J Biomol NMR (2007) 39:275–289
DOI 10.1007/s10858-007-9197-z

ARTICLE

Combined chemical shift changes and amino acid specific chemical shift mapping of protein–protein interactions

Frank H. Schumann · Hubert Riepl · Till Maurer · Wolfram Gronwald · Klaus-Peter Neidig · Hans Robert Kalbitzer

Received: 26 February 2007 / Accepted: 31 August 2007 / Published online: 23 October 2007

Abstract Protein–protein interactions are often studied by chemical shift mapping using solution NMR spectroscopy. When heteronuclear data are available the interaction interface is usually predicted by combining the chemical shift changes of different nuclei to a single quantity, the combined chemical shift perturbation $\Delta\delta_{\text{comb}}$. In this paper different procedures (published and non-published) to calculate $\Delta\delta_{\text{comb}}$ are examined that include a variety of different functional forms and weighting factors for each nucleus. The predictive power of all shift mapping methods depends on the magnitude of the overlap of the chemical shift distributions of interacting and non-interacting residues and the cut-off criterion used. In general, the quality of the prediction on the basis of chemical shift changes alone is rather unsatisfactory but the combination of chemical shift changes on the basis of the Hamming or the Euclidian distance can improve the result. The corrected standard deviation to zero of the combined chemical shift changes can provide a reasonable cut-off criterion. As we show combined chemical shifts can also be applied for a more reliable quantitative evaluation of titration data.

F. H. Schumann · H. Riepl · W. Gronwald ·
H. R. Kalbitzer (✉)
Institute of Biophysics and Physical Biochemistry, University of Regensburg, 93040 Regensburg, Germany
e-mail: hans-robert.kalbitzer@biologie.uni-regensburg.de

T. Maurer
Analytical Sciences Department, Boehringer Ingelheim Pharma GmbH & Co. KG, 55216 Ingelheim am Rhein, Germany

K.-P. Neidig
Software Department, Bruker BioSpin GmbH, Silberstreifen 4, 76287 Rheinstetten, Germany

Keywords Chemical shift mapping · Combined chemical shift · NMR · Protein–protein interaction

Introduction

NMR spectroscopy is the only method to determine three-dimensional structures of biological macromolecules in solution. For the determination of very large structures of biological macromolecules or their complexes X-ray crystallography is still the superior method although TROSY techniques (Pervushin et al. 1997) have pushed the size-limit for NMR accessible biomolecular systems to several hundred kDa. While the determination of high resolution structures of large biomolecular systems is still a challenging task, the investigation of structural rearrangements or protein–ligand interactions by NMR spectroscopy is quite straightforward. Moreover, compared to other biochemical techniques like cross-linking or mutational studies, the observation of chemical shift changes induced by a perturbation is probably the most sensitive method for detection. As an example, the change of a hydrogen bond length by 10 pm leads to proton chemical shift changes of the order of 0.5 ppm (Li et al. 1998) when calculated according to Wagner et al. (1983), which can be easily detected in high-resolution NMR spectra. Due to this high sensitivity of chemical shifts, even small ligand induced changes and/or structural rearrangements within a macromolecule can be sensed at the atomic level. Because of these favourable properties chemical shift perturbation mapping often is the method of choice for detecting and characterizing ligand–protein interactions.

In most practical cases ^{1}H, ^{15}N and/or ^{13}C chemical shift data are available when protein–protein interactions are studied. Even though the information about residue specific changes within the molecule are represented by the respective shift differences in free and complexed state, a single quantity $\Delta\delta_{\mathrm{comb}}$ (called in the following combined chemical shift change) which comprises the information of all involved nuclei i of a given amino acid at position j in the respective protein is more convenient for data evaluation.

Several approaches have been proposed to obtain such a quantity. The simplest approaches use the normalized length of a vector $\mathbf{E}_j$ (Euclidean distance) with the components E_{ji} defined by the chemical shift differences $\Delta\delta_{ji}$ for the atoms i at a specific position j within the primary sequence of the protein (Farmer et al. 1996; Geyer et al. 1997; Terada et al. 1999; Mulder et al. 1999; Meininger et al. 2000; Gröger et al. 2003). $\Delta\delta_{\mathrm{comb}}$ for N_{a} types of atoms is then given by

$$\Delta\delta_{\mathrm{comb},j} = \sqrt{\frac{1}{N_{\mathrm{a}}}\sum_{i=1}^{N_{\mathrm{a}}}(w_i\Delta\delta_{ji})^2} \tag{1}$$

with w_i a weighting factor which accounts for differences in sensitivity of different resonances in an amino acid (e.g. amide ^{1}H and ^{15}N). The division by N_{a} is often omitted and does not play a role when N_{a} is the same for all residues j of the protein under consideration. In the following we will use an internally consistent representation of the different methods and thus the description is not identical to those in the original papers.

When chemical shifts are expressed in ppm (what we assume in this paper) then a suitable estimate for the weighting factors is given by (Geyer et al. 1997)

$$w_i = \frac{|\gamma_i|}{|\gamma_1|} \tag{2}$$

with γ_i and γ_1 the magnetogyric ratio of nucleus i and the proton, respectively. For ^{1}H, ^{15}N and ^{13}C one would obtain 1.000, 0.102, 0.251 as weighting factors. Note that this is equivalent to expressing the chemical shift changes in Hz instead of ppm. Correspondingly Terada et al. (1999) express the $^{1}\mathrm{H}^{\mathrm{N}}$, ^{15}N,$^{13}\mathrm{C}^{\alpha}$ and $^{13}C'$ chemical shift changes in Hz to describe the combined shift perturbation for each residue. A more involved scaling of the chemical shift changes was proposed by Mulder et al. (1999). They calculated atom type specific weighting factors w_i analogously to the calculation of a Z-score for ^{15}N from the ratio of the average standard deviations $\langle\sigma_{ik}^2\rangle^{1/2}$ of the corresponding chemical shifts stored in the BioMagResBank data base. Here, i ($i > 1$) is one of the above nuclei and k one of the 20 proteinogenic amino acids, that is

$$w_i = \frac{\langle\sigma_{1k}^2\rangle^{1/2}}{\langle\sigma_{ik}^2\rangle^{1/2}} = \frac{\sqrt{\frac{1}{19}\sum\limits_{k=1}^{19}\sigma_{1k}^2}}{\sqrt{\frac{1}{20}\sum\limits_{k=1}^{20}\sigma_{ik}^2}} \tag{3}$$

An exception is proline where no amide proton exists. The weighting factor for H^{N} is 1 per definition, for ^{15}N a weighting factor of 0.15 was obtained by Mulder et al. (1999). They also proposed an extension of this definition for individual amino acid types but did not follow this line (see below). According to that we calculated the weighting factor of $^{13}\mathrm{C}^{\alpha}$ and $^{13}\mathrm{C}'$ to 0.28 and 0.34. Farmer et al. (1996) used similar weighting factors of 0.17 and 0.39 for ^{15}N and $^{13}\mathrm{C}'$ estimated from the "spread" of the protein chemical shifts without defining exactly the source of the corresponding data. These factors were also used later by other authors (e.g. Meininger et al. 2000).

A different way to calculate the weighting factors without resorting to a data base was proposed by Gröger et al. (2003)

$$w_i = \frac{1}{\sigma_i} = \sqrt{\frac{N-1}{\sum\limits_{j=1}^{N}(\Delta\delta_{ji} - \langle\Delta\delta_{ji}\rangle)^2}} \tag{4}$$

where σ_i is the standard deviation of the ^{15}N (or ^{13}C) chemical shift changes $\Delta\delta_{ji}$ observed in the given protein.

An alternative definition of the combined chemical shift $\Delta\delta_{\mathrm{comb}}$ was introduced by Heitmann et al. (2003) using the Hamming distance

$$\Delta\delta_{\mathrm{comb},j} = \frac{1}{N_{\mathrm{a}}}\sum_{i=1}^{N_{\mathrm{a}}}|w_i\Delta\delta_{ji}| \tag{5}$$

together with weighting factors calculated according to Eq. 2.

There are basically two different functions used in literature for the combination of chemical shifts $\Delta\delta_{\mathrm{comb}}$,the square root of the sum of the weighted squares of the chemical shift values (Euclidean distance, Eq. 1) and the sum of weighted absolute chemical shift changes (Hamming distance, Eq. 5). These can be combined with different ways to calculate the weighting factors.

Several issues about the use of the combined chemical shifts for evaluating the protein–protein interactions are still unclear, (1) in which instances and why the proposed definitions of the combined shifts are optimal for discriminating between interacting and non-interacting residues and (2) what cut-off values should be used. In the following we will deal with these questions and propose alternatives to the methods already known.

Materials and methods

Chemical shift data for the calculation of chemical shift ranges for individual atoms and amino acid types were taken from the BioMagResBank database (URL http://www.bmrb.wisc.edu/). The chemical shift changes induced in the Ras-binding domain (Raf-RBD) of Raf by binding of wildtype Ras or Ras(D30E,E31K) complexed with Mg.GppNHp were taken from Terada et al. (1999). The corresponding X-ray structure of Rap1A (E30D,K31E) and the GTP analogue GppNHp in complex with the Ras-binding domain of c-Raf1 was solved by Nassar et al. (1995). As a second example the chemical shift data for the turkey ovomucoid third domain/bovine chymotrypsin A_α complex were taken from Song and Markley (2001) and Fujinaga et al. (1987). As next example the interaction of the PDZ2 domain of human phosphatase hPTP1E with a C-terminal peptide from the Fas receptor was analyzed (Ekiel et al. 1998; Kozlov et al. 2000). The corresponding BioMagResBank access codes for the chemical shifts of the free and complexed PDZ2 domain are bmr4123.str and bmr4124.str, respectively. As last example the interaction of the N-terminal domain of enzyme I (EIN) with HPr was analyzed. The NMR solution structure of the EIN–HPr complex from *E. coli* was taken from (Garrett et al. 1999) and the chemical shifts of free HPr and free EIN are stored under access codes bmr2371.str and bmr4106.str, respectively. Chemical shifts for the complete HPr–EIN complex can be found under bmr4246.str.

Residues involved in protein–protein interaction were defined by two criteria, (1) the water accessible surface of the residue in question decreased in the complex by more than 5% and (2) at least one atom of the residue was closer than 0.5 nm to an atom of the interacting protein in the complex. Before this calculation, the protons missing in the X-ray structure were generated. With this definition 16 residues out of 75 residues of Raf were found to be involved in the interaction in the Ras–Raf complex. The corresponding number of interacting residues in the ovomucoid–chymotrypsin complex is 10 from 47 residues in total. For the PDZ2 domain 15 out of 96 residues, for free HPr 25 from 85 residues and for free EIN 30 from 249 residues were found to be interacting. The amino acid specific calculation of $\Delta\delta_{\text{comb}}$ is implemented in the program AUREMOL-INTERACT contained in the program package AUREMOL http://www.auremol.de).

The Pearson correlation coefficient C for a vector $\mathbf{M}$ that a residue is member of the class C_i and a vector $\mathbf{D}$ that contains the binary data (0 = false, 1 = true) is given by

$$C(\mathbf{D},\mathbf{M}) = \frac{\text{TN}\bullet\text{TP} - \text{FN}\bullet\text{FP}}{\sqrt{(\text{TP}+\text{FN})(\text{TP}+\text{FP})(\text{TN}+\text{FP})(\text{TN}+\text{FN})}} \tag{6}$$

with TP, TN, FP, FN the number of true positive predictions, true negative predictions, false positive predictions and false negative predictions, respectively (Baldi and Brunak 2001). In the form of Eq. 6 the Pearson correlation coefficient is called Matthews correlation coefficient (MCC) in literature.

Theoretical considerations

Determination of the interaction sites

As obvious from the Introduction there are various ways to define the combined chemical shift changes $\Delta\delta_{\text{comb}}$ but it is not clear which is the most efficient way to discriminate interaction sites from non-interacting sites. At least three different quantities have to be defined, the general form of $\Delta\delta_{\text{comb}}$,the definition of the weighting factors for different nuclei and/or atoms, and the threshold of $\Delta\delta_{\text{comb}}$ for the assignment of residues to one of the two classes C_1 (interacting residue) or C_2 (non-interacting residue).

Probability distributions

Since the database is usually not sufficient to determine the multidimensional probability distribution of chemical shift changes introduced by complexation, a reduction to lower dimensions is required. Here, two limiting cases exist, either the components of the vector $\mathbf{E}$ are statistically independent or they are strongly correlated. In the first case the distribution for the conditional probability $p(\mathbf{E_j}|\mathbf{C_i})$ to find $\mathbf{E}_j$ if a residue j is involved in an interaction (class C_1) or not involved in an interaction (class C_2) can be written as

$$p(\mathbf{E}_j|C_i) = \prod_{k}^{N_a} p(E_{kj}|C_i) \tag{7}$$

with E_{kj} the components of $\mathbf{E_j}$. When the general form of the probability distribution is independent of the component k but its width is varying that is

$$p(\tilde{E}_{kj}|C_i) = p(w_k E_{kj}|C_i) = p(w_1 E_{1j}|C_i) \tag{8}$$

then

$$p(\tilde{\mathbf{E}}_j|C_i) = \prod_{k}^{N_a} p(\tilde{E}_{kj}|C_i). \tag{9}$$

In literature, the probability distributions of the weighted chemical shift changes are usually defined by the weighted Euclidian or Hamming distances as

$$p(|\tilde{\mathbf{E}}_j||C_i) = p\left(\sqrt{\frac{1}{N_a}\sum_{k=1}^{N_a}(w_k E_{kj})^2}|C_i\right) \tag{10}$$

or

$$p(||\tilde{\mathbf{E}}_j|||C_i) = p\left(\frac{1}{N_a}\sum_{i=1}^{N_a}|w_i \Delta\delta_{ji}|\right) \tag{11}$$

are taken as a measures for distinguishing residues that are either involved in binding or not. The other extreme case would assume that the individual properties are strongly coupled. In the simplest case (as done in discriminant analysis) they are represented by the linear relation

$$E_j = \sum_{k=1}^{N_a} w_k E_{kj} \tag{12}$$

The corresponding probability is

$$p(E_j|C_i) = p\left(\sum_{k=1}^{N_a} w_k E_{kj}|C_i\right). \tag{13}$$

Calculation of weighting factors

The weighting factors w_j can be estimated in different ways as described in Introduction. As a generalization of Eq. 3 the weighting factors can be calculated not only for specific atom types i but also for specific amino acid types k that is by omitting the averaging over the amino acid types. This was proposed earlier by Mulder et al. (1999) for ^{15}N shifts but discarded as not significant. As in Eq. 4 the normalisation to the proton chemical shift can (and is) omitted in the following that is

$$w_{ik} = \frac{1}{\sigma_{ik}} = \sqrt{\frac{N_{ik}-1}{\sum_{j=1}^{N_{ik}}(\Delta\delta_{jik} - \langle\Delta\delta_{jik}\rangle)^2}} \tag{14}$$

with the summation performed over the number N_{ik} of chemical shifts found for atom type i in the amino acid k. The corresponding values calculated for the structures contained in the BioMagResBank data base are presented in Table 1. One can expect that the information content increases using BioMagResBank data; in the worst case it will be unchanged.

Table 1 Amino acid type specific weighting factors[a]

Amino acid	H^N	N	C^α	C'
Ala	1.67	0.264	0.498	**0.45**
Arg	1.64	0.263	0.42	0.481
Asp	**1.75**	0.253	0.478	**0.556**
Asn	1.54	0.240	0.513	0.546
Cys	1.52	0.216	**0.292**	0.49
Glu	1.67	0.275	0.469	0.493
Gln	**1.70**	0.266	0.461	0.505
Gly	1.47	0.243	**0.752**	0.521
His	1.45	0.238	0.407	0.481
Ile	1.45	0.228	0.362	0.503
Leu	1.54	0.252	0.461	0.49
Lys	1.64	0.257	0.448	0.469
Met	1.67	0.276	0.448	0.478
Phe	**1.39**	0.236	0.375	0.49
Pro	–	**0.082**	**0.617**	**0.641**
Ser	1.67	0.267	0.459	**0.556**
Thr	1.59	0.203	0.369	0.552
Trp	**1.25**	0.225	0.38	0.498
Tyr	**1.35**	0.225	0.382	0.493
Val	1.45	0.206	**0.339**	0.515
Mean	1.55	0.236	0.447	0.510
σ	0.132	0.041	0.099	0.042

[a] The weighting factors w_{ik} were calculated from the spread in chemical shifts contained in the BMRB data base according to Eq. 14. Weighting factors deviating by more than one standard deviation from the mean are written in bold letters. The weighting factors have the unit 1/ppm

The values obtained for the amino acid type specific weighting factors vary significantly for some of the amino acids; a deviation by more than one standard deviation from the mean was found for 11 out of the 20 amino acids (Ala, Cys, Asp, Gly, Glu, Phe, Pro, Ser, Trp, Tyr, Val). For these amino acids the introduction of amino acid specific weighting factors is expected to have an influence on the combined chemical shift. In general, the observed relative variations of the weighting factors are much smaller for amide resonances than for C^α and H^N-atoms.

The calculation of the weighting factors is independent of the functional form of $\Delta\delta_{\text{comb}}$ and therefore the five methods proposed to calculate the weighting factors are tested for the different cases. In the following we use the compact notation NX where N designates the used function and X determines the way the weighting factors are calculated. Form 1 would be the linear expression defined by Eq. 11, form 2 would be the Hamming distance defined by Eq. 5, form 3 would be the Euclidian distance defined by Eq. 1. Correspondingly, (A) defines the weighting factors calculated from the gyromagnetic ratio (Eq. 2), (B) the empirical factors given by Meininger et al. (2000), (C) the weighting factors calculated from the atom specific chemical shift distribution from the BMRB data base (Eq. 3), (D) those calculated from the individual chemical shifts of the data under investigation (Eq. 4) and (E) the amino acid specific weighting factors introduced in this paper (Eq. 6).

Cut-off values

The final decision that should follow from the calculated $\Delta\delta_{\mathrm{comb}}$ values is the assignment of a given residue of one of the two classes C_1 and C_2. This is usually done by arbitrarily defining a threshold value that separates the assignment to the two classes. It is obvious that the choice of this threshold value strongly determines the outcome of the procedure. Even if the general chemical shift distribution of the two classes is known, the choice of such a value is not trivial since the chemical shift ranges usually overlap (see below).

Use of combined chemical shifts in titration studies

Often combined chemical shifts are used to obtain dissociation constants from a titration of a protein with a ligand. With the definitions given above this is permitted as long as the chemical shift part can be separated from the part of the equation that describes the binding of the ligand that is $\Delta\delta_{i,j}$ of a nucleus i in an amino acid in position j can be written as

$$\Delta\delta_{i,j}(c_{\mathrm{P}}^{\mathrm{total}}, c_{\mathrm{L}}^{\mathrm{total}}) = a_{i,j} f(c_{\mathrm{P}}^{\mathrm{total}}, c_{\mathrm{L}}^{\mathrm{total}}) \tag{15}$$

with $c_{\mathrm{P}}^{\mathrm{total}}$ and $c_{\mathrm{L}}^{\mathrm{total}}$ the total concentration of protein P and ligand L and a_{ij} a constant independent of these concentrations. This is true for a simple equilibrium with a dissociation constant K_{D} under fast exchange conditions where $\Delta\delta_{i,j}$ is given as

$$\Delta\delta_{i,j} = \delta_{i,j} - \delta_{\mathrm{P},i,j} = (\delta_{\mathrm{PL},i,j} - \delta_{\mathrm{P},i,j}) \cdot \frac{(c_{\mathrm{P}}^{\mathrm{total}} + c_{L}^{\mathrm{total}} + K_{\mathrm{D}}) - \sqrt{(c_{\mathrm{P}}^{\mathrm{total}} + c_{\mathrm{L}}^{\mathrm{total}} + K_{\mathrm{D}})^2 - 4 \cdot c_{\mathrm{P}}^{\mathrm{total}} \cdot c_{\mathrm{L}}^{\mathrm{total}}}}{2 \cdot c_{\mathrm{P}}^{\mathrm{total}}} \tag{16}$$

with $\delta_{\mathrm{PL}\ i,j}$ and $\delta_{\mathrm{P}\ i,j}$ the chemical shifts in the fully complexed state and in the ligand free state, respectively. The combined chemical shift $\Delta\delta_{\mathrm{comb},j}$ is then given by

$$\Delta\delta_{\mathrm{comb},j} = \Delta\delta_{\mathrm{P-L},j} \cdot \frac{(c_{\mathrm{P}}^{\mathrm{total}} + c_{\mathrm{L}}^{\mathrm{total}} + K_{\mathrm{D}}) - \sqrt{(c_{\mathrm{P}}^{\mathrm{total}} + c_{\mathrm{L}}^{\mathrm{total}} + K_{\mathrm{D}})^2 - 4 \cdot c_{\mathrm{P}}^{\mathrm{total}} \cdot c_{\mathrm{L}}^{\mathrm{total}}}}{2 \cdot c_{\mathrm{P}}^{\mathrm{total}}} \tag{17}$$

Using the Hamming distance (Eq. 5) one obtains

$$\Delta\delta_{\mathrm{P-L},j} = \frac{1}{N_{\mathrm{a}}} \sum_{i=1}^{N_{\mathrm{a}}} |w_i(\delta_{\mathrm{PL}ij} - \delta_{\mathrm{P}ij})| \tag{18}$$

Using the Euclidian distance (Eq. 1) one obtains

$$\Delta\delta_{\mathrm{P-L},j} = \frac{1}{N_{\mathrm{a}}} \sqrt{\sum_{i=1}^{N_{\mathrm{a}}} (w_i(\delta_{\mathrm{PL}ij} - \delta_{\mathrm{P}ij}))^2} \tag{19}$$

Results and discussion

Chemical shift distributions

Reasonable estimates of the probability distributions could be (and optimally are) obtained from a large data set where chemical shifts of free proteins and complexed proteins as well as the 3D-structures of the corresponding complexes are known. However, such a data set is not available yet in public data bases. A reasonable assumption for the non-interacting residues j of the class C_2 is that the corresponding chemical shift changes $\Delta\delta_{kj}$ are sufficiently well described by a Gaussian distribution with the expectation value $\langle\Delta\delta_{kj}\rangle = 0$. For residues of class C_1 larger chemical shift changes with an almost symmetrical bimodal distribution and an expectation value $\langle\Delta\delta_{kj}\rangle = 0$ are also probable.

Figure 1 shows the chemical shift distributions (four nuclei taken into account) of the two classes for two protein complexes: the complex of Raf-1-RBD with Ras(D30E/E31K) and the turkey ovomucoid third domain/bovine chymotrypsin A_α-complex. Amino acids directly involved in the interaction (class C_1) are depicted by black bars, and those not located in the interaction site in white (class C_2). As expected the C_2 residues can be rather well represented by a normal distribution. The expectation values $\langle\Delta\delta_{kj}\rangle$ for H^N, N, C^α and C′ of –0.0075, –0.0486, 0.058 and –0.04 ppm are close to 0 in case of the Ras-complex.; the standard

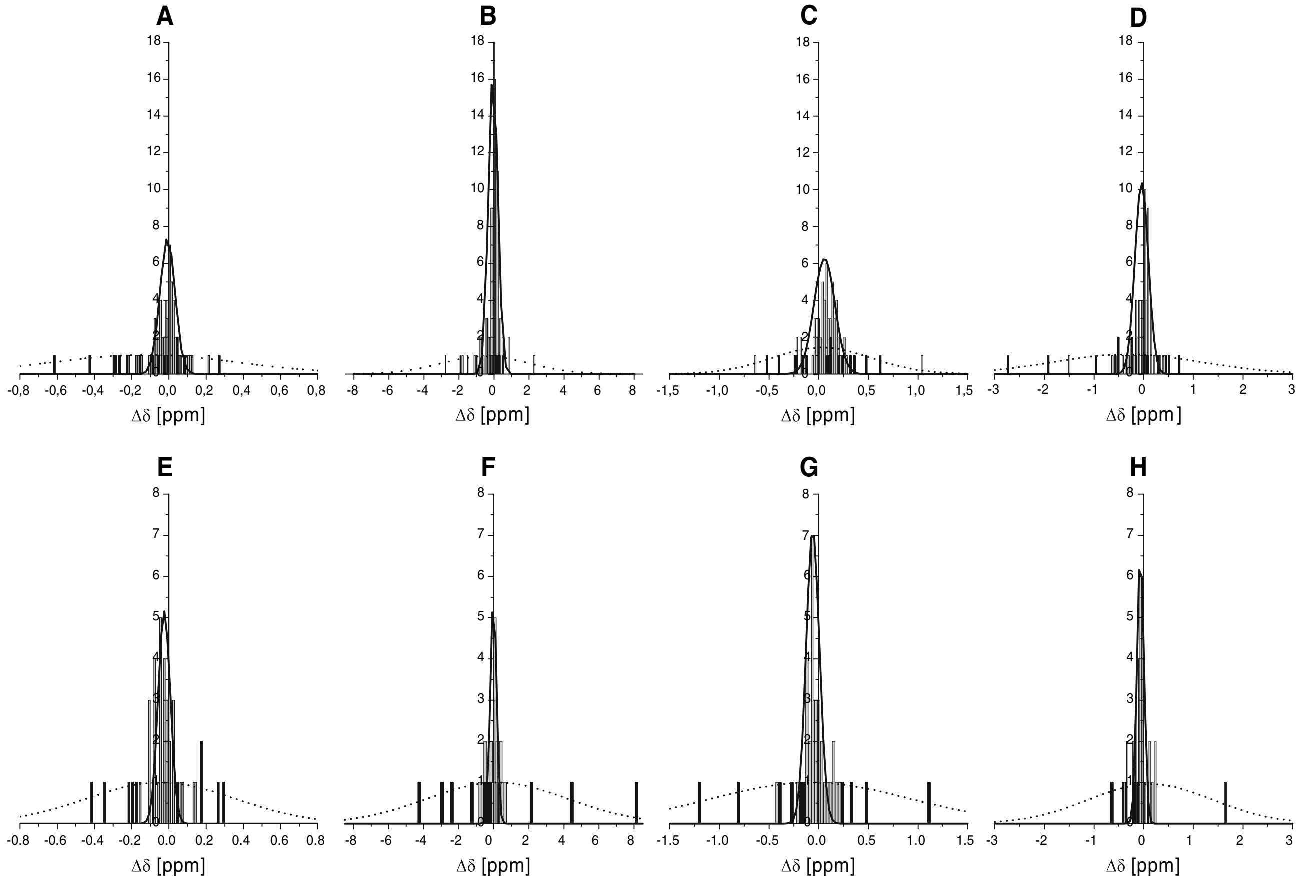

Fig. 1 Chemical shift distributions of interacting and non-interacting residues: Upper row: chemical shift distributions of (**A**) H^N, (**B**) N, (**C**) C^α and (**D**) C' of classes C_1 (black) and C_2 (white) of Raf-RBD by binding of Ras.Mg^{2+}.GppNHp (Terada et al. 1999). Lower row: chemical shift distributions of (**E**) H^N, (**F**) N, (**G**) C^α and (**H**) C' of classes C_1 (black) and C_2 (white) of the turkey ovomucoid third domain in complex with chymotrypsin A_α (Song and Markley 2001). Gaussians with corresponding expectation values and standard deviations are depicted in red for the two classes

deviations σ_{kj} are 0.081, 0.53, 0.211 and 0.269 ppm. For the ovomucoid/bovine chymotrypsin A_α-complex $\langle \Delta\delta_{kj} \rangle$ is with –0.025, 0.023, –0.06 and –0.063 ppm again approximately 0. The standard deviations σ_{kj} are with 0.065, 0.35, 0.134 and 0.137 ppm of similar magnitude to those observed in the Ras-complex.

For a discrimination of interaction induced shifts of class C_1 from noise dependent shifts of class C_2 the chemical shift distributions of the two classes must be known. However, they are not known in general. A reasonable assumption is that the chemical shift distributions of class C_1 can be sufficiently well approximated by a probability distribution with expectation value of 0. The shape of this distribution is not clear a priori; in Fig. 1a Gaussian was assumed. Since the two distributions overlap severely, it is a problem to assign the residues to one of the two classes using their chemical shift change only. However, relatively large absolute chemical shift changes are expected and are indeed observed for residues of class C_1.

Candidates for class C_1 are those amino acids whose combined chemical shift lies outside a given probability threshold defined by the width (standard deviation σ) of the C_2 distribution. This assumption is inherently used in the measures proposed in the literature although not often stated explicitly. The direct inspection of the data shows that there is always a considerable overlap of the distributions of class C_1 and class C_2 shift changes indicating that a distinction between the two classes on the basis of chemical shift changes solely has a considerable error probability, especially in the overlap region of the distributions.

Correlations between chemical shift changes of different atoms of a given amino acid

As outlined in the introduction, in literature different quantities are plotted as a function of the sequence position for the prediction of interaction sites. However, although it is usually not explicitly expressed, the definitions of combined chemical shifts are mostly based on the assumption that the chemical shift changes of different atoms in the amino acids are statistically independent.

Obviously, a correlation of chemical shift changes of neighbouring spins is not unlikely since the chemical shift changes are mainly induced by local conformational changes. The existence of such a correlation will also influence the statistical evaluation. There are two cases that may occur: in the first case both the size and the sign of the chemical shift change or in the second case only the size (absolute value) may be correlated. Tables 2 and 3 are showing the evaluation of pair-wise correlations of the backbone atoms for both cases. The chemical shift changes $\Delta\delta_{ij}$ (Table 2) of both classes are in general only weakly correlated. The exception is the correlation between the H^N and N chemical shift changes of the non-interacting residues of class C$_2$. Here, the correlation coefficient is with 0.60 rather high but with 0.04 it is almost vanishing for the same atoms of class C$_1$. The corresponding correlation coefficient of class C$_2$ shifts is virtually identical if only the absolute values $|\Delta\delta_{ij}|$ are considered (Table 3). In contrast, the corresponding correlation coefficient increases by 0.47–0.51 for the same pair of atoms in class C$_1$. This indicates that two different processes influence the ^{1}H and ^{15}N shifts in the two classes, the chemical shift variations seen for the amide groups of non-interacting residues preserve magnitude and direction, whereas for the amide groups of interacting residues the sign of the ^{1}H and ^{15}N shift change is not correlated. A plausible explanation for the last case can be derived from the known main sources of ^{1}H^N and ^{15}N chemical shift variations (Asakura et al. 1995): for the amide protons it is mainly the variation of the strength of the hydrogen bonding, for the nitrogen shifts it is mainly a change of the backbone configuration defined by the backbone torsion angles. The main factor inducing chemical shift changes by binding is the induced local conformational change; a related change of the dihedral ϕ, ψ -angles will also change the strength of a corresponding amide hydrogen bond but the signs of the corresponding ^{1}H^N and ^{15}N shift change are not correlated. Since the shift changes in the non-interacting class behave differently a different physical mechanism must apply that is not based mainly on a local structural change of the backbone dihedral angles.

Table 2 Correlations between chemical shift changes $\Delta\delta_{ji}$ of atoms i in individual amino acids j[a]

	$\Delta\delta$(H^N)	$\Delta\delta$(N)	$\Delta\delta$(C$^\alpha$)	$\Delta\delta$(C′)
C$_1$				
$\Delta\delta$(H^N)		0.04	–0.36	0.19
$\Delta\delta$(N)			0.20	–0.09
$\Delta\delta$(C$^\alpha$)				–0.43
$\Delta\delta$(*C*′)				
C$_2$				
$\Delta\delta$(H^N)		**0.60**	–0.08	0.29
$\Delta\delta$(N)			0.10	0.35
$\Delta\delta$(C$^\alpha$)				0.32
$\Delta\delta$(C′)				

[a] Correlation coefficients were calculated for Raf–RBD complexed with Ras.Mg^{2+}.GppNHp. Correlation coefficients larger than 0.5 are depicted in bold letters

Table 3 Correlations between chemical shift changes of atoms $i|\Delta\delta_{ji}|$ in individual amino acids j[a]

	$|\Delta\delta|$(H^N)	$|\Delta\delta|$(N)	$|\Delta\delta|$(C$^\alpha$)	$|\Delta\delta|$(C′)
C$_1$				
$|\Delta\delta|$(H^N)		**0.51**	**0.51**	**0.72**
$|\Delta\delta|$(N)			0.15	0.14
$|\Delta\delta|$(C$^\alpha$)				0.28
$|\Delta\delta|$ (C′)				
C$_2$				
$|\Delta\delta|$(H^N)		**0.62**	0.15	0.23
$|\Delta\delta|$ (N)			0.03	0.24
$|\Delta\delta|$(C$^\alpha$)				0.42
$|\Delta\delta|$ (C′)				

[a] Correlation coefficients were calculated for Raf-RBD complexed with Ras.Mg^{2+}.GppNHp. Correlation coefficients larger than 0.5 are depicted in bold letters

The correlations are somewhat higher for the correlations of $|\Delta\delta_{ij}|$ for some of the class C$_2$ signals (Table 3), but they are much higher for most class C$_1$ signals. All correlations of the H^N shift changes increase greatly when only the absolute value is considered. A maximum correlation coefficient of 0.72 is found for the correlation to the C′ chemical shift changes. Since some correlations exist especially in those amino acids that are assumed to be part of the interaction sites, the above derived quantities for the description of the combined chemical shift change are neither ideal nor suitable to describe the properties of the system perfectly. However, for practical purposes the aim is to identify the most efficient description.

The choice of a proper cut-off value deciding to what class a particular residue can be assigned is not trivial but necessary in the standard analysis of chemical shift perturbations. The choice of such a value would be straightforward when the chemical shift probability distributions of the classes C$_1$ and C$_2$ would be known. Then only a confidence level would have to be selected. In a first approximation the chemical shift distributions of the individual backbone atoms of non-interacting residues can be assumed to be normal distributed with a mean of zero. The data shown in Fig. 1 are in line with this approximation.

The standard deviation to zero (σ_0) of the chemical shift differences is hence a reasonable measure for the class C_2 and can be used to predict the probability that a residue with a given total chemical shift change $\Delta\delta_{\mathrm{comb}}$ is likely to belong to class C_2 residues. The standard deviation σ_0 will also be used in the following as approximation for the derived quantities defined above.

Specificity and sensitivity of different combined chemical shift definitions

Because of the cut-off problem, an accurate and unbiased comparison of the different chemical shift mapping methods is only feasible with a general procedure which does not use an explicit cut-off criterion. In practice, when using a chemical shift cut-off value one has always to choose between high sensitivity and specificity. Therefore, we calculated the statistical measures sensitivity and specificity as function of the cut-off value and plotted them against each other (Fig. 2). In addition, the corresponding distributions for the different definitions of combined chemical shifts $\Delta\delta_{\mathrm{comb}}$ are depicted in Fig. 3 for the Ras-complex. It is obvious (and to be expected) that at low sensitivities all methods provide high specificity. However, practical applications do require high sensitivity and specificity simultaneously. Here, the tested methods show significant differences. Quantitatively, clear differences are also seen for the two complexes and the predictive power depends not surprisingly on the system actually investigated. This is due to the general data structure of the two cases being rather different. In general, much larger shift changes through binding are observed for the ovomucoid–chymotrypsin complex. As a consequence, for the ovomucoid–chymotrypsin complex, all discussed procedures lead to a better prediction of the interface residues. Both examples so far nicely represent two extremes that might occur in protein interaction studies, one with a quite severe overlap of total chemical shift distributions of interacting and non-interacting residues on one hand and well separated distributions on the other hand.

From the plots (Fig. 2) for both protein–protein complexes discussed here it becomes obvious that method **1** including the sign of the chemical shifts, generally provides non-optimal results independent of the definition of the weighting factors. This is also apparent from the data of Tables 4 and 5 where sensitivity and specificity are summarized for cut-off values of σ_0^{corr} and $2\sigma_0^{\mathrm{corr}}$. This mainly is a consequence of a low correlation between chemical shift changes of resonances of interacting residues discussed above (Tables 2, 3). Methods **2** and **3** deliver for the Raf–Ras complex the first false positive residues later (at higher sensitivities), after 4 or 5 correctly spotted amino acids.

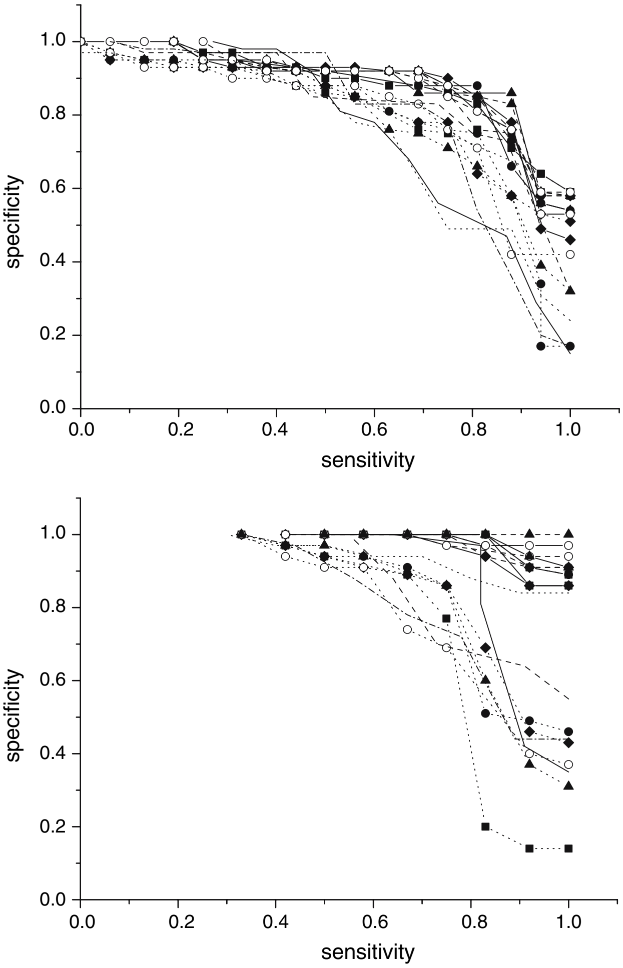

Fig. 2 Specificity and sensitivity of different chemical shift-mapping methods. The specificity and the sensitivity of recognizing interacting residues were calculated as a function of the chemical shift cut-off value and plotted against each other for the different methods 1A–3E. In addition, it is also calculated for the chemical shifts changes of the individual atom types. The upper plot shows the results for the Ras-binding domain (Raf-RBD) of Raf by binding of Ras.Mg^{2+}.GppNHp and the lower one for the turkey ovomucoid third domain in complex with chymotrypsin A$_\alpha$. 1A (■), 1B (◆), 1C (●), 1D (○), 1E (▲) combined with (...), 2A (■), 2B (◆), 2C (●), 2D (○), 2E (▲) combined with (– – –), 3A (■), 3B (◆), 3C (●), 3D (○), 3E (▲) combined with (——), ^{1}H^N (——), ^{15}N (– – –), ^{13}C$^\alpha$ (... ...) and ^{13}C′ (–...–)

At a cut-off value giving the highest sensitivity of 1.0, detecting all interface residues of the Ras–Raf complex, specificity values of 0.63 can be obtained (Fig. 2). For the combined chemical shift changes $\Delta\delta_{\mathrm{comb}}$ in the second example an optimal specificity value of 1.00 at a sensitivity of 1.00 is possible.

The next question to answer is what combination of weighting factors and calculation procedures is best suited. In the sensitivity range up to 0.5 methods 2 and 3 reach quite similar results, giving specificity values of higher than 0.9. A closer look at the plots reveals that in example 2 method 2E using the sum of amino acid specific weighted

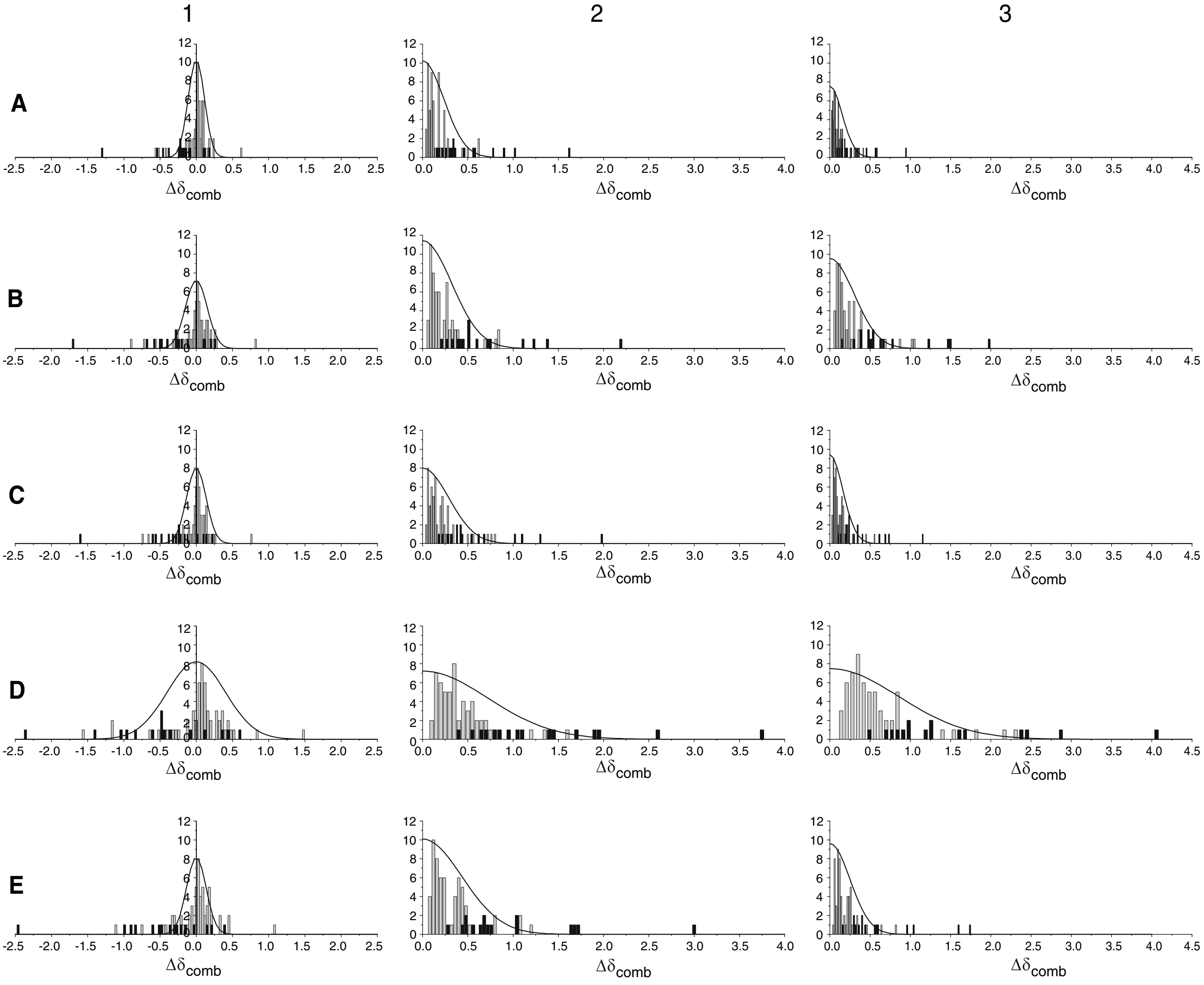

Fig. 3 Distributions of the combined chemical shift perturbations in Ras-binding domain of Raf by binding of Ras.Mg^{2+}.GppNHp. The bars represent frequencies of the calculated combined chemical shift differences $\Delta\delta_{\text{comb}}$ for methods 1A–3E. Residues involved and not-involved in the in the protein–protein interaction are indicated as black and white bars, respectively. Gaussians with one standard deviation to zero σ_0 are depicted. Note that σ_0 was calculated from the chemical shift changes of all residues since in practice the assignments to class C_1 and C_2 is not known

absolute chemical shift changes shows an ideal correlation of sensitivity and specificity, with a specificity of 1 at a sensitivity of 1. The corresponding procedure 3E, applying the square root of the sum of amino acid specific weighted squares of the chemical shift values, has only a specificity of 0.9 at a sensitivity of 1. Looking at example 1 both methods show the best performance of all procedures at a sensitivity of 0.88 with a specificity of 0.83 (2E) and 0.86 (3E), all other methods give a specificity lower than 0.8 at this sensitivity. For other sensitivity/specificity pairs the picture is not apparent but the amino acid type weighting give similar or equal results than other weighting schemes.

Until now we were only speaking about combined chemical shifts being used as a measure for interacting residues in proteins. Why should we combine the shifts and not inspect the chemical shift deviations for every nucleus separately? In Fig. 2 the sensitivity/specificity properties of H^N-, N-, C^α- and C'-shifts are also shown together with the combination of chemical shifts $\Delta\delta_{\text{comb}}$. For both examples separate shifts have a similar predictive power as method 1 but in general a lower one than methods 2 and 3.

Influence of the cut-off value on the quality of the prediction

As mentioned above, the chemical shift distributions of non-interacting residues (class C_2) are sufficiently well described by a normal distribution with a mean of zero (Fig. 1), hence the calculation of the standard deviation to zero σ_0 is reasonable. A practical problem is always that the chemical shift distribution of the interacting residues is not known. If we calculate σ_0 using all residues (class C_1 and C_2), the value is strongly biased by the large chemical shift values of interacting residues. In order to account for that problem we use an iterative procedure to calculate a corrected standard deviation to zero σ_0^{corr}. This is done in the following way: in the first step one calculates σ_0 for all $\Delta\delta_{\mathrm{comb}}$ values, all values outside three times σ_0 (0.5 % of the residues do not belong to the distribution) are removed and a first corrected standard deviation σ_0^{corr} is obtained for the remaining $\Delta\delta_{\mathrm{comb}}$ values. If there are $\Delta\delta_{\mathrm{comb}}$ values present larger than three times of that new σ_0^{corr} value they will be excluded again and another σ_0^{corr} is calculated. This procedure is repeated as often as no $\Delta\delta_{\mathrm{comb}}$ value larger than three times of the actual σ_0^{corr} remains. That final σ_0^{corr} value is taken as cut-off criterion. In Fig. 3 Gaussians with the corrected standard deviation σ_0^{corr} are shown. Looking at the distributions of combined chemical shifts $\Delta\delta_{\mathrm{comb}}$ this still seems to represent a good approximation for methods **1** (note that for clarity the vertical scale of the Gaussians has been set to rather high values). For methods **2** and **3** the lowest shift deviations cannot be perfectly represented by a Gaussian around zero since different small shift changes cannot cancel out because of the suppression of the sign of individual contributions. In addition, the chemical shift distributions of non-interaction residues in the Ras–Raf complex are characterised by two local minima, a feature not found in the ovomucoid–chymotrypsin complex. However, in the absence of other practical alternatives the general description of the distributions using Gaussians appears acceptable.

Taking σ_0^{corr} or $2\sigma_0^{\mathrm{corr}}$ as cut-off value would mean that 31.6 % and 4.5 % of non-interacting residues are expected to have values larger than the cut-off values, respectively. These residues would be wrongly interpreted as interacting. Since Raf contains 59 non-interacting residues one would expect 19 and 2–3 false positively predicted residues, respectively. The number of false negative residues depends strongly on the overlap of the C_1 and C_2 distributions. In the Raf–Ras complex the two distributions are not well separated. There is one residue (Asn71) involved in the interaction that shows almost no chemical shift changes after interaction with Ras. It can barely be recognized by combined chemical shifts; unless a very low threshold value is used resulting in a very low specificity of the prediction.

Another method to assess the quality of predictions is the use of MCC. It should be 1 for an optimal prediction. Plotting the correlation coefficient as a function of multiples of σ_0^{corr}one can obtain the optimal cut-off value according to that criterion for the different methods (Fig. 4). As to be expected the optimal cut-of value is dependent on the data as well as on the definition of the combined chemical shift value. In the two cases optimal values are obtained in the range from 0.9 to 1.2 σ_0^{corr} indicating that the corrected standard deviation to zero is a quite good measure for the cut-off value. However, if the optimal value would be exactly known for a given data set in advance (what is not true in practice) then the performance of the different methods is rather similar. In the given example, method 3B reaches the maximum performance, the methods 3A and 2D are least efficient.

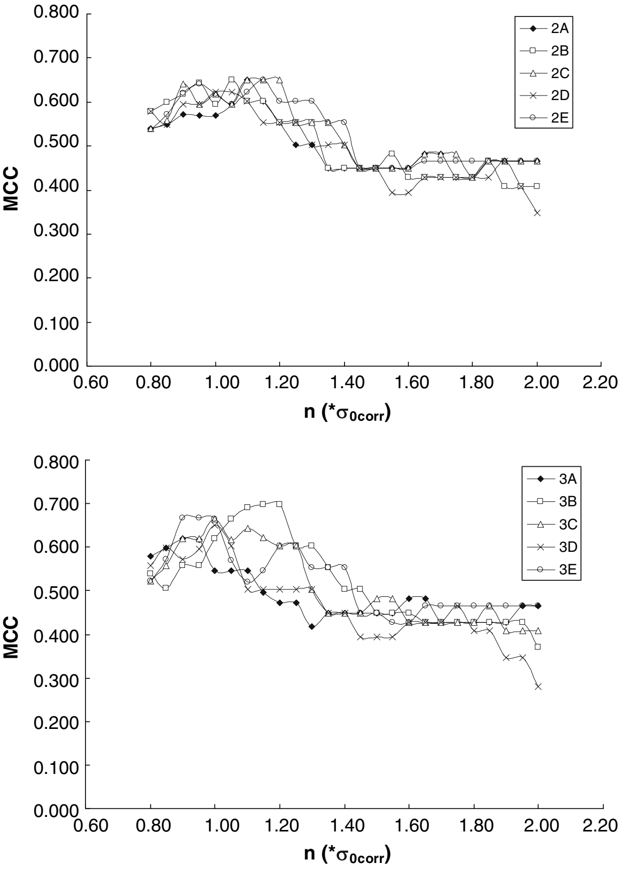

Fig. 4 Quality of the prediction as a function of the cut-off value. The MCC is plotted against *n* times the corrected standard deviation to 0 (σ_0^{corr}) for Raf-RBD using all four nuclei to calculate $\Delta\delta_{\mathrm{comb}}$ (according to Table 4)

Tables 4 and 5 compare specificity, sensitivity, positive predictive value (PPV) and negative predictive value (NPV) obtained for every chemical shift mapping procedure using σ_0^{corr} and $2\sigma_0^{\mathrm{corr}}$ (see also above). For the Raf–Ras complex (Table 4) the combined chemical shift values with a cut-off at σ_0 methods 3B, 3C and 3E gives the highest sensitivity (0.88) together with specificities between of 0.83 and 0.86. For ovomucoid–chymotrypsin complex (Table 5) the combined chemical shift values with a cut-off at σ_0^{corr} methods 2B, 2E, 3B and 3E give the highest sensitivity possible (1.00) with specificities varying

Table 4 The significance of combined chemical shifts for the identification of interacting residues in Ras-binding domain (Raf–RBD) of Raf by binding of Ras.Mg^{2+}.GppNHp[a]

Method	Sensitivity	Specificity	PPV	NPV	MCC
1A	0.63(0.19)	0.86(0.93)	0.56(0.43)	0.89(0.81)	0.47(0.17)
1B	0.56(0.28)	0.86(0.93)	0.53(0.56)	0.88(0.83)	0.42(0.31)
1C	0.63(0.31)	0.86(0.93)	0.56(0.56)	0.89(0.83)	0.47(0.31)
1D	0.63(0.25)	0.83(0.93)	0.50(0.50)	0.89(0.82)	0.42(0.24)
1E	0.56(0.31)	0.86(0.93)	0.53(0.56)	0.88(0.83)	0.42(0.31)
2A	0.75(0.44)	0.86(0.95)	0.60(0.70)	0.93(0.86)	0.57(0.47)
2B	0.75(0.38)	0.88(0.95)	0.63(0.67)	0.93(0.85)	0.60(0.41)
2C	0.81(0.38)	0.86(0.95)	0.62(0.67)	0.94(0.85)	0.62(0.41)
2D	0.75(0.31)	0.90(0.95)	0.67(0.63)	0.93(0.84)	0.62(0.35)
2E	0.75(0.44)	0.86(0.95)	0.60(0.70)	0.93(0.86)	0.57(0.47)
3A	0.69(0.38)	0.88(0.97)	0.61(0.75)	0.91(0.85)	0.55(0.45)
3B	**0.88**(0.38)	0.83(0.93)	0.58(0.60)	**0.96**(0.85)	0.62(0.37)
3C	**0.88**(0.38)	0.86(0.95)	0.64(0.67)	**0.96**(0.85)	**0.67**(0.41)
3D	0.75(0.25)	**0.92**(0.95)	**0.71**(0.57)	0.93(0.82)	0.65(0.28)
3E	**0.88**(0.44)	0.86(0.95)	0.64(0.70)	**0.96**(0.86)	**0.67**(0.47)
^{1}H^{N}	0.53(0.47)	0.81(0.95)	0.42(0.70)	0.87(0.87)	0.35(0.52)
^{15}N	0.80(0.47)	0.74(0.95)	0.44(0.70)	0.93(0.87)	0.47(0.52)
^{13}C$^{\alpha}$	0.63(0.25)	0.76(0.95)	0.42(0.57)	0.88(0.82)	0.34(0.28)
^{13}C′	0.75(0.50)	0.83(0.97)	0.55(0.80)	0.92(0.88)	0.52(0.56)
1A	0.63(0.19)	0.86(0.93)	0.56(0.43)	0.89(0.81)	0.35(0.52)

[a] Sixteen residues are directly involved in the interaction of the Ras-binding domain (Raf-RBD) of Raf with Ras.Mg^{2+}.GppNHp. Shown are the statistical values calculated for every method using σ_0^{corr} and $2\sigma_0^{\text{corr}}$ (in brackets) as cut-off criterion. Sensitivity is defined as the number right positive predicted residues of C_1, divided by the number of right positive and false negative predicted residues of C_1; specificity, number right negative residues divided by the number of right negative and false positive predictions; PPV, number right positive prediction divided by the number of right positive and false positive prediction; NPV, number of right negative predictions divided by the number of right negative and false negative predictions; MCC, Matthews correlation coefficient. The optimal values are highlighted by bold letters

between 0.76 and 0.88. Another quality criterion is the PPV that tells what part of the predicted events is true. A maximum value is obtained for the Raf–Ras-complex for method 3D, in the second example for methods 2A and 2C. The NPV gives information on the quality of the class C_2 prediction. The best values are again obtained for the first example (Table 4) for methods 3B, 3C and 3E, for the second example 2B, 2D, 3B, 3E.

Using the MCC as quality measure for the prediction together with the cut-off value σ_0^{corr} the best value of 0.67 is found for methods 3C and 3E for the Raf–Ras complex (Table 4). For the ovomucoid–chymotrypsin complex an almost perfect correlation of 0.89 is found for methods 2A and 2C (Table 5).

Predictive power for a reduced set of nuclei (^{1}H and ^{15}N)

Since the majority of all chemical shift perturbation studies are based in ^{1}H,^{15} N-HSQC spectra, in many cases only the information of these two nuclei are available. It is to be expected that the prediction power decreases since less information is available. In fact, the Matthews coefficient drops substantially for all methods when only this information is used (Tables 4–6). As shown before, in the Ras–Raf-complex method 3E remains optimal and in the ovomucoid–chymotrypsin complex methods 2A and 2D. However, in the last case method 3D now is also optimal.

Unfortunately, there are not much data found in literature where structural and shift data are available. We included in Table 6 also data from the PDZ2 domain of human phosphatase hPTP1E complexed with a C-terminal peptide from the Fas and from HPr from *E. coli* complexed with N-terminal domain of enzyme I (EIN). Here methods 2A and 3E show the highest predictive power.

Structural interpretations

In the left part of Fig. 5 the interacting residues obtained from the X-ray structure of cRaf1–RBD complexed with Rap1A(E30D,K31E) are shown in red. A crystal structure of the complex with Ras is not available yet. What the Raf-interaction concerns (Nassar et al. 1995) this Ras-like Rap1A mutant (often called Raps) is assumed to be a

Table 5 The significance of combined chemical shifts for the identification of interacting residues in the turkey ovomucoid third domain/bovine chymotrypsin A$_{\alpha}$-complex[a]

Method	Sensitivity	Specificity	PPV	NPV	MCC
1A	0.67(0.33)	0.83(1.00)	0.54(1.00)	0.85(0.81)	0.40(0.52)
1B	0.58(0.25)	0.89(1.00)	0.64(1.00)	0.86(0.83)	0.48(0.45)
1C	0.58(0.25)	0.91(1.00)	0.70(1.00)	0.86(0.80)	0.53(0.45)
1D	0.75(0.25)	0.83(1.00)	0.60(1.00)	0.91(0.80)	0.54(0.45)
1E	0.50(0.25)	0.91(1.00)	0.60(1.00)	0.84(0.80)	0.46(0.45)
2A	0.83(0.50)	**1.00**(1.00)	**1.00**(1.00)	0.94(0.85)	**0.89**(0.65)
2B	**1.00**(0.67)	0.76(1.00)	0.60(1.00)	**1.00**(0.89)	0.68(0.77)
2C	0.83(0.50)	**1.00**(1.00)	**1.00**(1.00)	0.94(0.85)	**0.89**(0.65)
2D	1.00(0.50)	0.89(1.00)	0.75(1.00)	**1.00**(0.85)	0.82(0.65)
2E	**1.00**(0.67)	0.82(1.00)	0.67(1.00)	1.00(0.89)	0.74(0.77)
3A	0.83(0.42)	0.97(1.00)	0.91(1.00)	0.94(0.83)	0.83(0.59)
3B	**1.00**(0.67)	0.82(1.00)	0.67(1.00)	**1.00**(0.89)	0.74(0.77)
3C	0.92(0.58)	0.88(1.00)	0.73(1.00)	0.97(0.87)	0.75(0.71)
3D	0.92(0.42)	0.94(1.00)	0.85(1.00)	0.97(0.83)	0.84(0.59)
3E	**1.00**(0.67)	0.88(1.00)	0.75(1.00)	**1.00**(0.89)	0.82(0.77)
$^{1}H^{N}$	0.82(0.36)	0.90(1.00)	0.75(1.00)	0.93(0.82)	0.73(0.59)
^{15}N	0.82(0.64)	0.70(0.94)	0.47(0.78)	0.92(0.89)	0.48(0.65)
$^{13}C^{\alpha}$	0.90(0.50)	0.84(0.97)	0.64(0.83)	0.96(0.86)	0.71(0.64)
$^{13}C'$	0.56(0.33)	0.88(1.00)	0.56(1.00)	0.88(0.84)	0.55(0.65)

[a] Twelve residues are directly involved in the interaction in the turkey ovomucoid third domain/bovine chymotrypsin A$_{\alpha}$-complex. Shown are the statistical values calculated for every method using σ_0^{corr} and $2\sigma_0^{\mathrm{corr}}$ (in brackets) as cut-off criterion. The definition of the sensitivity, specificity, the PPV and the NPV is explained above. MCC, Matthew correlation coefficient. The optimal values are highlighted by bold letters

well-suited analog for Ras itself since all amino acids in the binding site are identical to that of Ras. This structure was also the basis for a simulation of the complex structure of Raf-RBD with Ras.Mg.GppNHp by Gohlke et al. (2003). Accordingly, in the structure of the complex modelled for Ras the same residues of Raf are interacting with Ras. On the right side the residues are depicted in red that are recognized from the chemical shift changes using σ_0^{corr} (0.435) as the cut-off value. A comparison of the two plots reveals that by using this threshold two residues (Asn71, Lys84) of the interaction surface are not recognized applying method 2E. In addition, a few false positive assignments occur with Phe61 and Asn74 that show a rather large shift changes. Furthermore, residues Ile58, Val60, Leu62, Leu78, Cys81, Leu82 and Ala85 are wrongly recognized from the chemical shift changes alone. However, they can be excluded easily from the predicted class C_1 residues when a 3D structure is available since they have a low solvent accessibility (<5%). The additional residues that are assigned to class C_1 with a threshold of σ_0^{corr} are depicted in yellow in Fig. 5.

In Fig. 3 one can observe a pronounced second maximum of the combined chemical shift distribution using the amino acid specific weighting factors (2E, 3E). This second maximum is present but not so apparent using atom specific weighting factors only. A third maximum at high chemical shift changes is barely visible. The 16 amino acids with the largest chemical shift changes can be assigned to the third peak in the distribution and are shown in red in Fig. 6. Eleven of them are involved in the interaction. The elimination of residues that are not located on the surface of the uncomplexed protein (low-solvent accessibility) but predicted as interacting residues by their combined chemical shifts improves much the specificity. The reason is evident from Fig. 6: almost all residues (Ile58, Val60, Leu62, Leu82) of the third maximum of the distribution which are not interacting directly are located at the back side of β-strand B1 and are thus probably influenced by local conformational changes of this β-strand induced through binding. The second maximum contains 25 amino acids (Fig. 6, green). Five residues (Thr57, Asn71, Lys84, Lys87, Gly90) of them are located in the interaction interface. The remaining residues belong either to a 2nd layer of the interaction sphere, six are solvent inaccessible (Val72, Leu78, Cys81, Ala85, Leu86, Leu126) or are located at the borders of the interaction interface (Asn74-Met76, Asp80, Met83, Leu91, Gln92, Glu94, Ala118, Ile122, Glu124, Gln127, Val128). The main exception is His105 located opposite to the interaction surface.

In example 2 the interaction is provided by a loop region, leading to huge chemical shift perturbations of the involved

Table 6 Quality of the prediction for amide ^{1}H and ^{15}N shift perturbations[a]

Method	Raf	Ovomucoid	PDZ2	HPr
1A	0.31(0.47)	0.62(0.47)	0.31(0.15)	0.25(0.24)
1B	0.33(0.47)	0.68(0.62)	0.27(0.15)	0.25(0.24)
1C	0.33(0.47)	0.68(0.62)	0.24(0.15)	0.24(0.24)
1D	0.33(0.47)	0.47(0.55)	0.26(0.24)	0.27(0.24)
1E	0.35(0.41)	0.69(0.55)	0.23(0.15)	0.14(0.17)
2A	0.44(**0.59**)	**0.82**(0.69)	**0.37**(0.31)	**0.49**(**0.30**)
2B	0.36(**0.59**)	0.75(0.62)	0.34(0.31)	0.44(0.24)
2C	0.38(**0.59**)	0.62(**0.82**)	0.35(0.31)	0.46(0.24)
2D	0.36(0.51)	**0.82**(0.69)	0.23(0.28)	0.46(0.24)
2E	0.38(0.53)	0.59(**0.82**)	0.31(0.31)	0.38(0.24)
3A	0.45(0.47)	0.58(0.75)	0.32(0.35)	0.47(0.24)
3B	0.44(0.49)	0.50(0.62)	0.34(**0.44**)	0.36(0.17)
3C	0.40(0.53)	0.51(**0.82**)	0.35(0.38)	0.38(0.17)
3D	0.44(0.47)	**0.82**(0.75)	0.26(0.41)	0.31(0.17)
3E	**0.46**(0.47)	0.51(**0.82**)	**0.37**(0.36)	0.33(0.17)

[a] Matthews correlation coefficient was calculated for the ^{1}H and ^{15}N-shifts only. Raf, Raf binding domain complexed with Ras.Mg.GppNHp, Ovomucoid, turkey ovomucoid third domain complexed with bovine chymotrypsin A$_\alpha$, PDZ2, PDZ2 domain of human phosphatase hPTP1E with a C-terminal peptide from the Fas receptor, HPr, HPr from *E. coli* complexed with N-terminal domain of enzyme I (EIN). As cut of value σ_0^{corr} and $2\sigma_0^{\mathrm{corr}}$ (in brackets) were used. The optimal values are highlighted by bold letters

residues. The combined chemical shift perturbations show no second maximum and hence no second interaction sphere is observed in comparison to the Ras–Raf complex.

Conclusion and outlook

Possible improvements of the prediction

From the different ways to calculate combined chemical shifts for the prediction of interaction sites, it is clear that method 1 including the sign of the chemical shift changes gives always sub-optimal results. The same is true for method 4. Depending on the actual system studied either the combination of chemical shifts on the basis of the Hamming distance (method 2) or the Euclidian distance (method 3) is slightly superior. Using the Matthew correlation coefficient to assess the quality of the prediction in the Ras–Raf-complex methods 3C and 3E are superior, in the ovomucoid–chymotrypsin complex methods 2A and 2C. This means the weighting on the basis of the gyromagnetic ratio (A), on the basis of the atom specific chemical shift spread (C) and on the basis of the atom and amino acid specific chemical shift spread (E) gives better results from case to case when the shifts from all backbone atoms are used. When only ^{1}H and ^{15}N shifts are used then sometimes the weighting calculated from the data set under consideration (Gröger et al. 2003) (Table 6) also gives optimal results. Only the empirical weighting factors (B) given by Meininger et al. (2000) give never the optimal results in our small data base of interacting proteins, but it is very likely that cases exist where this weighting will be superior. Since the differences in performance of the different methods are small, only a large, unbiased data base (that does not exist yet) could lead to a selection of an optimal method in the statistical sense.

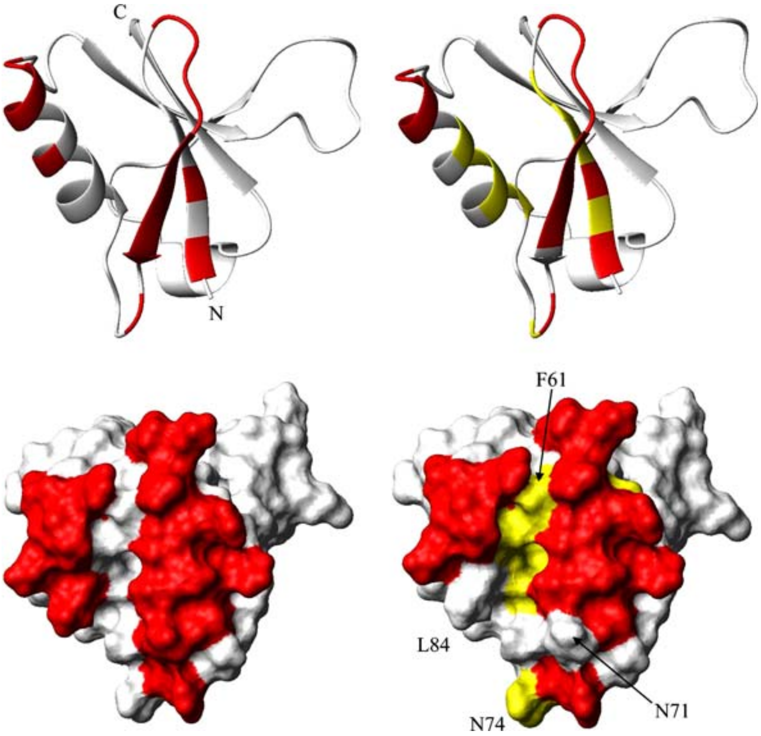

Fig. 5 Contact surfaces of Raf-RBD with Ras.Mg^{2+}.GppNHp. The left side shows the ribbon and surface representation of interacting residues following from the X-ray data and the described criteria (indicated in red). The right column depicts the selected residues (right positive in **red**, and false positive in **yellow**) using the standard deviation σ_0^{corr} as cut-off criterion for method **2E**

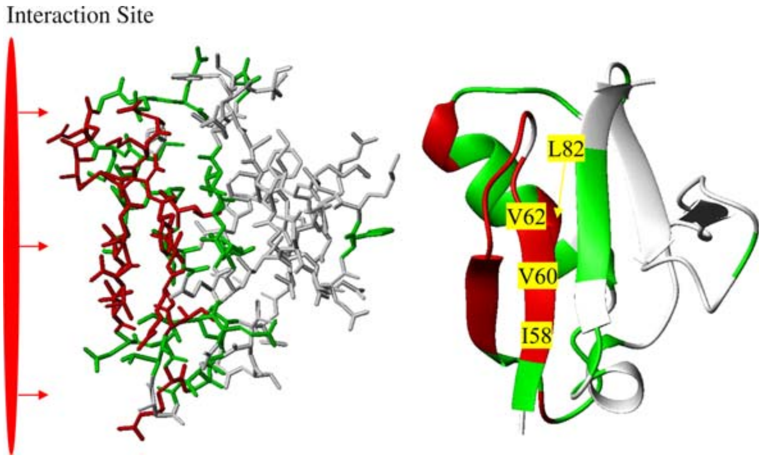

Fig. 6 First and second interacting sphere of Raf-RBD with Ras.Mg^{2+}.GppNHp. Sixteen residues located in the 1st interaction sphere (**red**): 11 of them are recognized by method 2E are true positive (Arg59, Asn64-Val70, Arg73, Val88, Arg89), 4 solvent inaccessible (Ile58, Val60, Leu62, Leu82), 1 false positive (Phe61). 25 residues are in the 2nd sphere (**green**): 5 are right positive (Thr57, Asn71, Lys84, Lys87, Gly90), 6 solvent inaccessible (Val72, Leu78, Cys81, Ala85, Leu86, Leu126) and 14 not bound (Asn74-Met76, Asp80, Met83, Leu91, Gln92, Glu94, His105, Ala118, Ile122, Glu124, Gln127, Val128). The chemical shifts of the residues depicted in white are not significantly perturbed

Using the iterative procedure described above to obtain the corrected standard deviation to zero (σ_0^{corr}) as cut-off value gives a better representation of the distribution of non-interacting residues (class C$_2$), since large $\Delta\delta_{\mathrm{comb}}$ values are eliminated being expected to be part of the interaction sphere. This cut-off criterion ensures to pick the majority of interacting residues (sensitivity of more than 0.88 and up to 1.00) with a reasonable specificity. It represents so far a quite ideal cut-off criterion serving as threshold for the two rather different examples described here (Fig. 4).

However, it is obvious from the histograms shown here that from the chemical shift changes alone, all interacting residues cannot be identified safely since the distributions of classes C$_1$ and C$_2$ overlap severely. As a consequence, in many cases a considerable part of the predicted interacting residues is not involved in interaction and vice versa. Therefore, a clear result is that the quality of the prediction of interacting residues from chemical shift changes is surprisingly low and that the performance further decreases when only the ^{1}H^N and ^{15}N shifts are used for the prediction.

It is obvious that additional information is required for a more reliable prediction. The results can be improved when external information is added. (1) As far as the 3D structure of the protein under investigation is known one can remove residues with low water accessibility, although interaction induced conformational changes can possibly expose internal residues to the interaction partner. (2) Drawing all residues on the 3D structure can help to exclude possible false positive residues, which are too distant from the main surface patch. These filters are also applied by the protein docking program HADDOCK (Dominguez et al. 2003). (3) Another information for excluding wrong positives is the fact that in a simple 1 to 1 binding mode the K_{D} calculated from the combined chemical shifts should be equal in the limits of error. When titration data a evaluated with Eqs. 15–19 the smallest K_{D} obtained is most probably the correct one and only residues with the same K_{D} are located in the primary interaction site.

Acknowledgment This work was supported by the European Union (Structural Genomics in Europe), the Fonds der chemischen Industrie, the Deutsche Forschungsgemeinschaft and the BMBF.

References

Asakura T, Taoka K, Demura M, Williamson MP (1995) The relationship between amide proton chemical shifts and secondary structure in proteins. J Biomol NMR 6(3):227–36

Baldi P, Brunak S (2001) Bioinformatics. The machine learning approach, 2nd edn. MIT Press, London, England, pp 158, 159

Dominguez C, Boelens R, Bonvin AMJJ (2003) A Protein–protein docking approach based on biochemical or biophysical information. J Am Chem Soc 125:1731–1737

Ekiel I, Banville DL, Shen SH, Slon-Usakiewicz JJ, Koshy A, Gehring K (1998) Main-chain signal assignment for the PDZ2 domain from human protein tyrosine phosphatase hPTP1E and its complex with a C-terminal peptide from the Fas receptor. J Biomol NMR 12:455–456

Farmer BTII, Constantine KL, Goldfarb V, Friedrichs MS, Wittekind M, Yanchunas JJ, Robertson JG, Mueller L (1996) Localizing the NADP+ binding site on the MurB enzyme by NMR. Nat Struct Biol 3:995–997

Fujinaga M, Sielecki AR, Read RJ, Ardelt W, Laskowski M Jr, James MNG (1987) Crystal and molecular structures of the complex of α-chymotrypsin with its inhibitor Turkey ovomucoid third domain at 1.8 Å resolution. J Mol Biol 195:397–418

Garrett DS, Seok Y-J, Peterkofsky A, Gronenborn AM, Clore GM (1999) Solution structure of the 40,000 Mr phosphoryl transfer complex between the N-terminal domain of enzyme I and HPr. Nat Struct Biol 6:166–173

Geyer M, Herrmann C, Wohlgemuth S, Wittinghofer A, Kalbitzer HR (1997) Structure of the Ras-binding domain of RalGEF and implications for Ras binding and signalling. Nat Struct Biol 4:684–698

Gohlke H, Kiel C, Case DA (2003) Insights into protein–protein binding by binding free energy calculation and free energy decomposition for the Ras–Raf and Ras–RalGDS complexes. J Mol Biol 330:891–913

Gröger C, Möglich A, Pons M, Koch B, Hengstenberg W, Kalbitzer HR, Brunner E (2003) NMR-spectroscopic mapping of an engineered cavity in the I14A mutant of HPr from *Staphylococcus carnosus* using Xenon. J Am Chem Soc 125:8726–8727

Heitmann B, Maurer T, Weitzel JM, Strätling WH, Kalbitzer HR, Brunner E (2003) Solution structure of the matrix attachment region-binding domain of chicken MeCP2. Eur J Biochem 270:3263–3270

Kozlov G, Gehring K, Ekiel I (2000) Solution structure of the PDZ2 domain from human phosphatase hPTP1E and its interactions with C-terminal peptides from the Fas receptor. Biochemistry 39:2572–2580

Li H, Yamada H, Akasaka K (1998) Effect of pressure on individual hydrogen bonds in proteins. Basic pancreatic trypsin inhibitor. Biochemistry 37(5):1167–73

Meininger DP, Rance M, Starovasnik MA, Fairbrother WJ, Skelton NJ (2000) Characterization of the binding interface between the e-domain of *Staphylococcal* protein A and an antibody Fv-fragment. Biochemistry 39:26–36

Mulder FA, Schipper D, Bott R, Boelens R (1999) Altered flexibility in the substrate-binding site of related native and engineered high-alkaline Bacillus subtilisins. J Mol Biol 292(1):111–123

Nassar N, Horn G, Herrmann C, Scherer A, McCormick F, Wittinghofer A (1995) The 2.2 Å crystal structure of the Ras-binding domain of the serine/threonine kinase c-Raf1 in complex with RaplA and a GTP analogue. Nature 375 (6532): 554–560

Pervushin K, Riek R, Wider G, Wüthrich K (1997) Attenuated *2T* relaxation by mutual cancellation of dipole–dipole coupling and chemical shift anisotropy indicates an avenue to NMR structures of very large biological macromolecules in solution. Proc Natl Acad Sci USA 94:12366–12371

Song J, Markley JL (2001) NMR chemical shift mapping of the binding site of a protein proteinase inhibitor: changes in the 1H, 13C and 15N NMR chemical shifts of turkey ovomucoid third domain upon binding to bovine chymotrypsin A_{α}. J Mol Recognit 14:166–171

Terada T, Ito Y, Shirouzu M, Tateno M, Hashimoto K, Kigawa T, Ebisuzaki T, Takio K, Shibata T, Yokoyama S, Smith BO, Laue ED, Cooper JA (1999) Nuclear magnetic resonance and molecular dynamics studies on the interactions of the Ras-binding domain of Raf-1 with wild-type and mutant Ras proteins. J Mol Biol 286:219–232

Wagner G, Pardi A, Wüthrich K (1983) Hydrogen bond length and proton NMR chemical shifts in proteins. J Am Chem Soc 105:5948–5949