0.768	−0.750	−0.736	−0.854
Ile	260		248	247		0.210	0.217		0.132	0.139		0.707	0.732		−0.643	−0.770
Thr	206		199	197		0.203	0.208		0.124	0.127		0.656	0.672		−0.325	−0.441
Pro	157		149	149		0.165	0.172		0.102	0.098		0.542	0.579		0.047	−0.006
Ala	515	491	493	496	0.167	0.172	0.179	0.099	0.100	0.107	0.592	0.627	0.633	−0.258	−0.228	−0.329
Ser	256		244	245		0.162	0.172		0.104	0.108		0.560	0.626		−0.043	−0.143
Gly	0															

[a] See footnote "a" to Table 2

conclude that Ala is a particularly good model to describe Leu. Depending slightly on the region investigated, the ICS(u,w) surfaces of Lys, Asn, Gly, Pro, Asp, and Glu are modelled well by Ala. As have been observed before, (Iwadate et al. 1999) the experimental ICS points of Cys fit worst to our Ala surfaces. Thus we confirm that the ICS values of Cys should be excluded from statistical analyses. It is not too surprising either that Phe, Met, Tyr, Trp, and His are somewhat less precisely described by the Ala surfaces.

It is common practice to try to group amino acids according to their physico-chemical properties. This grouping does not seem to work in the present case as statistical measures of amino acids in the same group show very limited similarities. However, using Iwadate's grouping scheme (Iwadate et al. 1999) one does find similarities between statistical measures regarding agreement with the Ala surface results among amino acid residues within the same group. Therefore, this grouping is used in Tables 2–6 and in the figures as it allows an easier presentation of the results obtained.

Overall, it can be concluded that Ala, with the exception of Cys, provides a good to excellent model for the characterization of secondary structure–ICS relations of residues.

ICS(u,w) for $^{13}C^{a}$ The α-helix region

The empirical surfaces along with the filtered experimental points in the α-helix region can be seen in Fig. 2. for a selection of amino acid residues, providing one example for each of the 5 groups advocated by Iwadate.

As to the deviations of the experimental points from the empirical α-helix surfaces, one can find a relatively impressive fit. Experimental points found in the "diagonal" of the region fit considerably better to the empirical surface, and the largest deviations are observed at the edges of the region. This is in clear agreement with expectation. Two problems are noted which might account for some of the deviations. First, it is not checked whether the residue for which the deviation is large is in the middle of an α-helix, i.e., it is not checked what the backbone dihedral angles of its neighbours are. Second, the characteristic dihedral angles of other secondary structure types, e.g., certain residues of loops, also fall into the large and square-shaped "helix" region defined. These tests were not done as it was felt that further filtering of the experimental points would jeopardize the validity of the findings of the present joint experimental and theoretical study.

Table 5 Statistics concerning the experimental $^{1}H^{\alpha}$[a] chemical shifts as compared to the computed, fitted surfaces of mono- (M), tri- (T) and penta-peptide (P) models of the β-strand region[a]

AA	No. of Exp points	No. of filtered exp. points			Average of Abs[diffs] (ppm)			MAD of Abs[diffs] (ppm)			Max of Abs[diffs] (ppm)			Shift (ppm)		
		M	T	P	M	T	P	M	T	P	M	T	P	M	T	P
Trp	32	32	32	32	0.339	0.340	0.340	0.175	0.168	0.168	0.69	0.70	0.70	0.263	0.442	0.452
Leu	170	166	166	166	0.320	0.320	0.320	0.170	0.172	0.172	0.95	0.96	0.97	0.041	0.236	0.246
Met	42		40	40		0.319	0.319		0.184	0.184		0.80	0.80		0.317	0.328
Phe	94		94	94		0.399	0.399		0.238	0.238		1.19	1.19		0.276	0.286
Tyr	94	91	91	90	0.431	0.433	0.424	0.290	0.288	0.278	1.45	1.44	1.42	0.097	0.282	0.308
Cys	48	47	47	47	0.342	0.337	0.338	0.244	0.245	0.244	0.96	0.95	0.94	0.230	0.422	0.432
Glu	107	107	107	107	0.326	0.326	0.326	0.166	0.169	0.169	0.94	0.93	0.94	−0.067	0.124	0.135
Gln	62		61	61		0.343	0.343		0.212	0.212		1.01	1.02		0.133	0.143
Lys	131	130	130	130	0.343	0.337	0.338	0.180	0.183	0.183	1.00	0.99	0.99	−0.068	0.123	0.133
Arg	90	90	90	90	0.379	0.378	0.379	0.202	0.200	0.200	1.05	1.03	1.03	−0.047	0.148	0.159
His	40	40	40	40	0.329	0.328	0.327	0.192	0.189	0.190	0.95	0.98	0.98	0.163	0.356	0.367
Asp	33		32	32		0.285	0.285		0.158	0.158		0.91	0.91		0.322	0.329
Asn	27		27	27		0.323	0.323		0.164	0.164		0.81	0.81		0.530	0.540
Val	281	276	277	277	0.333	0.332	0.333	0.192	0.195	0.195	1.02	1.04	1.04	−0.301	−0.108	−0.097
Ile	224	223	222	222	0.350	0.345	0.345	0.206	0.203	0.203	1.05	0.97	0.97	−0.194	0.008	0.019
Thr	196		193	193		0.341	0.341		0.192	0.192		1.00	1.00		0.194	0.205
Pro	0															
Ala	50	48	48	48	0.356	0.350	0.350	0.229	0.223	0.224	1.17	1.14	1.13	0.196	0.380	0.390
Ser	58	58	58	58	0.372	0.373	0.373	0.211	0.215	0.215	0.90	0.94	0.94	0.216	0.402	0.413
Gly	0															

[a] See footnote "a" to Table 2

The same qualitative picture, namely that there is a seemingly good correspondence between experiment and theory and that the experimental points differ more from the surface in the edge region, applies more or less to all residues. Not surprisingly, (Iwadate et al. 1999) Cys is an exception once again. It has the worst fit, the experimental points on average have the greatest errors, with the highest absolute deviation from the average, for Cys.

The β-strand region

An analysis similar to that reported in the previous sub-section was carried out for the β-strand region of the Ramachandran map. The computed Ala $^{13}C^{\alpha}$ surfaces were compared to experimental ICS values of 19 residues, as experimentally no Pro residues were found in this region.

By visual inspection, the deviation of the experimental points from the computed surfaces looks a bit worse than that for the α-helix region, although statistically the overall picture is just very slightly worse. The shift is almost the same here as for the α-helix region, or at least the trends are the same. For example, for Ala and Ser the α-helix shifts are −4 and +2 ppm, while the β-strand shifts are −3 and +3 ppm, respectively. Figure 3 shows the experimental points falling into the β-strand regions for certain amino

Table 6 Statistics concerning the experimental $^{13}C^{\alpha}$[a] chemical shifts as compared to the computed, fitted surfaces of mono- (M), tri- (T) and penta-peptide (P) models of the polyproline II region. The amino acids are ordered according to a scheme given by Iwadate et al. (1999)

	Leu	Cys	Glu	Lys	Arg	Asp	Val	Ile	Thr	Pro	Ala	Ser
No. of Exp. Points	155	16	73	106	57	83	91	69	63	203	100	82
No. of filtered exp. points	149	16	70	103	55	79	85	67	60	198	97	80
Average of Abs[diffs] (ppm)	0.81	3.28	0.91	0.68	0.73	0.83	0.82	1.30	1.15	0.60	0.74	1.09
MAD of Abs[diffs] (ppm)	0.52	1.34	0.59	0.42	0.47	0.54	0.51	0.64	0.68	0.33	0.40	0.71
Max of Abs[diffs] (ppm)	2.57	6.67	2.96	2.23	2.30	3.15	2.73	2.84	3.76	1.94	2.05	3.21
Shift (ppm)	−0.12	1.68	1.35	1.30	1.17	−0.64	7.66	5.78	7.85	7.06	−3.18	3.03

Fig. 2 Empirical $^{13}C^{a}$ ICS(u,w) surfaces, based on the related computed surface of the For-Gly-(Ala)$_2$ peptide model, of given residues in the a-helix region showing the experimental data points, as well

acids along with the empirical surface. Clearly, there is no subregion with an excellent fit such as the diagonal part of the a-helix region. The larger scatter of the points could partially be explained by the diversity of b-strand structures; for example, there are parallel and antiparallel b-sheets. As in the case of the a-helical region, the Cys surface is the worst in terms of the deviation of the experimental points from the empirical surface.

The a_D region

For the purposes of the present study we considered the amount of experimental information for a residue significant if more than 20 experimental points fall into the region of interest. The lack of statistically significant number of experimental points falling to the left-handed a-helix region hinders a thorough analysis in this region. Thus, we report results for just a few amino acid residues: Gly and Asn in case of $^{13}C^{a}$ and Asn in case of $^{1}H^{a}$. The partial results obtained are given on Fig. 4. The data available do not allow a thorough discussion.

Poly-proline II region

Although the catchment regions of the PPII and b-strand secondary structure types are highly similar, (Vlasov et al. 2001) the PPII region was studied separately. Residues were selected strictly on the basis of their f and w angles. There is limited experimental information available for certain residues; therefore, only a few interesting examples

Fig. 3 Empirical $^{13}C^{a}$ ICS(u,w) surfaces of given residues in the strand region showing the experimental data points, as well

are given. For almost all residues investigated, the deviation of the experimental points from the empirical surface is larger than for the strand region (see Table 6 and Fig. 5).

ICS(u,w) for $^{1}H^{a}$ shifts

The other nucleus besides $^{13}C^{a}$ for which changes in the ICS(u,w) surface might principally be the result of local torsional effects is $^{1}H^{a}$.

Repeating the same analysis as performed for $^{13}C^{a}$ in case of the computed and experimental $^{1}H^{a}$ chemical shifts resulted in the following conclusions: (1) The simple function used for the $^{13}C^{a}$ shifts is sufficient for describing the interesting regions of the ICS(u,w) surfaces for $^{1}H^{a}$; the statistical measures are quite similar, being between 0.97 and 0.99 for $^{1}H^{a}$. (2) The number of outlying points (vide infra) was about the same for $^{1}H^{a}$ and $^{13}C^{a}$. (3) For both nuclei, more or less the same residues had to be filtered out. (4) Similarly to $^{13}C^{a}$, there is no significant difference in the utility of the tri- and pentapeptide ICS(u,w) surfaces. These results suggest the considerable utility of $^{1}H^{a}$ ICS information in structural studies on proteins (Figs. 6 and 7).

Fig. 4 Empirical $^{13}C^{a}$ and $^{1}H^{a}$ ICS(ϕ,ψ) surfaces of given residues in the left-handed α-helix region showing the experimental data points, as well

Outlying points

As shown above, the ICS surfaces of the Ford(AD)$_2$NH$_2$ model describe quite well the experimental information available for most residues. This is especially impressive if one keeps in mind the oversimpliÞed selection procedure used to allocate the residues to a given secondary structure type. The underlying assumption is that those chemical shifts are useful which are determined almost exclusively by local structural effects. Nevertheless, as discussed above, exclusion of some experimental points cannot be avoided.

One might expect that the claimed resolution of the X-ray structure determination correlates with the number of outlying residues. Figure 5 shows the resolution of the protein structure determination with respect to the number of outlying residues found in that protein. Somewhat surprisingly, the Þgure shows no correlation, suggesting that the real problem is not with the X-ray structural determination of the proteins but with the suspected uncertainties of the ICSs. Some of the experimental points eliminated automatically following the protocol described above were checked by hand. The detailed analysis of all the ÔproblematicÕ residues exceeds the scope of this study. Nevertheless, upon visual inspection of some of the outlying experimental points, they were usually found to correspond to residues of one or more of the following type:

(1) The residue is actually not part of the secondary structure type that their ϕ and ψ angles suggest, but is part of a loop or a turn instead.
(2) The residue looks like part of the given secondary structure, but actually it is part of a short structural element, e.g., a short interconnecting helix, turn or loop.
(3) The residue, or a nearby residue, binds a ligand, is a catalytic residue, or there is a metal or DNA chain nearby.
(4) The residue is the Þrst/last residue of the secondary structure unit. In these cases it depends on the classiÞcation algorithm whether that residue is part of an α-helix, for example, or part of a loop.
(5) The residue is in the outer surface or in an inner pocket of the protein exposed to unusual interactions.

Based on the simple statistical measures discussed, the following proteins possess the largest number of outlying residues (given in parentheses) when the $^{13}C^{a}$ and $^{1}H^{a}$ nuclei are both considered: 1D8CA (26), 1G6TA (25), 1A7TB (21), 1CEX0 (21), 1G6SA (21), and 1CDLB (20).

Fig. 5 The original computed surface of the For-(Ala)$_3$-NH$_2$ peptide model and empirical $^{1\beta}C^{\alpha}$ ICS(ϕ,ψ) surfaces of given residues in the polyproline II region showing the experimental data points, as well

Note that 1G6TA and 1G6SA are the same protein chains with different ligands, so it is not surprising that they possess a similar number of outlying points. The dataset of all filtered-out points of the tri- and penta-peptide models for all nuclei and regions is available in the Supplementary Material.

Perhaps it is worth discussing a few outlying points. The (ϕ,ψ) = (−59.8, −46.4) dihedral angles of the 192nd residue of 1G6TA, a Lys, make it fall into the α-helix region. When investigating the actual secondary structure types around this residue, it turns out that this residue is part of a β-sheet (see Fig. 9), in fact it is in the F sheet, according to the PROMOTIF classification scheme, of residues 187–195. The same observation holds for the 192nd residue of chain A of 1G6S. This protein, furthermore, has a ligand bound to residues 194–200, making the chemical shifts of residue 192 deviate substantially from the empirical ICS surface. The following incomplete list of residues of the given proteins became outlying points due to the fact that they are at the end or at the beginning of a structural unit, a β-strand or an α-helix, and therefore cannot be regarded as regular residues of that region: Lys-390 (1D8C), Val-119 (1D8C), Val-535 (1D8C), Glu-709 (1D8C), and Lys-198 (1G6T).

Experimental C^{α}–H^{α} correlation plots

Let us finally analyse the experimental ICS results of the Ala residues from a different perspective, by trying to make ICS–ICS correlation plots introduced in Perczel and Csa

Fig. 6 Empirical $^{1}H^{a}$ ICS(u,w) surfaces of given residues in the α-helix region, showing the experimental data points, as well

Fig. 7 Empirical $^{1}H^{a}$ ICS(u,w) surfaces of given residues in the β-strand region showing the experimental data points, as well

Fig. 8 The number of residues found to be outliers within a protein with respect to the resolution of the structure determination of that protein

Fig. 9 The structure of the protein of PDB accession code 1G6T, chain A. Color codes: blue = sheet, red = helix, and green = Lys-192

Fig. 10 Experimental ICS(C^{α})–ICS(H^{α}) correlation plots of Ala residues. (**A**) All experimental points falling to the a-helix and b-strand regions are plotted. (**B**) Same as (**A**) except that the outlying points are filtered out

2000, 2001, 2002; Czinki et al. 2003; Perczel et al. 2003 and Czinki et al. 2007. An interesting correlation plot would concern experimental $^{1\beta}C^{a}$ and $^{1}H^{a}$ ICSs. Such plots were shown before to allow unambiguous distinction between secondary structure types for small model peptides.

Using all experimental points of Ala residues falling into the (above described) a-helix and b-strand regions, irrespective of their other structure or sequence qualifications, the a-helix and b-strand regions can almost fully be distinguished on the ICS($^{13}C^{a}$)–ICS($^{1}H^{a}$) correlation plot (Fig. 10). There is a minuscule area where the shifts of the two regions overlap. Apart from this area, the a-helix and b-strand regions are distinct. The distinction of the two secondary structure types becomes even more pronounced when those experimental data which correspond to outlying points by the statistical procedure described above are filtered out. Importantly, the $^{13}C^{a}$–$^{1}H^{a}$ correlation plots of other residues show the same characteristics. When the filtered datasets are plotted, the helical and strand regions become clearly distinguishable, in the same manner as in the case of the experimental points of the Ala residue.

Summary

A database of proteins containing experimentally determined backbone torsion angles, obtained from the PDB database, and isotropic chemical shifts, obtained from the RefDB database, has been assembled. The database

determined, presented in Table 1, contains 175 proteins and 16,341 residues and includes ICS(ϕ,ψ) data for all common amino acids.

This database can be used for establishing ICS(ϕ,ψ) surfaces. Although the changes in ϕ and ψ can be regarded as the main source for variations in the ICSs of the nuclei $^{13}C^{a}$ and $^{1}H^{a}$, due partly to the considerable scatter in the experimental data and partly to the lack of sufficient data and the existence of other structure–ICS effects, construction of experimental ICS(ϕ,ψ) surfaces for these and other nuclei proved impossible. Clearly, ICS(ϕ,ψ) surfaces can only be determined with help from accurate ab initio computations. After an appropriate fitting, the surfaces computed at the DFT(B3-LYP) level for the four most populous regions of the Ramachandran surface, α-helix, β-strand, left-handed α-helix (α_D), and polyproline-II of three simple alanine-containing model systems, For–(L-Ala)$_n$–NH$_2$, where n = 1, 3, 5, were used, together with the experimental ICS(ϕ,ψ) information, for constructing empirical ICS(ϕ,ψ) surfaces for all residues. In turn, these empirical surfaces provide a nice check toward assessing the usefulness of ab initio ICS values.

An important conclusion from this study is that the computed alanine ICS(ϕ,ψ) surface describes well basically all amino acid residues, except Cys. As to the model size effect on the empirical ICS(ϕ,ψ) surfaces, the shapes and the corresponding statistical measures of the surfaces based on the For–(Ala)$_3$–NH$_2$ or For–(L-Ala)$_5$–NH$_2$ models were found to be basically identical. The adequacy of a tripeptide model, when only the middle residue and especially its C^{a} and H^{a} nuclei are of interest, was proved in this study beyond reasonable doubt.

Confirming the choice made for our simple model, i.e., using only Ala as a building block, variation in the side chain characterizing the different residues has a relatively small effect on the empirical ICS surfaces. The experimental points coming from different sequence environments fit well the shifted Ala surface and show a mostly random deviation from this surface. The computed surface proved also useful for a simple automated exclusion of "problematic" experimental points.

Experimental ICS($^{13}C^{a}$)–ICS($^{1}H^{a}$) correlation plots for all common residues, determined so far only ab initio for small model peptides, (Perczel and Csizmadia 2000, 2001, 2002; Czinki et al. 2003; Perczel et al. 2003) differentiate between the two most important secondary structure types, α-helix and β-strand. Thus, the use of ICS-based correlation plots, probably both in two and three dimensions, hold promise in providing important constraints for structure determination of proteins based on NMR spectroscopy.

Acknowledgments The work described has been supported by the Scientific Research Fund of Hungary (Grant No. OTKA T047185).

References

Becke AD (1988) Density-functional exchange-energy approximation with correct asymptotic-behavior. Phys Rev A 38:3098

Beger RD, Bolton PH (1997) Protein phi and psi dihedral restraints determined from multidimensional hypersurface correlations of backbone chemical shifts and their use in the determination of protein tertiary structures. J Biomol NMR 10:129–142

Bernstein FC, Koetzle TF, Williams GJB, Meyer EF, Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M (1977) Protein data bank – computer-based archival file for macromolecular structures. J Mol Biol 112:535–542

Braun D, Wider G, Wuthrich K (1994) Sequence-corrected N-15 random coil chemical-shifts. J Am Chem Soc 116:8466–8469

Czinki E, Csizmadia AG, Perczel A (2003) A theoretical case study of type I and type II beta-turns. Chem-Eur J 9:1182–1191

Czinki E, Csizmadia AG (2004) On NMR isotropic chemical shift surfaces of peptide models. J Mol Struct (THEOCHEM) 675:107–116

Czinki E, Csizmadia AG, Magyarfalvi G, Schreiner PR, Allen WD (2007) Secondary structures of peptides and proteins via NMR chemical-shielding anisotrophy (CSA) parameters. J Am Chem Soc 129:1568–1577

Dalgarno DC, Levine BA, Williams RJP (1983) Structural information from NMR secondary chemical-shifts of peptide Alpha-C-H protons in proteins. Biosci Rep 3:443–452

Ditchfield R (1974) Self-consistent perturbation theory of diamagnetism. Mol Phys 27:789–807

Frisch MJ et al (1998) Gaussian98, A11, Gaussian Inc., Pittsburgh, PA

Frisch MJ et al (2004) Gaussian03, Rev C02, Gaussian Inc., Wallingford, CT

Gronenborn AM, Clore GM (1994) Identification of N-terminal helix capping boxes by means of C-13 chemical shifts. J Biomol NMR 4:455–458

Helgaker T, Jaszunski M, Ruud K (1999) Ab initio methods for the calculation of NMR shielding and indirect spin-spin coupling constants. Chem Rev 99:293–352

Ho BK, Thomas A, Brasseur R (2003) Revisiting the Ramachandran plot: Hard-sphere repulsion, electrostatics, and H-bonding in the alpha-helix. Protein Sci 12:2508–2522

Hovmoller S, Zhou T, Ohlson T (2002) Conformations of amino acids in proteins. Acta Crystallogr Sect D – Biol Crystallogr 58:768–776

Iwadate M, Asakura T, Williamson MP (1999) C-alpha and C-beta carbon-13 chemical shifts in proteins from an empirical database. J Biomol NMR 13:199–211

Kleywegt GJ, Jones TA (1996) Phi/psi-chology: Ramachandran revisited. Structure 4:1395–1400

Le HB, Pearson JG, de Dios AC, Oldfield E (1995) Protein-structure refinement and prediction via NMR chemical shifts and quantum-chemistry. J Am Chem Soc 117:3800–3807

Lee C, Yang W, Parr RG (1988) Development of the colle-salvetti correlation – Energy formula into a functional of the electron-density. Phys Rev B 37:785

Neuhaus D, Williamson M (1989) The nuclear overhauser effect in structural and conformational analysis. VCH, New York

Osapay K, Case DA (1991) A new analysis of proton chemical-shifts in proteins. J Am Chem Soc 113:9436–9444

Osapay K, Case DA (1994) Analysis of proton chemical-shifts in regular secondary structures of proteins. J Biomol NMR 4:215–230

Perczel A, Csizmadia AG (2000) Toward direct determination of conformations of protein building units form multidimensional NMR experiments part I: A theoretical case study of For-Gly-NH$_2$ and For-L-Ala-NH$_2$. J Comp Chem 21:882–900

Perczel A, Csaszar AG (2001) Toward direct determination of conformations of protein building units form multidimensional NMR experiments part II: A theoretical case study of Formyl-L-Valine amide. Chem Eur J 7:1069–1083

Perczel A, Csaszar AG (2002) Toward direct determination of conformations of protein building units form multidimensional NMR experiments part III: A theoretical case study of For-L-Phe-NH_2. Eur Phys J D 20:513–530

Perczel A, Furey AK, Csaszar AG (2003) Toward direct determination of conformations of protein building units form multidimensional NMR experiments part V: NMR chemical shielding analysis of N-formyl-serinamide, a model for polar side-chain containing peptides. J Comp Chem 24:1157–1171

Ramachandran G, Sasikekharan V (1968) Conformation of polypeptides and proteins. Adv Protein Chem 23:283–437

Red–eld C, Dobson CM (1990) H-1-NMR studies of human lysozyme B: Spectral assignment and comparison with hen lysozyme. Biochemistry 29:7201–7214

Schafer A, Huber C, Ahlrichs R (1994) Fully optimized contracted gaussian-basis sets of triple zeta valence quality for atoms Li to Kr. J Chem Phys 100:5829–5835

Seavey BR, Farr EA, Westler WM, Markley JL (1991) A relational database for sequence speci–c protein NMR data. J Biomol NMR 1:217

Spera S, Bax A (1991) Empirical correlation between protein backbone conformation and C-alpha and C-beta C-13 nuclear-magnetic resonance chemical shifts. J Am Chem Soc 113:5490–5492

Szilagyi L (1995) Chemical-shifts in proteins come of age. Progr Nucl Magn Res Spectr 27:325–443

Szilagyi L, Jardetzky O (1989) Alpha-proton chemical-shifts and secondary structure in proteins. J Magn Reson 83:441–449

Sun HH, Sanders LK, Old–eld E (2002) Carbon-13 NMR shielding in the twenty common amino acids: Comparisons with experimental results in proteins. J Am Chem Soc 124:5486–5495

Vlasov PK, Kilosanidze GT, Ukrainskii DL, Kuzmin AV, Tumanyan VG, Esipova NG (2001) Left-handed conformation of poly-L-proline II type in globular proteins. Sequence speci–city. Bio–zika 46:573–576

Wang YJ (2004) Secondary structural effects on protein NMR chemical shifts. J Biomol NMR 30:233–244

Wishart DS, Sykes BD, Richards FM (1991) Relationship between nuclear-magnetic resonance chemical-shift and protein secondary structure. J Mol Biol 222:311–333

Wishart DS, Sykes BD, Richards FM (1992) The chemical-shift index – A fast and simple method for the assignment of protein secondary structure through NMR-spectroscopy. Biochemistry 31:1647–1651

Wishart DS, Case DA (2001) Use of chemical shifts in macromolecular structure determination. Meth Enzymol 338:3–34

Wolinski K, Hinton JF, Pulay P (1990) Ef–cient implementation of the gauge-independent atomic orbital method for NMR chemical-shift calculations. J Am Chem Soc 112:8251–8260

Xu XP, Case DA (2001) Automated prediction of N-15, C-13(alpha), C-13(beta) and C-13Õ chemical shifts in proteins using a density functional database. J Biomol NMR 21:321–333

Xu XP, Case DA (2002) Probing multiple effects on N-15, C-13 alpha, C-13 beta and C-13Õ chemical shifts in peptides using density functional theory. Biopolymers 65:408–423

Zhao J, Dyson HJ, Wright PE (1997) Chemical shift dispersion and secondary structure prediction in unfolded and partly folded proteins. FEBS Lett 419:285–289

Zhang HY, Neal S, Wishart DS (2003) RefDB: A database of uniformly referenced protein chemical shifts. J Biomol NMR 25:173–195