2 A New Method for the Assignment of Amino Acid Types

2.1 Difficulties with the assignment of large proteins

The scope of NMR-structure determination has expanded to larger and larger proteins. Simultaneously, new challenges in assigning even the backbone resonances of those large proteins have arisen. While it has been shown that it is in principle possible to achieve resonance assignments (Salzmann et al., 1999) and even solve the structures of fairly large proteins by NMR (Tugarinov et al., 2005), all these methods are based on a near perfect deuteration of the protein, which make classical approaches to the backbone assignment via CBCA(CO)NH based strategies difficult. In the presence of deuterated samples, experiments probing the carbon frequencies of C^{α} or C^{β} have to be performed as 'out and back' sequences that start from HN-magnetisation and end with HN detection. Alternatively, they could rely on carbon-detection (Hu et al., 2003). Both approaches suffer from low sensitivity, either due to the length of the pulse sequence in case of the 'out and back' experiments or the intrinsic low sensitivity of detecting carbons. Other approaches link the amino acid-backbone based on 4D experiments (Tugarinov et al., 2004) that need long measurement times to achieve the required resolution in all four dimensions.

NMR experiments probing general features of amino acids, like carbon resonances, are likely to fail in the context of large, fully labelled proteins due to spectral overlap or sensitivity problems. To solve some of these problems, amino acid selective techniques, either in selective labelling or in the design of the pulse sequences may be applied. Selective labelling can be implemented in a very straightforward way. Usually, a specifically labelled sample containing a single amino acid labelled with ^{15}N is produced to guide the backbone resonance assignment (Arora et al., 2001). When it is required to obtain information for each of the 20 amino acids, the production of the samples is labourious and expensive. Also, it is not always possible to produce specifically labelled samples for each amino acid due to problems with loss of the specific labelling (scrambling) or the changed growth conditions of the bacterial cultures. Another intrinsic disadvantage of single amino acid labelling schemes is the fact that almost all sequential information is lost. This can in principle be overcome by the use of combinatorial labelling strategies, but will always multiply the number of samples required.

Recently, NMR techniques recording spectra for single amino acid types or groups of amino acids have been introduced applying a variety of selective NMR techniques (Schubert et al., 1999; Schubert et al., 2001a; Schubert et al., 2001b). Designed for the automatic backbone assignment of small proteins, these approaches usually fail for larger proteins. The long and complicated pulse sequences are not sufficiently sensitive to probe larger molecules. Additionally, the selection schemes rely on very uniform NMR properties of all residues in the sequence. This is usually not found in large proteins. Perhaps the largest drawback is that most of these experiments will not work in deuterated samples. They relate properties of the amino acid sidechains to the backbone NH and their pulse-sequences would therefore be very long if performed as "out and back" experiments. Also, one of the most potent selection mechanisms in these methods is MUSIC that relies on protons in the sidechains for its selection mechanism.

Another strategy employs differentially labelled samples (Parker et al., 2004). The authors of this study used five samples labelled to different degrees at the N and CO position of selected amino-acids. The labelling degree was then measured by analyzing ^{15}N-HSQC and 2D-HNCO spectra. Since the information of all ten spectra is required to determine the amino acid type connected to a single HSQC peak, this method is called "combinatorial selective labelling" (CSL). The labelling required for this method is achieved by adding selectively labelled amino acids to the medium. To avoid scrambling, the authors suggest to use an *in vitro* expression system to produce the samples.

In this chapter, a simple method to infer the identity of amino acids will be introduced. It is based on the observation that the labelling pattern originally introduced by growing the expression host on either 1,3-^{13}C-glycerol or 2-^{13}C-glycerol (LeMaster et al., 1996) contains information about the amino acid type. An alternating ^{13}C labelling pattern is observed that is dependent on the metabolic synthesis pathway of the particular amino acid. Originally introduced to dilute the coupling network of ^{13}C in solid samples, proteins produced by this procedure have properties that also make them interesting for investigations of very large systems by solution NMR.

2.2 Assignment Based on Labelling Pattern

2.2.1 Principle of the Method

The enrichment of ^{13}C at the C$^{\alpha}$ and CO positions in the 2-^{13}C-Glycerol labelled sample should in principle allow to distinguish five different groups of amino acids, based on the labelling patterns of their different biogenetic precursors (Figure 2.1). All amino acids with a simple labelling pattern are synthesized from glycolytic precursors. Those with a complex labelling pattern are synthesized from precursors that are components of the citric-acid cycle (either oxaloacetate or α-ketogluterate). The only amino acids with unique labelling pattern are leucine and lysine, reflecting their unique biosynthesis. Thus, five groups can be detected: the "glycogenic group" (G, H, S, C, A, W, F, Y and V), the "aspartic acid group" (D, N, M, T and I), the "glutamic acid group" (E, Q, P and R), lysine and leucine. Leucine is the only single amino acid with a unique simple labelling pattern.

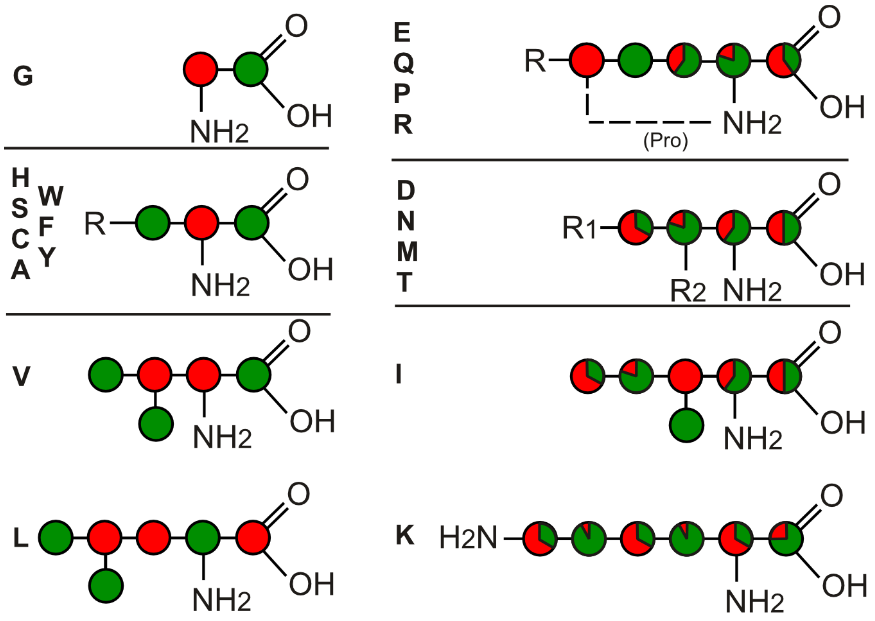

Figure 2.1 Labelling pattern derived from 1,3-^{13}C-glycerol

All carbon atoms depicted in green are labelled when the samples are produced from E.coli grown on 1,3-^{13}C-glyecerol. The carbons depicted in red will remain ^{12}C. The opposite labelling is achieved from 2-^{13}C-glycerol. Figure adopted from (Castellani et al., 2002).

In addition to the differential labelling patterns, the chemical shift information that is recorded in the HNCA can be used for further discrimination. Statistical information obtained from the BMRB about C$^{\alpha}$ chemical shifts in proteins (Figure 2.2) can clearly separate glycines from the "glycogenic group", prolines from the "glutamic acid group" and isoleucines together with threonines from the "aspartic acid group".

Since the assignment strategy is aimed at large proteins it requires robust and sensitive NMR methods that provide the data. Therefore HNCO and HNCA pulse-sequences were used that are amongst the most sensitive 3D protein NMR experiments. These methods are of particular interest since they can be performed in constant time schemes and are also compatible to TROSY techniques. This should in principle allow to apply this approach to very large proteins.

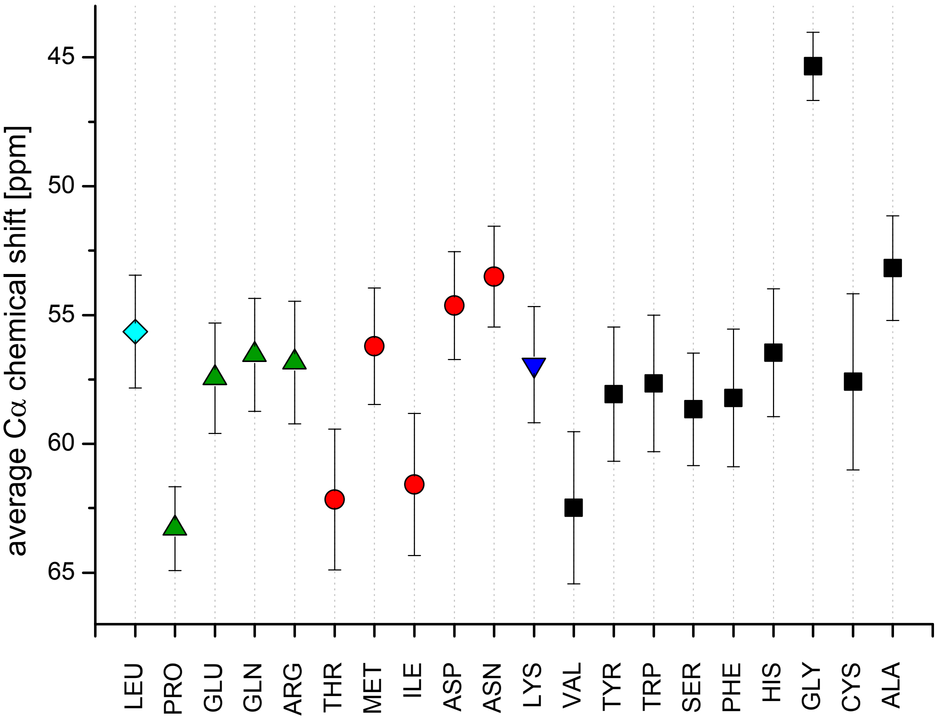

Figure 2.2 Distribution of the average C^{α} chemical shifts

The distribution of the average C^{α} chemical shift is plotted against the amino acids sorted to their respective labelling group. While glycine and proline have unique chemical shifts within their group, threonine and isoleucine have approximately the same chemical shift distribution. Valine might not be always distinguishable from the glycogenic group but amino acids in this group with chemical shifts larger than 63 ppm should almost certainly be valines. Black squares represent the "glycogenic group", blue down-triangles lysines, cyan diamonds leucines. The aspartic and glutamic acid groups are represented by red circles and green up-triangles, respectively. This colouring scheme will be used throughout this section.

Special care has to be taken in acquisition and processing of the spectra of the up to three samples to ensure their comparability, since the experiments are going to be compared in a quantitative way. It is advisable to record spectra that are to be directly compared on the same spectrometer keeping all parameters constant.

For an isolated spin the normalized Volume V should correspond to the fraction of labelled carbons following equation (2.1).

$$F = bV \tag{2.1}$$

'b' accounts for individual properties of that spin within the molecule such as relaxation, motion exchange and others. The fraction of spins leading to a single peak labelled in the 2-glycerol labelled sample (F_2) can thus be expressed as

$$F_2 = \frac{bV_2}{bV_2 + abV_{1,3}} \qquad (2.2)$$

with 'a' accounting for differences in the concentrations of the two samples. Removing all terms contributing equally to both volumes in equation 2.2 leads to equation 2.3.

$$F_2 = \frac{V_2}{V_2 + aV_{1,3}} \qquad (2.3)$$

For non-isolated spins, for example in the HNCA, the observed volume of a single peak is related to both fractions of labelled spins that are coupled to an individual backbone nitrogen via ^{1}J- and ^{2}J-CN couplings. The observable product operator prior to detection is given in equation 2.4

$$I_x \cos\Omega_N t_1 \left(\cos\Omega_{CA} t_2 \left(\sin\pi {}^1J\tau \cos\pi {}^2J\tau\right)^2 + \cos\Omega_{CA'} t_2 \left(\sin\pi {}^2J\tau \cos\pi {}^1J\tau\right)^2\right) \qquad (2.4)$$

thus, we expect the volume of the C$^\alpha$-N-H crosspeaks to be modulated with

$$V_i \sim F_i\left(1 + F_{i-1}\left(\left(\cos\pi {}^2J\tau\right)^2 - 1\right)\right) \qquad (2.5)$$

yielding equation 2.7 and 2.8.

$$\begin{aligned} x &= \left(\left(\cos\pi {}^2J\tau\right)^2 - 1\right) \\ y &= \left(\left(\cos\pi {}^1J\tau\right)^2 - 1\right) \end{aligned} \qquad (2.6)$$

$$V_i = b\left(F_i + xF_iF_{i-1}\right) \tag{2.7}$$

$$V_{i-1} = b\left(F_{i-1} + yF_{i-1}F_i\right) \tag{2.8}$$

Applying equations 2.7 and 2.8 to equation 2.2 together with the assumption that the relation of fractions obtained from the 2-glycerol labelled sample (F_2) to the fractions obtained from the 1,3-glycerol labelled sample is constant ($F_{1,3} = 1 - F_2$) and that a is already included in the volumes, this gives the "observable" fraction F* (eq. 2.9).

$$F_i^* = \frac{F_i + xF_iF_{i-1}}{(F_i + xF_iF_{i-1}) + \left(\left(1-F_i\right) + x\left(1-F_i\right)\left(1-F_{i-1}\right)\right)} \tag{2.9}$$

Analogously

$$F_{i-1}^* = \frac{F_{i-1} + yF_{i-1}F_i}{(F_{i-1} + yF_{i-1}F_i) + \left(\left(1-F_{i-1}\right) + y\left(1-F_{i-1}\right)\left(1-F_i\right)\right)} \tag{2.10}$$

The index indicating that this corresponds the fraction of labelling in the 2-glycerol sample (F_2) is omitted for clarity.

2.2.2 Results on α-Spectrin SH3

The degree of labelling can be expressed as the contribution of the peak volume from the spectra obtained from the 2-glycerol labelled sample to the sum of the volumes for the same peak from both samples according to equation 2.3.

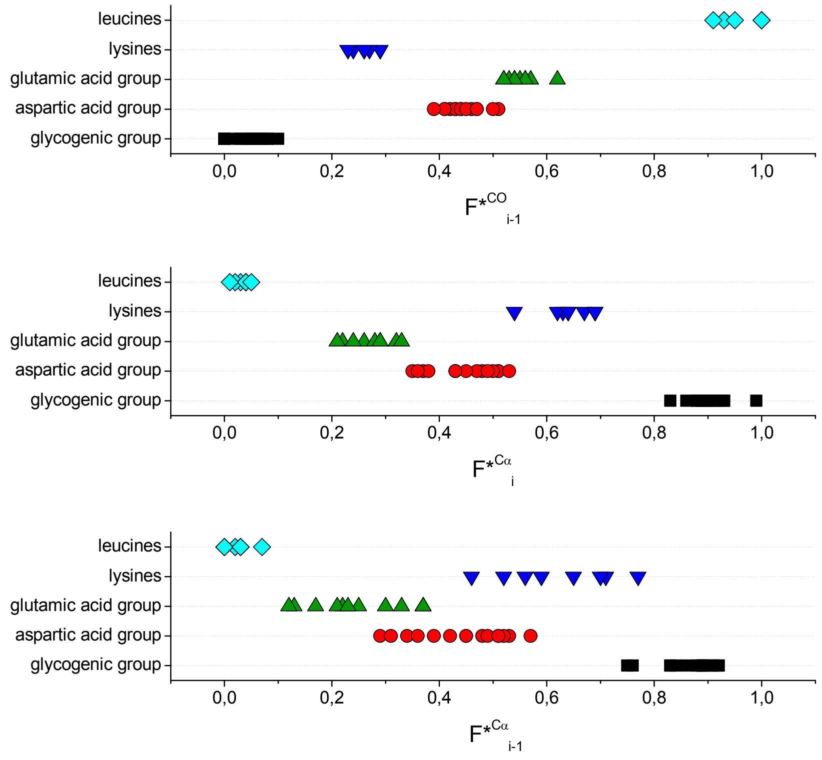

Figure 2.3 Overview over the F* derived from all datasets
Volumes from all peaks in the 2D-HNCO and 3D HNCA spectra were treated according to equation 2.3 and plotted against the group identity of the concerning amino acid.

This should remove most absolute differences in peak volumes stemming from differences in the relaxation properties within the same protein that may arise from internal motion. The calibration factor 'a' can be obtained by comparing spectra from the two different samples that should give comparable results (equation 2.12). In this study the ^{15}N-HSQC spectra (recorded with carbon decoupling) were used for this purpose.

$$a = V_{(\mathrm{HSQC}\ 1,3-\mathrm{glyc})} / V_{(\mathrm{HSQC}\ 2-\mathrm{glyc})} \tag{2.12}$$

The volumes in eqauation 2.12 were taken from HSQC spectra from both samples. Averaged over all peaks, this gave a correction factor of 1.02. The assignments for this analysis was taken from earlier works (Pauli et al., 2001). For the work presented here, the assignment of the peak identities in the 3D HNCA spectra, where two C$^{\alpha}$-peaks occur per ^{1}H-^{15}N couple, was also taken from the existing assignment.

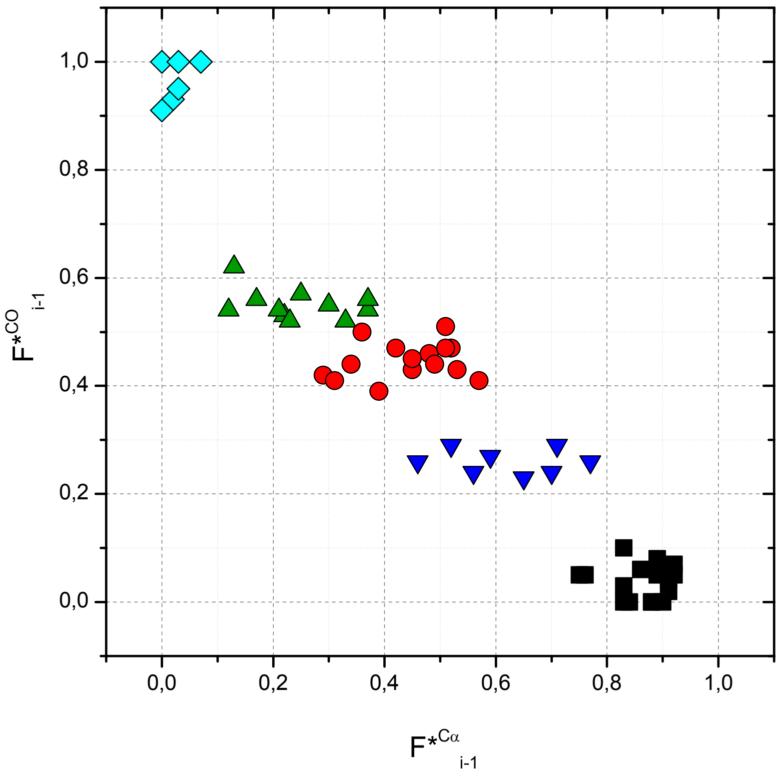

Figure 2.4 Plot of F*^{CO} against F*$^{C\alpha}$

Data obtained from identical amino acids via HNCO and HNCA spectra is highly correlated. This is due to the almost perfectly alternating ^{13}C-labeling obtained from the glycerol (Compare Figure 2.3).

In a real case, were the assignment is not known, a HN(CO)CA from a fully ^{13}C,^{15}N-labelled sample would help to identify, which of the two peaks belongs to the i-1 amino acid and thus also which one belongs to the detected ^{1}H-^{15}N couple. To obtain a measure of the performance of the method, all peaks were assigned using the pre-existing assignments and the discrimination of the five expected groups of amino acids was evaluated.

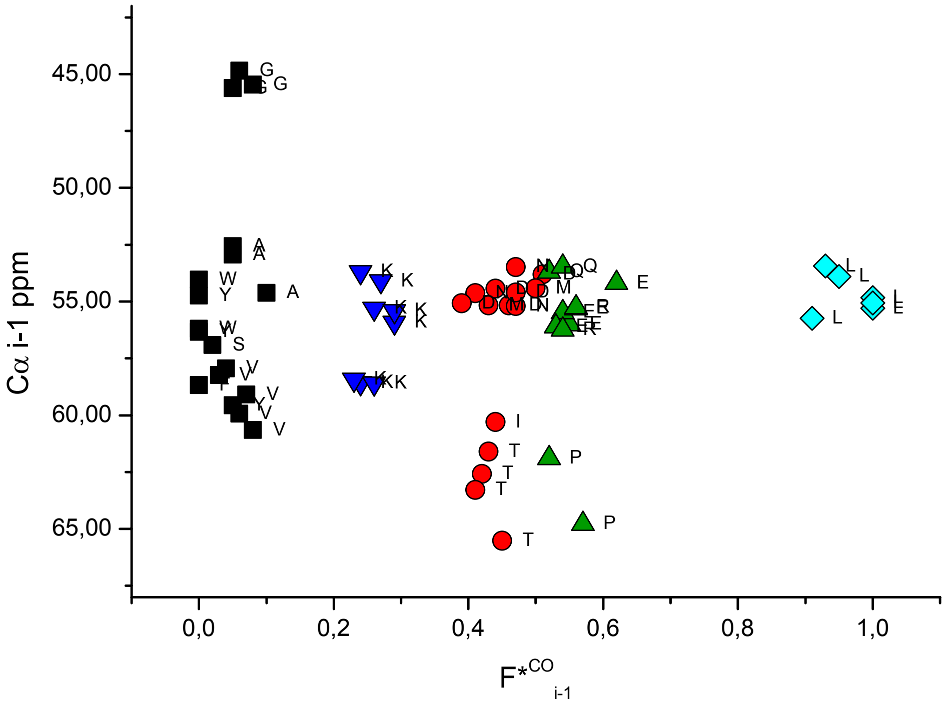

Figure 2.5 Plot of the C^{α} chemical shift from the i-1 amino acid against F^{*CO}
The chemical shift from the same amino acid is plotted against F^{CO}. This demonstrated clearly that the C^{α} chemical shifts of I/T, P and G are separated from the rest of their respective groups. To some extend valine can also be separated.

Obviously, the data derived from the HNCO spectra contains sufficient information to distinguish the five observable groups (Figure 2.3). It should therefore be possible to assign each amino acid X from any (X)Y-pair in the sequence to the correct group. The scattering in the HNCA spectrum is much larger and seems to depend on the peak intensities as the $F^{*C\alpha_i}$ values stemming from the more intense peaks scatter less than the $F^{*C\alpha_{i-1}}$ values. In both cases only the leucines and the glycogenic group are separated from the rest of the others fairly well.

As expected, the fractional labelling data corresponding to one ^1H-^{15}N-pair is highly correlated (Figure 2.4). This is due to the fact that the labelling pattern obtained from the two labelled forms of glycerol is almost perfectly alternating in the protein backbone. Once again the large spread of the C^{α} data is visible, especially when compared to the CO data. Taking the well resolved CO data and relating it to the C^α chemical shift (Figure 2.5) demonstrates the power of the chemical shift information to separate the otherwise quite homogenous groups. Clearly, G, I, T and P can be separated from the rest of the amino acids in their respective groups. Although the valines show chemical shifts around 60 ppm, other amino acids from that group show similar shifts making it difficult to separate all of them. So special care has to be taken, identify the valines using the chemical shift information.

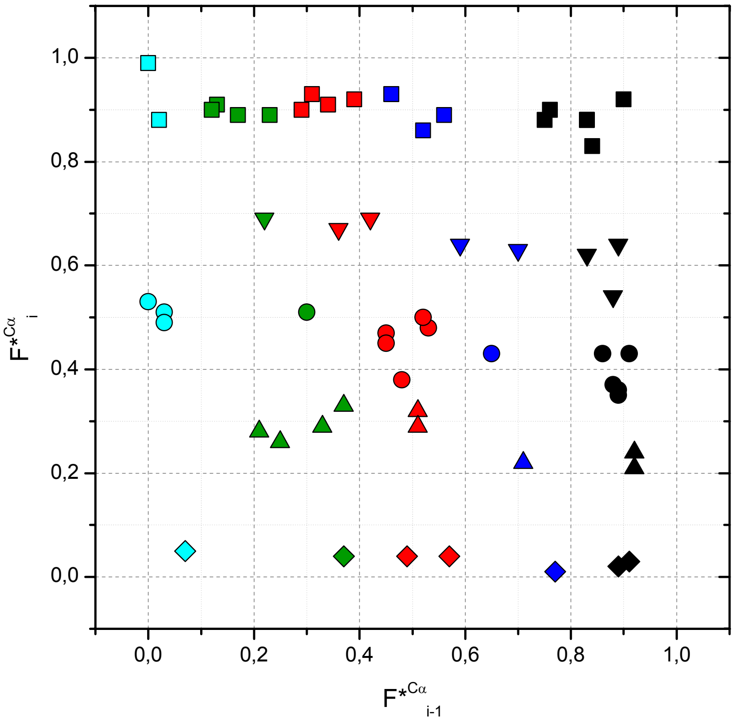

Figure 2.6 Plot of $F^{*C\alpha}$i-1 against $F^{*C\alpha}$i

Plotted are the $F^{*C\alpha}$-values that can be obtained from a single HSQC peak. The shapes of the symbols give the group of the i-1 amino acid, the colours give the identity of the i amino acid. Obviously, there is a dependency between those values that does not have its origin in the type of the amino acids.

The data shown in Figure 2.3 indicates that an isolated analysis of a single C^α F* value will be difficult because of the observed scattering. On the other hand, only the

HNCA contains information about the Y amino acids from the XY pairs. Thus, a method, which allows analysing the data despite the scattering, would be a great step towards an assignment of amino acid pairs. Plotting the data of both C^{α} resonances visible at a single ^{1}H-^{15}N resonance pair in the HNCA (Figure 2.6) reveals a strong correlation expected from equation 2.10. Clearly, the F* values obtained, for example, from lysines (blue for C^{α}_{i-1}, down-triangles for C^{α}_{i} in (Figure 2.6) strongly depend on the F* values obtained from the other peak correlated to the same ^{1}H-^{15}N resonance pair. Thus, instead of a rectangular grid, a deformed pattern is expected that takes into account the intensity dependencies of C^{α} peaks.

Based on this observation, it should be possible to derive grid-points for the expected areas from the internal dependencies in the HNCA spectrum. Equation 2.9 allows calculating the predicted areas in which the 25 possible amino acid pairs should appear in the $F^{*C\alpha_{i}}$ vs. $F^{*C\alpha_{i-1}}$ plot. The corners of those areas calculated with coupling constants of 10 Hz and 6 Hz and $F^{C\alpha}$ of 0.95, 0.6 0.5 0.2 and 0.05 are indicated in Figure 2.7 as blue crosses. For the calculations, it is assumed that the borders lie exactly in the middle between the theoretically assumed fractional labelling values.

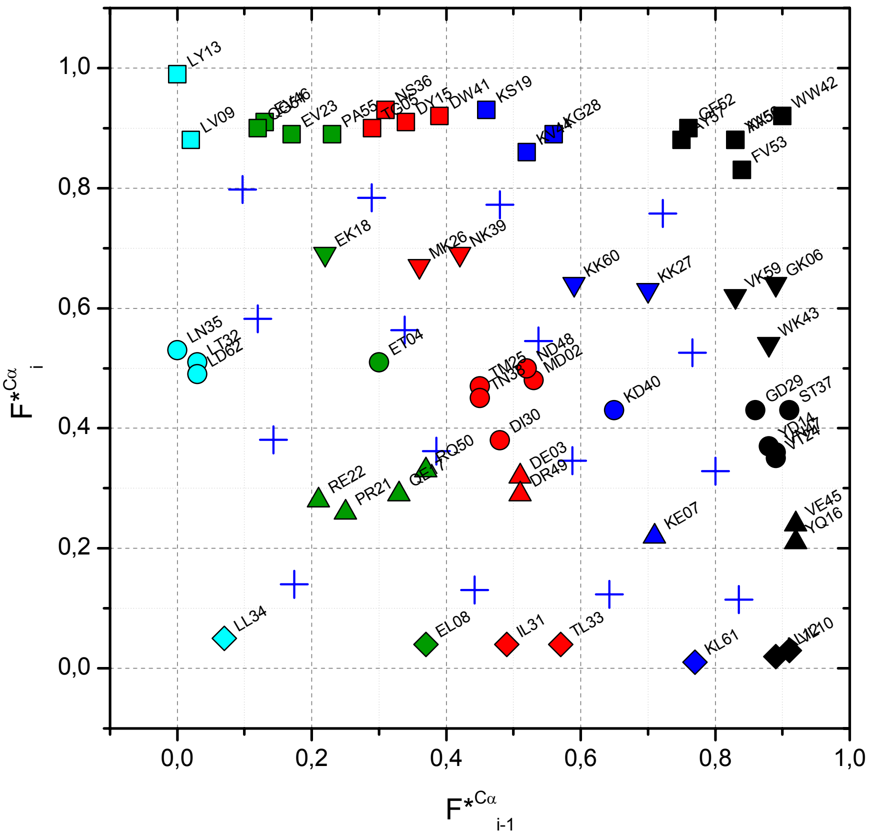

Figure 2.7 Simulation of dependencies for $F^{C\alpha}$

Plotted are the $F^{*C\alpha}$-values that can be obtained from a single HSQC peak. Shapes give the group of the i-1 amino acid, the colour give the identity of the i amino acid. The number in the labels corresponds to the i position. Blue crosses indicate the corners of those areas were the 25 combinations of groups are expected from eq. 9 calculated with coupling constants of 10 Hz and 6 Hz and $F^{C\alpha}$ of 0.95, 0.6 0.5 0.2 and 0.05.

All pairs of $F^{*C\alpha}$ values from the SH3 domain fall into the expected areas. This shows that the quality of the data was high enough to assign every amino acid Y to the correct (X)Y pair based only on the fractional labelling data. This implies that all leucines and lysines were identified. Furthermore, the distribution of pairs in SH3 allowed to unambiguously assign 7 pairs to their correct sequence position and left only two possible sequence positions for 14 other pairs. Taking the C^{α} chemical shift of both the X and Y amino acids into account (Figure 2.8) another 12 amino acids can be assigned to their correct position. Valines that might or might not exhibit shifts larger than 60 ppm were considered as not identifiable and disturbing the assignment of other > 60 ppm amino acids. A very extensive analysis of the amino acid pairs that are expected from a combination of groups might even allow to improve the assignment further.

For example, there are five pairs of "glutamic acid group" residues and "aspartic acid group" residues: GD, ST, VT, VN and YD. Of these five pairs, GD is easily identifiable from its C^{α}_{i-1} chemical shift. Both the VT and VN pair should exhibit $C\alpha_{i-1}$ chemical shifts around 60 ppm but from this data alone they are not distinguishable. The C^{α}_{i} chemical shift highlights the ST and the VT pair. Again they are not distinguishable from this data alone. Taken together, the VT pair shows two identifiable chemical shifts and can thus be separated from the ST as well as the VN pair. This way all these three pairs can be assigned. The YD pair is the only pair that remains with unassigned from chemical shift data. This way all five pairs of this combination of groups can be unambiguously assigned.

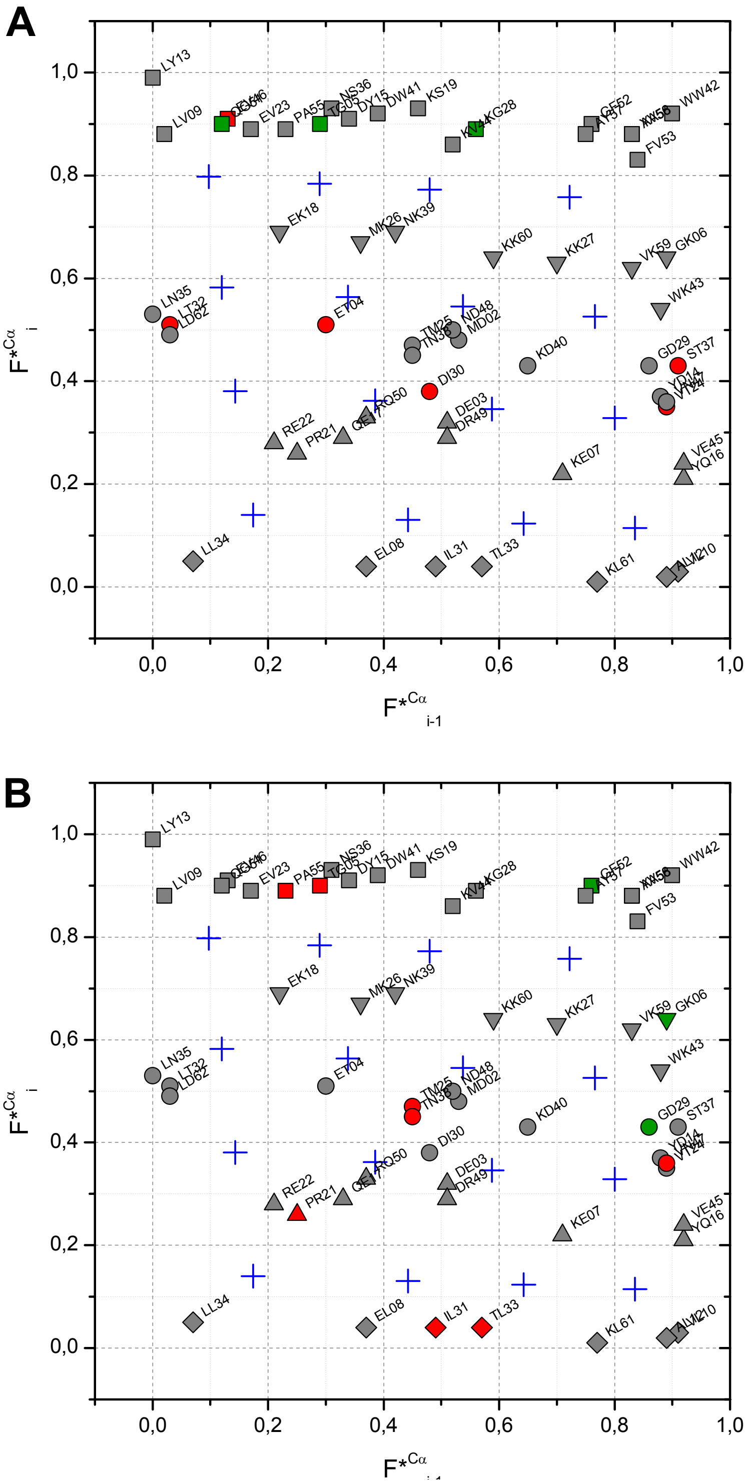

Figure 2.8 Indexing of “unusual” C$^{\alpha}$ chemical shifts

Plotted is the same data as in Figure 2.7. Additionally, the own C$^{\alpha}$ chemical shift **(A)** or that of the preceeding residue **(B)** is indicated by the colour of the symbols. Chemical shifts bigger than 60 ppm are indicated in red, those smaller than 50 ppm indicated in green. All other chemical shifts are coloured grey.

2.2.3 Discussion and Outlook

Using the outlined procedure, about a third (22) of the 62 amino acids in the α-spectrin SH3 domain can be assigned unambiguously. In the context of a larger protein this number is likely to decrease due to the smaller chance that a combination of groups and chemical shifts is unique.

	Leucine	Glut. Group	Asp. Group	Lysine	Glyc. Group
Glyc. Group	4,0	6,3	9,3	2,4	16,3
Lysine	0,6	0,9	1,4	0,4	2,4
Asp. Group	2,3	3,4	5,4	1,3	9,4
Glut. Group	2,0	3,5	4,7	1,2	8,3
Leucine	0,9	1,6	2,2	0,6	3,8

Table 2.1 Probability of group pairs in %
The table gives the probability of all possible pairs of the five groups that can be distinguished by their labelling pattern. The data was taken from (Cserzo et al., 1989).

Still, the ability to identify all leucines and lysines makes this method very interesting concerning a use as a tool for the initial assignment of protein backbones. Together both amino acids are found in ~16% of all amino acid positions of currently known protein sequences, so statistically every sixth amino acid should be either a lysine or a leucine. Leucine is the most abundant amino acid found in 10% of all positions. 16 of the 25 possible combinations of amino acid groups involve either a leucine or a lysine or both. These groups represent 28.1% of all occuring pairs (Table 2.1). In addition, pairs in which only one of these amino acids occurs, can be unique even in 30 – 40kD proteins, since they have a very low probability. The 62.3% of amino acids that occur in combination with the "glycogenic group" can be differentiated by two different chemical shifts. Since both glycine and in some cases valine can be separated from the group, at least the 13.2% of pairs that contain glycine can be identified and lower the number of pairs for which the assignment remains very ambiguous. Interestingly, the non-statistical distribution of amino acids in pairs leads to differences in the occurrence between AB and BA pairs. For example, the combination of "glutamic acid group"-"glycogenic group" occurs 1.2 times more often than the reverse combination.

Compared to other methods that yield amino acid specific information, the use of the method described here requires a reduced number of samples. The application of amino acid specific labelling implies that a different sample is needed for every additional information. Glycerol based labelling allows to distinguish 25 types of amino acids from three samples. It also requires a smaller number of samples that the "CSL" method, where five samples are needed, while offering the potential of gaining more information. The NMR experiments required are compatible even with very large molecules.

One of the most important aspects rendering this approach unique is the fact that sequential information is obtained. Although no linkage to another HSQC-peak is achieved, the method yields information about the preceding residue. Lysines and leucines can be identified and all other amino acids can be assigned to their respective groups. More importantly, this sequential information potentially allows to reduce the number of possible assignment options for a single pair. In the 62 amino acid protein domain α-spectrin SH3, it was possible to assign 22 amino acids unambiguously, seven additional ones were left with only two possibilities. In larger proteins, the proportion of unambiguous assignments will be lower but still a large fraction of amino acid pairs will be assigned to only two or three positions.

Finally, many steps of the method could be automated. Primary sequence analysis can reveal, which pairs of groups are to be expected and even which pairs might be unambiguously identified. All this can easily be automated and integrated into an analysis/display tool performing the necessary calculations and providing results compatible with common assignment software

2.3 References for Chapter 2

Arora,A., Abildgaard,F., Bushweller,J.H., and Tamm,L.K. (2001). Structure of outer membrane protein A transmembrane domain by NMR spectroscopy. Nat. Struct. Biol. 8, 334-338.

Castellani,F., van,R.B., Diehl,A., Schubert,M., Rehbein,K., and Oschkinat,H. (2002). Structure of a protein determined by solid-state magic-angle-spinning NMR spectroscopy. Nature 420, 98-102.

Cserzo,M., and Simon,I. (1989). Regularities in the primary structure of proteins. Int. J. Pept. Protein Res. 34, 184-195.

Hu,K., Eletsky,A., and Pervushin,K. (2003). Backbone resonance assignment in large protonated proteins using a combination of new 3D TROSY-HN(CA)HA, 4D TROSY-HACANH and 13C-detected HACACO experiments. J. Biomol. NMR 26, 69-77.

LeMaster,D.M., and Kushlan,D.M. (1996). Dynamical mapping of E-coli thioredoxin via C-13 NMR relaxation analysis. J. Am. Chem. Soc. 118, 9255-9264.

Parker,M.J., ulton-Jones,M., Hounslow,A.M., and Craven,C.J. (2004). A combinatorial selective labeling method for the assignment of backbone amide NMR resonances. J. Am. Chem. Soc. 126, 5020-5021.

Pauli,J., Baldus,M., van,R.B., de,G.H., and Oschkinat,H. (2001). Backbone and side-chain 13C and 15N signal assignments of the alpha-spectrin SH3 domain by magic angle spinning solid-state NMR at 17.6 Tesla. Chembiochem. 2, 272-281.

Salzmann,M., Pervushin,K., Wider,G., Senn,H., and Wuthrich,K. (1999). [13C]-constant-time [15N,1H]-TROSY-HNCA for sequential assignments of large proteins. J. Biomol. NMR 14, 85-88.

Schubert,M., Oschkinat,H., and Schmieder,P. (2001a). MUSIC and Aromatic Residues: Amino Acid Type-Selective (1)H-(15)N Correlations, III. J. Magn Reson. 153, 186-192.

Schubert,M., Oschkinat,H., and Schmieder,P. (2001b). MUSIC, selective pulses, and tuned delays: amino acid type-selective (1)H-(15)N correlations, II. J. Magn Reson. 148, 61-72.

Schubert,M., Smalla,M., Schmieder,P., and Oschkinat,H. (1999). MUSIC in triple-resonance experiments: amino acid type-selective (1)H- (15)N correlations. J. Magn Reson. 141, 34-43.

Tugarinov,V., Choy,W.Y., Orekhov,V.Y., and Kay,L.E. (2005). Solution NMR-derived global fold of a monomeric 82-kDa enzyme. Proc. Natl. Acad. Sci. U. S. A 102, 622-627.

Tugarinov,V., Hwang,P.M., and Kay,L.E. (2004). Nuclear magnetic resonance spectroscopy of high-molecular-weight proteins. Annu. Rev. Biochem. 73, 107-146.