J Biomol NMR (2012) 54:43–51
DOI 10.1007/s10858-012-9649-y

ARTICLE

Automated protein backbone assignment using the projection-decomposition approach

Jonas Fredriksson · Wolfgang Bermel · Doroteya K. Staykova · Martin Billeter

Received: 16 April 2012 / Accepted: 22 June 2012 / Published online: 18 July 2012

Abstract Spectral projection experiments by NMR in conjunction with decomposition analysis have been previously introduced for the backbone assignment of proteins; various pulse sequences as well as the behaviour with low signal-to-noise or chemical shift degeneracy have been illustrated. As a guide for routine applications of this combined tool, we provide here a systematic analysis on different types of proteins using welldefined run-time parameters. As a second result of this study, the backbone assignment module SHABBA was extensively rewritten and improved. Calculations on ubiquitin yielded again fully correct and nearly complete backbone and CHβ assignments. For the 128 residue long azurin, missing assignments mostly affect Hα and Hβ. Among the remaining backbone (plus Cβ) nuclei 97.5 % could be assigned with 1.0 % differences to a reference. Finally, the new SHABBA algorithm was applied to projections recorded for a yeast histone protein domain at room temperature, where the protein is subject to partial unfolding: this leads to unobservable resonances (about a dozen missing signals in a normal 15N-HSQC) and extensive degeneracy among the resonances. From the clearly observable residues, 97.5 % of the backbone and CHβresonances could be assigned, of which only 0.8 % showed differences to published shifts. An additional study on the protein MMP20, which exhibits spectral difficulties to an even larger extent, explores the limitations of the approach.

J. Fredriksson · D. K. Staykova · M. Billeter (✉)
Department of Chemistry and Molecular Biology, University of Gothenburg, Box 462, 405 30 Gothenburg, Sweden
e-mail: martin.billeter@chem.gu.se

W. Bermel
Bruker BioSpin GmbH, Silberstreifen, 76287 Rheinstetten, Germany

Keywords Algorithm · Automated resonance assignment · Decomposition · Spectral projections

Introduction

Sparse sampling, for example in the form of projection spectroscopy, serves the needs for speeding up NMR characterisation on proteins and achieving very high dimensionality while maintaining high resolution in each dimension (Kupče and Freeman 2008; Billeter and Staykova 2009; Kazimierczuk et al. 2010; Orekhov and Jaravine 2011). Application areas include efficient large scale protein characterisations (structure, dynamics, interactions etc.) as well as studies of intrinsically disordered or partly denatured proteins (Uversky and Dunker 2010). The PRODECOMP-SHABBA approach for backbone assignment accepts spectral projections from one or several experiments, together with a list of peaks from a ^{15}N-HSQC spectrum; the latter is transformed into a list of intervals along the directly detected dimension for individual decomposition calculations (Malmodin and Billeter 2005; Staykova et al. 2008a, b). These calculations determine for a set of connected (typically J-coupled) spins (a "spin system") the spectral traces along each dimension. As an illustration consider a 3D ^{15}N-NOESY-HSQC: A decomposition yields for each N–H moiety a set of three traces, one for the HN-dimension with one peak, one for the N-dimension with also one peak, and one for the NOE-dimensions with peaks for every proton involved in a NOE interaction with the given HN. The traces are referred to as shapes, and each set of shapes for a given N–H moiety is called a component (Orekhov et al. 2001). The multi-way decomposition implemented in PRODECOMP is a general procedure applicable to various types

of projection experiments (Malmodin and Billeter 2005), e.g. experiments used for backbone or side-chain assignment, or for extraction of NOE distance limits, or any combination of these. On the other hand, the analysis of the resulting components depends on the type of the NMR experiments. Similar to earlier studies (Staykova et al. 2008a), we combine here two experiments for backbone assignments for simultaneous decomposition by PRODECOMP.

Here, we present an improved version of the backbone assignment tool SHABBA and demonstrate its efficiency and robustness on different proteins. The general relation of PRODECOMP-SHABBA to other approaches has been addressed previously (Staykova et al. 2008a). Improvements from the earlier version include a new peak picker and a more sophisticated sequential assignment procedure, which is suitable also for more difficult proteins, e.g. proteins with unusually high extent of chemical shift degeneracy. While the earlier published ubiquitin is shortly mentioned, the larger azurin serves for illustrating the novel improvements of the algorithm. The GI domain of yeast histone H1 (Ali et al. 2004), which was measured at partly denaturing temperatures, shows the behaviour of the approach in the presence of severe spectral overlap. Finally, the limitations of the approach are explored on a protein, MMP20, which exhibits extensive difficulties due to unobservable signals, and which resisted complete assignment also with a conventional approach (Arendt et al. 2007).

Methods

Algorithms

The first part of the projection-decomposition approach for backbone assignment, implemented in the program PRODECOMP, has been described earlier in detail (Staykova et al. 2008a). In short, it consists of obtaining reduced dimensionality experiments, splitting of the resulting data into individual two-dimensional projections, which are then Fourier transformed. Next, these projection spectra are jointly decomposed (Staykova et al. 2008a), yielding for each non-proline residue a nine-dimensional component that in turn consists of one-dimensional shapes for the following eleven nuclei types (*i-1* indicates the residue preceding the H_N nucleus of the directly detected dimension): H_N, N, CO(*i-1*), Cα/Cβ(*i-1*), Hα/Hβs(*i-1*), Cα, Cβ, Hα and Hβs. The latter four shapes may contain, besides signals for *i* nuclei, also signals for *i-1* nuclei; Cα/Cβ(*i-1*) means that signals for both Cα and Cβ of residue *i-1* occur in a single shape, and similar for Hα/Hβs(*i-1*). The decomposition calculation is not specific to any particular type of protein characterisation, e.g. backbone or side-chain assignment or structural studies.

However, the second step, the analysis of the components and their shapes, differs for each type of protein characterisation. Thus, a specific tool called SHABBA was introduced for backbone assignment (Staykova et al. 2008a). In short, it consists of glycine detection based on the absence of signals from β-nuclei, followed by sequential connections of components using correlations between the Cα/Cβ(*i-1*) shape of one component and the Cα and Cβ shapes of a second (and similar for α and β hydrogens). The resulting chains of components are an important intermediate result (referred to as "chains"). It has long been known that chemical shifts of both Cα and Cβ nuclei of proteins exhibit a strong correlation with amino acid types (Grzesiek and Bax 1993), these correlations have been used in many applications to identify amino acid types or to position spin systems on the protein sequence. In a similar way, the above chains of components are positioned on the protein sequence by comparing the Cα and Cβ chemical shifts of the components with expected chemical shifts for the protein sequence based on statistical values for each residue type compiled at the BMRB (Ulrich et al. 2007). The differences between component shifts and expected (BMRB) shift are collected for an entire chain of components and expressed as the root of the mean of their sum (RMSD). A final peak picking yields the assignment.

This earlier described SHABBA procedure was not sufficient for more demanding proteins, as it could not handle errors in preliminary chains of components; errors occur typically when a component for a residue next to a proline (for which no component can be observed) seeks a connection to another component, and thus chains located next to prolines or the N- or C-terminus may erroneously become connected. A flow-chart of the new SHABBA procedure is presented in Fig. 1. Detection of errors in the chains of components is achieved by considering also shortened chains when sliding a component chain along the protein sequence (Fig. 1). A chain is shortened stepwise by removing one component at a time, first from one end and then from the other end. Significant improvements of the match between the experimental shifts and the BMRB estimates for a shortened chain with respect to the full-length chain lead to acceptance of the shortened chain. The strategy used relies on the standard deviations of chemical shifts reported in the BMRB, averaged over all residue types (these values may depend on updates at the BMRB, we used values of 1.71 ppm for Cα and 2.06 ppm for Cβ). Considering only Cβ, a full-length chain is accepted without attempt of shortening if the RMSD of the match is already $<$2.06 ppm. Shortening of a chain must improve the RMSD by at least this same value. Furthermore,

shortened chains are directly accepted once their RMSD drops below 2.06 ppm (this in order to avoid shortening to only a few components). Finally, at the end of iteration 3 (see below; at this stage only a few unassigned chains of components remain, all with typical lengths of 1–3 components), user interaction may be required to fill these into the remaining few and short gaps. Typically shortening of a chain occurs when a sizeable but false correlation, caused by shift similarities of two components, connects a chain to one or a few additional components; the full-length chain may exhibit a good average fit due to the many correct components, but removing one or a few incorrect terminal components will further improve this fit. Thus, each component chain is simultaneously positioned along the protein sequence and, if necessary, shortened. Shortening of the chains also allows splitting of erroneous chains such that each fragment may subsequently be properly positioned.

The new algorithm includes three iterations consisting of correlation calculations and sliding (Fig. 1). The first iteration is used to identify chains that are positioned next to prolines or the N- or C-terminus. Components located after a proline or the N-terminus are marked for zeroing of their correlation to potential preceding components in subsequent iterations. A similar treatment is implemented for components located before a proline or the C-terminus.

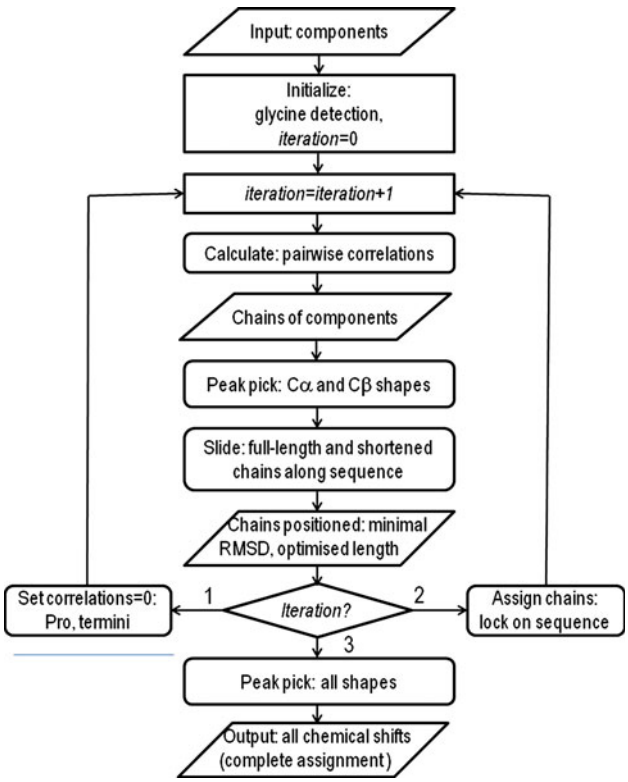

Fig. 1 Flow-chart illustrating the novel SHABBA backbone assignment protocol. Details for each step are provided in the text

After iteration 2, component chains with sufficiently low RMSD at their best position are sequence-specifically assigned; colliding assignments that may occur in this step are resolved by cutting off the conflicting components. For the remaining calculations of iteration 3 this means that their internal correlations are raised to 100 %, and that these chains are positioned on the protein sequence prior to the sliding of the remaining chains. Following iteration 3, all shapes are subjected to a final peak picking.

Besides the combination of chain sliding and chain shortening, additional improvements of the new SHABBA version include more accurate routines for glycine detection and peak picking, which also require fewer input parameters. Specifically, the use of chemical shift ranges obtained from the BMRB improved both detecting of Cα signals (in the Cβ-shapes of glycines) and avoiding of false peaks in general. More reliable identification of α- and β-signals was achieved by simultaneous consideration of shapes from neighbouring components providing higher reliability. Glycine detection could be simplified by dropping the requirement on observation of many large peaks in the shapes for α-nuclei (Staykova et al. 2008a, b). The following is a description of all parameters that govern the assignment as delineated in Fig. 1. (a) Iterations 1 and 2 consider only chains of components of minimal length six; shorter ones, which often have exhibit less variation of chemical shifts and are thus more difficult to position uniquely, are treated in iteration 3. (b) All correlations below 20 % are set to zero in the screening of the correlation matrix (Staykova et al. 2008a); this parameter was raised to 35 % for the GI domain of histone H1 (see "Results"). (c) The new peak picking routine, applied to the Cα and Cβ shapes prior to the sliding of component chains as well as for the final peak picking of all shapes, only requires parameters for noise determination; these are based on the standard deviation of the intensity distribution of all but the largest points in the corresponding shapes. (d) A penalty value of 50 ppm for pairings of a glycine with a non-glycine, or a proline with a non-proline is used for the RMSD calculations, i.e. whenever a sliding step gives rise to such a pairing, 50 ppm replaces the difference between the shift from the shape and the shift estimate from the BMRB; this corresponds to the maximal contribution of about 50 ppm occurring when a low chemical shift typical for an alanine is compared to a high chemical shift typical for a serine or threonine. For an objective evaluation of SHABBA and in order to allow comparisons of the applications to different proteins, none of these parameters, except for parameter (b) for the histone domain assignment, were varied among the different protein assignments. Also the option of manually adding or removing glycine components following their automated detection was not used.

Protein samples and NMR spectroscopy

For each protein, two projection experiments were performed: 5D HBHACBCACONH and 4D HBHACBCANH; the corresponding magnetisation pathways were presented in Grzesiek and Bax (1992, 1993). The faster recording technique proposed in Staykova et al. (2008a) was applied, reducing significantly the measurement time. The maximal number of 2D planes after splitting, p, is 40 for a 5D and 13 for a 4D experiments (for MMP20, not all planes were recorded) as calculated with the following expression:

$$\mathrm{p} = \sum_{k=1}^{n} \frac{n!}{(n-k)!k!} .2^{k-1} = \frac{3^n - 1}{2}$$

where n is the number of indirect dimensions. Each term in the sum provides the number of projections obtained with evolution on exactly k nuclei. This allows for example to determine the number of projections with evolution on all indirect nuclei; these have the lowest S/N and one may wish to omit their recording and decomposition (as we did for MMP20). Spectral data for ubiquitin were those described earlier (Staykova et al. 2008a). Protein concentrations were below 1.0 mM for azurin (exact value unknown due to sealed sample) and about 2 mM for ubiquitin and the histone domain, and all experiments were recorded at 298 K and with 60 complex points. 32 scans were used for azurin, yielding a measurement time per plane of 1.4 h. Corresponding numbers for the histone domain were 16 scans and 0.7 h. For MMP20, the first experiment was recorded as for azurin, whereas the number of scans was increased to 80 for the second experiment due to inherent problems with this protein (see "Results"), yielding 3.4 h per plane. All experiments were run at 600 MHz except for MMP20, where a 900 MHz instrument with cryoprobe was used. Comparison of the resulting chemical shifts with literature values were based on the BMRB entries 6,457 and 6,466 for ubiquitin, 6,161 for the histone GI domain, and 15,361 for MMP20; chemical shifts values reported in Leckner (2001) were used for azurin. The new SHABBA routines are available from the authors.

Results

Ubiquitin

Human ubiquitin is a 76 residue protein including three prolines, two of which are sequential neighbours, and six glycines (Hershko and Ciechanover 1998). As described earlier, 72 components resulting from decomposition of 30 projections were used, corresponding to all non-proline residues except for the (in a normal ^{15}N-HSQC) invisible Glu 24, plus an additional glycine present in our sample (Staykova et al. 2008a). The glycine detection in SHABBA yielded components for all six protein glycines plus an additional component. Since no difficulties were expected for this protein, and all but one component chains obtained in the first round of the correlation calculations exceeded the minimal length of six components normally required in iteration 1 (Fig. 1), this requirement was dropped in the hope to achieve a final sequential assignment in a single iteration. This was indeed the case, and the final peak picking in the component shapes yielded a complete and correct chemical shift table except for missing assignments for Hα of Leu 9 and Cβ of Thr 15, and no Hβs could be observed for threonines 7, 9 and 22 (in addition to those missing due to the invisible Glu 24; see also Table 1).

Azurin

The 128 residue long azurin contains four prolines and eleven glycines (Parr et al. 1976). For each of the 123 backbone peaks in the ^{15}N-HSQC, a decomposition interval was created. Glycine detection among the resulting components correctly detected ten glycines, missing the component for Gly 116. For the assignment, all three iterations of the flow-chart of Fig. 1 were performed. The assignment progress is illustrated in Fig. 2.

In the first iteration, the correlation calculation yielded nine chains of components, of which six reached the required minimal length of six components (thus, these six chains of components considered in this iteration span six or more residues). Peak picking of Cα and Cβ shifts in these components yielded 99, resp. 104, chemical shifts. Sliding the sequences with these chemical shifts along the expected chemical shifts, defined by the BMRB, for the azurin sequence (see "Methods") yielded the result shown by the top line of Fig. 2. Using the shortening mechanism for component chains and the RMSD calculations defined in Methods, the lowest RMSD values for the Cβ shift sequences all yielded correct assignments for either the full chain (second and last chain in the top line of Fig. 2) or for shortened chains (all others). One component chain contained two correct fragments: shortening of this chain from its N-terminus provided a component chain that yielded a low RMSD when positioned onto residues 55–62, while shortening from its C-terminal yielded a chain matching residues 63–72 (Fig. 2). Accordingly, a cut similar to those cuts next to prolines was introduced. Similar sliding for the Cα shift confirmed all results obtained for Cβ, except for one chain, for which no assignment suggestion resulted. The cutting points for chains with removed components are indicated by tilted arrows in Fig. 2. The vertical lines in the figure indicate that a terminal component of a chain has been assigned to a residue next to a proline or the N- or

Table 1 Missing assignments and differences to reference lists

Protein	Length	Missing assignments	Difference to reference[a]
Ubiquitin	76	Cβ 15; Hα 9; Hβ 7,9,22	–
Azurin	128		N 119; CO 69;
		Cα 37,41,48,56,64,82,88,98,116;	Cα 55,118;
		Cβ 21,46,51,61,82,93;	Cβ 27,109,121
		Hα 4,12,14,17,21,23,29,30,34,46,	Hα 84,87,117,126;
		48,52,85,96,98,107,108,113,118,120	
		Hβ 3,44,49,52,56,60,113,118,128	Hβ 31,43,61,62,84
Histone[b]	93	N 33;	
		Cα 54,66; Cβ 19,36;	
		Hα 36,45; Hβ 36,61	Hα 34; Hβ 57,70

[a] Differences do not necessarily reflect erroneous assignments. In particular for azurin, some of the current assignments presented here could be shown to be correct (see also text). Description of the references are given in the text

[b] Only residues 40–50, 54–60, 63–77, 81–83, 86–117, 122–130 (see text)

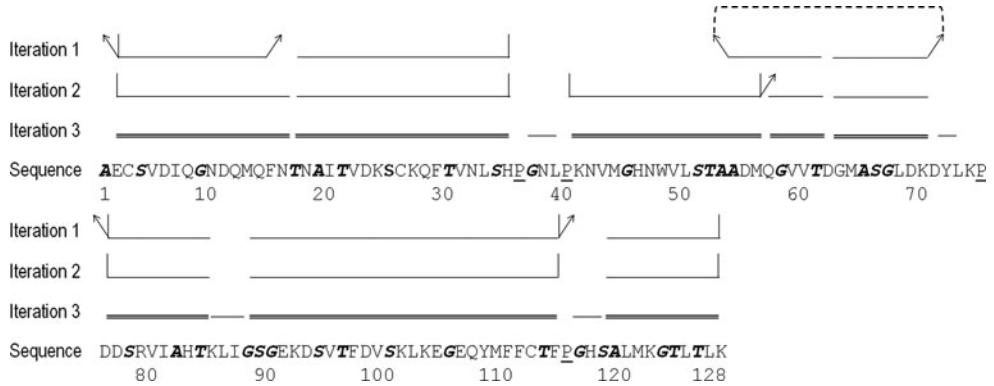

Fig. 2 Backbone assignment results for azurin of the three iterations of the flow-chart of Fig. 1. In the lines labelled "sequence", prolines are highlighted by *underlining*, and residue types with characteristic (or missing) Cβ chemical shifts (Gly, Ser, Thr, Ala) by *italic bold* fonts. For each of the three iterations, *horizontal lines* are component chains positioned on the protein sequence according to lowest RMSD (see "Methods"). *Arrows* indicate parts of component chains that are inconsistent with a low RMSD and were removed by shortening the chains. *Vertical lines* mark points were breaks are set in terms of zero correlations (see text). The *dashed* loop connecting chains 3 and 4 in the result for iteration 1 connects two fragments of the same component chain that could individually be positioned by shortening the chain either from its N-terminus or its C-terminus (see text). *Double lines* for iteration 3 indicate component chains assigned and locked after iteration 2. Thus, during iteration 3, only four chains of 2 or 3 components need to be assigned (e.g. one between Pro 36 and Pro 40)

C-terminus. As a consequence, correlations of these components are set to zero in the next iteration as described in "Methods". However, no final assignments are made during this first iteration.

Iteration 2 yielded, due to the zeroing of certain correlations as a result of iteration 1, twelve component chains, of which eight reached the minimal length of six required for this iteration. Chain sliding using chemical shift sequences, both Cβ and Cα, provided final assignments for the following residues: 2–16, 17–35, 41–56, 57–62, 63–72, 76–84, 88–114 and 119–128 (see Fig. 2, line 2). These assignments were locked in iteration 3, both during the correlation calculations and the chain sliding.

Finally, iteration 3 needed to assign the four remaining short component chains (less than six components), of which three had a length of three components and one a length of two components. Due to unique Cβ patterns involving glycine, serine, threonine and alanine, two chains could be directly assigned to two of the remaining four unassigned stretches in the protein sequence, and the remaining two chains were subsequently placed according to their length (three respectively two components).

Following this complete and correct positioning of chains of components on the protein sequence, a final peak picking of all shapes resulted in complete assignments list for H_N, N and CO. The nuclei of the α and β groups of prolines or residues preceding these were not reliable due to the missing (*i-1*) shapes. Among the other residues, chemical shifts were missing for 9 Cα, 6 Cβ and 20 Hα; for 9 residues no Hβ shift could be determined. To our

knowledge, no comprehensive list of chemical shifts has been deposited at the BMRB for azurin. We therefore compared our results to chemical shifts presented in a Ph.D. thesis (Leckner 2001); this list was obtained at a higher temperature, which may partly explain differences resulting from this comparison (e.g. we could confirm the correctness of our chemical shift for CO of Asp 69, which differs from the reference by 2.5 ppm). Differences include the shifts for N of Ala 119, CO of Asp 69, two Cα, three Cβ, four Hα and five Hβ (in the case of $C\beta H_2$ groups, an assignment was considered successful if at least one Hβ reported in the reference was identified). The large number of missing Hα and Hβ is an indication that signal-to-noise is low. Ignoring the aliphatic hydrogens, 97.5 % of the H_N, N, CO, Cα and Cβ nuclei could be correctly assigned, and 1 % of the resulting chemical shifts differ from the reference (these may still be correct, see above). These results are summarized in Table 1.

GI domain of histone H1

The GI domain of yeast histone H1 (henceforth called "histone domain") consists of 93 residues (residues 38–130), including extensive flexible tails (Ali et al. 2004). Backbone assignments were attempted using projection spectra recorded at room temperature (298 K), where the protein is partially unfolded resulting in extensive chemical shift degeneracy. Already a previous assignment at the lower temperature of 288 K exhibited significant chemical shift ambiguities: as Table 2 summarizes, several groups of 3–5 residues have in the published assignment (Ali et al. 2004) nearly identical shifts simultaneously for Cα, Cβ, Hα, and at least one Hβ (while they usually differ measurably in their N and H_N chemical shifts). The correlation calculation for sequential assignment in our approach relies on component shapes describing Cα, Cβ, Hα and Hβ, i.e. the atoms listed in Table 2. The extent of chemical shift degeneracy is likely to increase with the higher temperature used in the present experiment. Components from our nine-dimensional decomposition cover nuclei from two neighbouring residues. Even when considering α- and β-nuclei from two neighbouring residues, the chemical shift degeneracy remains in several cases. A typical example is the residue pair 40 and 128, both glutamic acids. The components for these two residues look very similar in all shapes of indirectly detected nuclei (Fig. 3). The only sizeable shift difference for this residue pair, 0.16 ppm, occurs for the H_Ns.

As a result from the above, it became obvious that a complete assignment would not be achievable. In addition, signals at expected locations according to Ali et al. (2004) in the ^{15}N-HSQC were unobservable. The 93 residue long domain contains five prolines; thus, removing also the N-terminal residue, 87 spin systems with a backbone H_N are expected. Careful peak picking in a normal ^{15}N-HSQC provided only 74 signals. These formed the starting point for the assignment by defining 74 decomposition intervals. The parameter defining a lower allowed limit for acceptable correlations between components was increased from 20 to 35 % relative to the ubiquitin and azurin assignments in order to adapt to overall higher correlations observed among the components of this histone domain; thus, the average correlation observed between all pairs of components for the histone domain was about three times as large as the corresponding average for azurin; consequently the use of 20 % yielded a few very long chains of components, which were incorrect combinations of shorter, correct chains.

Iteration 1 (see Fig. 1) yielded ten chains, seven of which were longer than the minimal length of six required in this iteration. Sliding of chemical shifts from these component chains along the expected chemical shifts for the protein sequence is discussed here only for Cβ shifts, since the Cα shifts gave no useful result due to the extensive overlap. Five chains were positioned without need for removing components; the two others obviously contained erroneous connections as illustrated by their best RMSD values that exceeded the limit of 2.06 ppm (see "Methods") more than 10-fold. One assigned chain covered the C-terminus of the protein, and two were positioned next to prolines. This allowed zeroing in the correlation matrix of the next iteration for three components. The same chains as in iteration 1 were obtained in iteration 2 except that one of the two chains with erroneous connections resulted now in two chains, both of which could be assigned unambiguously. At this point, assignments were achieved for the following protein residues: 40–48, 54–60, 68–73, 86–91, 92–103, 110–117 and 122–130. Due to colliding assignments, conflicting components at the end of the third chain were removed during the assignment to the protein sequence. Of the twelve chains of components from the last iteration, seven were already assigned in iteration 2. The RMSD values for three of the remaining five chains allowed only for one position each along the protein sequence (due to patterns defined by glycine, serine, threonine, and alanine). After this step, also the positioning of the final two chains of length two and three became unique.

The final sequential assignment thus revealed gaps in the sequence, for which no components were observed due to their complete overlap with other components; the assigned fragments consist of residues 40–50, 54–60, 63–77, 81–83, 86–117, 122–130 (this includes three prolines). Due to the extensive shift degeneracy, α- and β-nuclei were accepted as safe chemical shifts only when fulfilling the following condition: peak picking must be possible for both the shapes of the same residue as well as the corresponding (*i*-

Table 2 Residues in the GI domain of yeast histone H1 with similar chemical shifts at 288 K (Ali et al. 2004)

Group[a]	Residues	$d_{max}(C\alpha)$[b]	$d_{max}(H\alpha)$[b]	$d_{max}(C\beta)$[b]	$d_{max}(H\beta)$[b]	Types[c]
1	39,125,127	0.47	0.04	0.11	0.04	All Lys
2	40,41,122,126,128	0.28	0.04	0.19	0.05	All Glu
3	118,119,124,125	0.37	0.02	0.06	0.05	All Lys

[a] Residues are grouped together if their Cα and C*β* chemical shifts differ by <0.5 ppm, and their Hα and one of the H*β* chemical shifts differ by <0.05 ppm in the published assignment (Ali et al. 2004)

[b] $d_{max}(X)$ lists the maximal chemical shift difference within each group for nucleus X in ppm. Only the closest pair of shift for H*β* is considered

[c] Note that of the 93 residues of the histone domain, 18 are lysines and 11 are glutamic acids

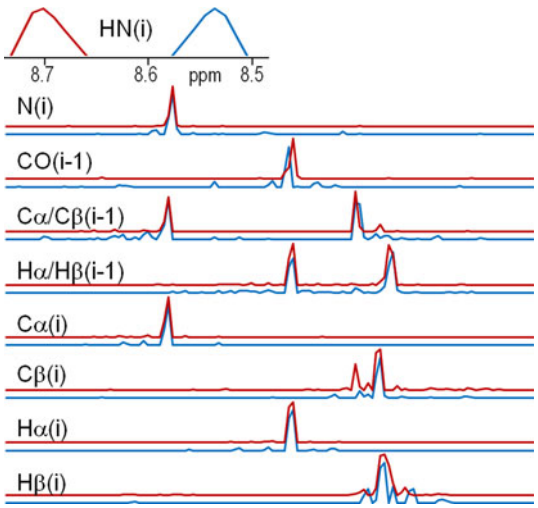

Fig. 3 Example of the chemical shift degeneracy in the GI domain of yeast histone H1 at 298 K: Glu 40 (*red*; i = 40) and Glu 128 (*blue*; i = 128). The *top panel* shows the component shapes for H_N. Because decompositions are performed for intervals along the direct dimension (H_N), only a narrow spectral range around 8.6 ppm is plotted. The remaining eight panels present all corresponding indirect shapes as indicated; here the full spectral width is plotted for each nucleus type. All shapes are plotted after scaling the maximal intensities to 1. A clear separation of chemical shifts for these two residues is only observed for the H_N nuclei, 0.16 ppm, while the signal for all other nuclei pairs strongly overlap

1)-shape from the sequentially following component. The final peak picking missed chemical shifts for one N and for two nuclei of each type Cα, C*β*, Hα and H*β* (only one H*β* per residue was considered). The reference assignment (Ali et al. 2004) provides no chemical shifts for CO. Differences between the current assignment (at 298 K) and the published reference (at 288 K) include shifts for one Hα and two H*β*. These results are summarized in Table 1.

Mmp20

MMP20, with 160 residues the largest protein used in this study, has like all other proteins examined here been previously characterised by conventional NMR methods (Arendt et al. 2007). This earlier study as well as inspection of the ^{15}N-HSQC revealed inherent difficulties: from the 148 expected H_N-N peaks (ignoring 11 prolines), only 130 can readily be identified in a normal ^{15}N-HSQC. Furthermore, when inspecting the published chemical shift list for the presence of chemical shifts for nuclei around each H_N, in particular for Cα, C*β*, Hα, H*β*, CO(*i-1*), Cα(*i-1*), C*β*(*i-1*), Hα(*i-1*), and at least one H*β*(*i-1*), only 121 H_Ns with shifts for all neighbours were found. This list of neighbouring nuclei represents however exactly what is used in the present backbone assignment approach. The absence of their chemical shifts in the published data indicates likely problems for the PRODECOMP-SHABBA approach, making it clear that MMP20 poses a major challenge for the present investigation. Peak picking in the projection ^{15}N-HSQC, using a low signal-to-noise threshold, yielded at best 127 signals that could be used for interval definition. Decomposition with PRODECOMP of these intervals resulted in 102 components with acceptable shapes, i.e. shapes with approximately the expected number of peaks. This implies the presence of many assignment gaps in the sequence, i.e. sequence fragments for which no components are available. SHABBA analysis resulted in 18 component chains, many of which were very short. However, three chains contained long stretches of components that were assigned based on the following arguments: alanines have uniquely low C*β* shifts, serines and threonines have uniquely high C*β* shifts, and glycines lack any C*β* shift; all these features are observables from the shapes of a component. Furthermore, prolines yield no component. Writing 'A' for alanines, 'B' for serines and threonines, and x for all "normal" residues (i.e. not Ala, Ser, Thr, Gly, Pro), the three component chains can be represented as follows: BxxxBxxB, xxxxxAxxxAxxAx and AxxxxBxx. Based on this distinction of residue types, the assignment of the three component chains is unique within the entire 160 amino acid sequence of MMP20. The three component chains can thus be unambiguously assigned to residues 10–17, 23–36 and 52–59. Finally, one may note that problematic regions are distributed unequally over the MMP20 sequence. Thus, an independent assignment

attempt using an artificially reduced list of components, which is restricted to the first third of the protein (residues 11–60), would allow more complete assignments by leaving only gaps for a total of eight non-proline residues.

Discussion

Projection spectroscopy coupled to decomposition analysis is a useful tool for many types of protein characterisations. Here, we discuss backbone assignments; other possible applications are side-chain assignments or structure elucidation (Fredriksson et al. 2012). Backbone assignments, implemented in the SHABBA algorithm (Staykova et al. 2008a, b), rely on correlations among components centred on neighbouring residues and sharing common atoms, in our case all Cα, Cβ, Hα, Hβ of the two residues. Following the sequential arrangement of the components based on these correlations, Cβ shifts, and in a supporting role also Cα shifts, are used to position component chains on the protein sequence. We have presented both a significantly improved SHABBA algorithm, as well as its application to various proteins, including a partially denatured histone domain and a protein that escaped full assignment also with conventional methods. The major improvement of the algorithm is a novel procedure to slide the sequence of Cα and Cβ shifts of component chains along the protein sequence, but significant improvements were also achieved for the peak picking routines and the glycine detection. Except for the change in the lower limit for acceptable correlations in the histone domain application (see "Results"), which was due to the extensive chemical shift degeneracy observed in this protein, all parameter choices were kept unchanged for all four proteins. Only four parameters are required; they regulate noise level detection in the component shapes (how many standard deviations of all but the strongest intensities in a shape define a noise cut-off), glycine detection (when is a shape pure noise due to missing β-nuclei), correlation calculation (rules for zeroing of correlations, see Staykova et al. 2008a, b) and RMSD calculations during the sliding step (e.g. penalties for mismatched glycines). Refraining from optimising these parameters for individual proteins allowed a more objective presentation, but in a normal application results may be improved when adapting some of these to the investigated proteins. Once a broader set of proteins is analysed with this approach, the suggestions of values for these parameters may be further optimised. This includes in particular the increase of the parameter defining a lower allowed limit for acceptable correlations between components from 20 to 35 % for the histone domain. This increase is motivated by the significantly larger average of all correlations, but a reliable rule for this parameter change requires more data sets with extensive shift degeneracy.

SHABBA is a tool for the analysis of projection spectra and does not concern experimental aspects of recording projections. Thus, it is best discussed in the context of other approaches such as PatternPicker (Moseley et al. 2004) or APSY (Hiller et al. 2005), while the GFT (Kim and Szyperski 2003) and Projection-Reconstruction (Freeman and Kupče 2003) approaches follow different routes. The absence of any peak picking in the often noisy projections represents the major difference (the S/N of the individual projections may approach 1 and still allow for successful decompositions of an entire set of projections (Malmodin and Billeter 2006)). On the other hand, direct peak picking in the projections may lead to large numbers of peak combinations to consider. Another feature of SHABBA is the possibility to combine projections from various experiments prior to any analysis (e.g. peak picking); this allows for example to improve the decomposition of NOESY-related projections, and therewith the identification of NOEs, with higher quality projections from backbone-related experiments (Fredriksson et al. 2012). Limitations of the approach include so called mixing of two components that have a high degree of degeneracy in the direct dimension and in addition very different amplitudes. Of less fundamental nature is the limitation to (what is often referred to as) 0°, 45° and 90° projections, which could at least partially be removed (Malmodin and Billeter 2006).

Currently, new experiments are explored, e.g. for sequential connections via CO or for replacing the current 5D backbone experiment with a 4D experiment. Ongoing work focuses also on an improved component correlation algorithm and on using more nuclei for the positioning of component chains on the protein sequence. Reliability measures for both the decomposition as well as the peak picker results will further support full automation. The occurrence of wrongly connected components chains into one large chain, e.g. due to missing components or also the presence of prolines, will be addressed by analysing correlations for each nuclei type individually rather than their average. Finally, SHABBA may be generalised to accept input from other types of experiments, for example by combining NOE-based and scalar-coupled spectral data for backbone assignments.

In conclusion, we have demonstrated that the projection-decomposition approach is an efficient and reliable automated tool for fast backbone and CβH$_n$ assignments. It can readily be applied to smaller proteins, including polypeptides with intrinsically disordered or denatured regions. As the yeast histone H1, GI domain, and MMP20 studies show, the approach is robust in the sense that it provides reliable partial answers also in difficult situations where complete assignments are not possible.

Acknowledgments This work was supported by the Swedish NMR Centre and the EU ExtendNMR project. JF thanks the Lawski Foundation for financial support. Data for MMP20 were recorded at CERM (support by the EU-NMR European Network of Research Infrastructures). We thank Tim Stevens and CERM for loans of NMR samples with the histone domain and MMP20, respectively.

References

Ali T, Coles P, Stevens T, Stott K, Thomas J (2004) Two homologous domains of similar structure but different stability in the yeast linker histone, Hho1p. J Mol Biol 338:139–148

Arendt Y, Banci L, Bertini I, Cantini F, Cozzi R, Del Conte R, Gonnelli L (2007) Catalytic domain of MMP20 (Enamelysin)—the NMR structure of a new matrix metalloproteinase. FEBS Lett 581:4723–4726

Billeter M, Staykova DK (2009) Rapid multidimensional NMR: decomposition methods and their applications. Encyclopedia of magnetic resonance

Fredriksson J, Bermel W, Billeter M (2012) Structural characterisation of a histone domain via projection-decomposition. J Magn Reson 217:48–52

Freeman and Kupče (2003) New methods for fast multidimensional NMR. J Biomol NMR 27:101–113

Grzesiek S, Bax A (1992) An efficient experiment for sequential backbone assignment of medium-sized isotopically enriched proteins. J Magn Reson 99:201–207

Grzesiek S, Bax A (1993) Amino-acid type determination in the sequential assignment procedure of uniformly C-13/N-15-enriched proteins. J Biomol NMR 3:185–204

Hershko A, Ciechanover A (1998) The ubiquitin system. Annu Rev Biochem 67:425–479

Hiller S, Fiorito F, Wüthrich K, Wider G (2005) Automated projection spectroscopy (APSY). Proc Natl Acad Sci USA 102: 10876–10881

Kazimierczuk K, Stanek J, Zawadzka-Kazimierczuk A, Kozminski W (2010) Random sampling in multidimensional NMR spectroscopy. Progr NMR Spectr 57:420–434

Kim S, Szyperski T (2003) GFT NMR, a new approach to rapidly obtain precise high-dimensional NMR spectral information. J Am Chem Soc 125:1385–1393

Kupče E, Freeman R (2008) Hyperdimensional NMR spectroscopy. Progr NMR Spectr 52:22–30

Leckner J (2001) Folding and structure of Azurin—the influence of a metal. PhD thesis, Chalmers university of technology. ISBN 91-7291-023-2

Malmodin D, Billeter M (2005) Multiway decomposition of NMR spectra with coupled evolution periods. J Am Chem Soc 127: 13486–13487

Malmodin D, Billeter M (2006) Robust and versatile interpretation of spectra with couple evolution periods using multi-way decomposition. Magn Reson Chem 44:S185–S195

Moseley HNB, Riaz N, Aramini JM, Szyperski T, Montelione GT (2004) A generalized approach to automated NMR peak list editing: application to reduced dimensionality triple resonance spectra. J Magn Reson 170:263–277

Orekhov V, Jaravine V (2011) Analysis of non-uniformly sampled spectra with multi–dimensional decomposition. Progr NMR Spectr. doi:10.1016/j.pnmrs.2011.02.002

Orekhov V, Ibraghimov IV, Billeter M (2001) MUNIN: a new approach to multi-dimensional NMR spectra interpretation. J Biomol NMR 20:49–60

Parr SR, Barber D, Greenwood C (1976) A purification procedure for the soluble cytochrome oxidase and some other respiratory proteins from Pseudomonas aeruginosa. Biochem J 157:423–430

Staykova DK, Fredriksson J, Bermel W, Billeter M (2008a) Assignment of protein NMR spectra based on projections, multi-way decomposition and a fast correlation approach. J Biomol NMR 42:87–97

Staykova DK, Fredriksson J, Billeter M (2008b) PRODECOMPv3: decompositions of NMR projections for protein backbone and side-chain assignments and structural studies. Bioinformatics 24:2258–2259

Ulrich E, Akutsu H, Doreleijers J, Harano Y, Ioannidis Y, Lin J, Livny M, Mading S, Maziuk D, Miller Z, Nakatani E, Schulte C, Tolmie D, Wenger R, Yao H, Markley J (2007) BioMagResBank. Nucleic Acids Res 36:D402–D408

Uversky VN, Dunker AK (2010) Understanding protein non-folding. BBA 1804:1231–1264