Theor Chem Acc (2001) 106:91–97
DOI 10.1007/s002140000217

Theoretical
Chemistry Accounts

Regular article

FIRE: predicting the spatial proximity of protein residues from 3D NOESY–HSQC

T. E. Malliavin[1], **P. Barthe**[2], **M. A. Delsuc**[3]

[1] Laboratoire de Biochimie Théorique, UPR 9080, IBPC, 11, rue P. et M. Curie, 75005 Paris, France
[2] Departament de Quimica Organica, Universitat de Barcelona, c/Marti i Franques, 1-11, 08028 Barcelona, Spain
[3] Centre de Biochimie Structurale, INSERM U 414, CNRS UMR 5048, Université de Montpellier I, Faculté de Pharmacie, 15, av. Ch. Flahault, 34060 Montpellier Cedex 2, France

Received: 5 July 2000/Accepted: 8 September 2000/Published online: 19 January 2001

Abstract. We address the problem of the prediction of residue spatial proximity in a protein, through the automatic processing of a 3D ^{15}N NOESY–HSQC. The spatial distance between residues is estimated from a spectral match value calculated using a comparison of the resonances involving the amide hydrogens. The method is shown to provide a good estimation of a large number of residue spatial proximities, in the case of two experimental 3D spectra, recorded on proteins of α and β secondary structures. It is tested on simulated data sets against the protein size, secondary structure and the quality of the signal. More than 70% of the sequential assignment is correctly predicted, and the prediction is better for the α than for the β secondary structure. The medium- and long-range correlations seem equally well predicted for all the secondary structures. The efficiency of the method is compared to a previously proposed spectral correlation approach.

Key words: Nuclear magnetic resonance – Structure prediction – Computer-aided assignment – Data processing

1 Introduction

Methods permitting rapid insight into geometric properties of proteins given the simple observation of their NMR spectrum are certainly very interesting and valuable tools. They are of particular interest in the frame of structural genomics studies as they could permit a rapid determination of the global fold and could eventually permit rapid protein structure determination in an unattended manner.

Methods have been proposed to get hints on the secondary structure from the observation of mere chemical shifts [1], but the determination of the global fold and of the tertiary structure has not been addressed yet. It is usually considered that this cannot be done with at least a partial assignment of the spectrum in order to relate to the peptide backbone the connectivities which are observed in the nOe spectra [2].

In the case of a double-labeled sample, methods for automatic determination of spin system and sequential assignments [3, 4, 5] have been proposed. However, in other cases, assignment is still very dependent on manual spectral analysis, thus presenting a major bottleneck for the use of NMR in structural genomics.

For a ^{15}N single-labeled protein, a possible way towards gaining information on protein geometry is by direct analysis of the 3D ^{15}N nuclear Overhauser enhancement spectroscopy (NOESY)–heteronuclear single quantum coherence (HSQC) experiment, which displays spatial proximity information while usually presenting little spectral superposition for the amide protons. The spectral correlation method [6] was a first attempt in that direction; here we explore this approach more extensively.

The protocol proposed here to process 3D ^{15}N NOESY–HSQC is made up of the following steps. For each protein residue observed on the 2D HSQC, the subspectrum containing the nOe correlation peaks involving the residue amide hydrogen is extracted from the 3D experiment. All pairs of subspectra are then compared two by two on the basis of the detected peaks; values are built from this comparison (called match values in the following) and stored in a match matrix. This match matrix shows strong resemblance with the contact matrix of the protein as the subspectrum pairs corresponding to residues close in space usually exhibit several peaks located at the same ^{1}H chemical shift and consequently produce large match values. This protocol

Correspondence to: T. E. Malliavin
e-mail: Therese.Malliavin@ibpc.fr

Contribution to the Symposium Proceedings of Computational Biophysics 2000

has been called FIRE standing for Fold Insight by nuclear ResonancE. It is evaluated here by running it on experimental as well as on simulated 3D ^{15}N NOESY–HSQC spectra obtained from a range of proteins presenting different sizes and topologies.

2 Systems and methods

Experimental data sets were recorded on P8^{MTCP11} and P13^{MTCP11}, two proteins of 68 and 107 amino acids respectively, and encoded by the MTCP1 oncogene [7]. The p13 polypeptide used for the NMR study contained 117 residues because of the clone construct [8]. The p8 main structural motif [9] consists of two antiparallel helices spanning residues 8–20 (αI) and 29–40 (αII), strapped in an α- hairpin motif by the two disulfide bridges 7–38 and 17–28. The third helix (αIII), spanning residues 48–63, is connected to the double-helix motif by a loop from residue 41–46, and a third disulfide bridge 39–50 links the top of helix αIII to the tip of helix αII. The 3D structure of p13 shows two roughly symmetrical motifs. Each motif is made of four-stranded antiparallel β sheets, containing two short strands and two long strands. The two β sheet motifs are connected by a large loop, which appears less defined, probably owing to increased flexibility. These two motifs wrap on an orthogonal β-barrel, forming the core of the protein. They delimit a cavity which is completely filled by the sidechains of hydrophobic residues on the β sheets. In each motif, the two longer strands form a larger two-stranded β sheet.

The 2D HSQC and 3D NOESY–HSQC spectra were recorded using an AMX600 Bruker spectrometer, using previously described acquisition parameters [10]. Processing and handling of assignment data were realized with the Gifa assignment module [11, 12].

Simulated data sets were also calculated for six proteins (Table 1), using Protein Data Bank (PDB) [13] and Bio Mag Res Bank assignment files [14]. Intensities were calculated using the program SPIRIT [15], which takes into account the efficiency of INEPT magnetization transfer and incomplete recovery of z magnetization between scans. A relaxation delay of 1.25 s, a first INEPT duration of 3.66 ms, a reverse INEPT duration of 1.83 ms, a model of isotropic rigid motion with a global correlation time of 4 ns, and a mixing time of 200 ms were used for all simulations. The spectra were calculated from simulated intensities larger than 0.002, as the sum of Lorentzian functions of frequencies corresponding to proton chemical shifts, and with linewidths of 25 or 35 Hz. It was supposed here that there are no peak superpositions in 2D HSQC (see Sect. 3.2), and only the proton chemical shifts were used to simulate the 3D spectrum. In order to check the reliability of the simulation, the data set simulated for protein p8 was compared to the experimental spectrum, and both gave rise to the same order of match values (data not shown).

The amount of noise, σ, in the experimental and simulated data sets was estimated as the standard deviation of the data points in an empty spectral zone; this noise was computed in each spectral column, in order to take into account the noise variation from one column to another.

The commands involved in the FIRE protocol were implemented in Gifa, using the Gifa macro language [11] and Fortran 77.

3 Algorithms

3.1 Estimation of the fold information contained in a 3D NOESY–HSQC experiment

For this study, it is a central question to know whether a 3D NOESY–HSQC experiment contains a sufficient amount of information to define or hint at the protein fold. Many protons of the protein are simply absent from this spectrum and have little or no impact on the final result. The distance information extracted from the measured signal is filtered through the amide hydrogens and thus reflects only very indirectly the 3D Euclidean distance between atoms.

To estimate how the distance information is coded in the 3D experiment, between all the non-proline residues, i and j, we used the parameter Δ_{ij}, which is equal to the minimum distance between an amide hydrogen and the hydrogens of the other residue:

$$\Delta_{ij} = \mathrm{Min}_{\mathrm{H}_i,\mathrm{H}_j}[d(\mathrm{HN}_i, \mathrm{H}_j), d(\mathrm{HN}_j, \mathrm{H}_i)] \ , \qquad (1)$$

where H_i, H_j are the hydrogens and HN_i, HN_j the amide hydrogens of residues i and j; d is the 3D Euclidean distance; $\mathrm{Min}_{\mathrm{H}_i,\mathrm{H}_j}$ is the minimum value on the overall set of hydrogens from residues i and j. The exchangeable hydrogens, aromatic hydrogens and Hϵ hydrogens from lysines are excluded from the calculation. Δ verifies the distance properties [16], and is related to the maximum distance information which can be extracted from an isolated 3D NOESY–HSQC.

The Δ distance and the 3D Euclidean distance between amide hydrogens are compared for proteins p8 and p13 (Fig. 1) by displaying their inverse (called proximities in the following). The 3D Euclidean proximity barely shows the long-range contacts of the structures and many sequential proximities are missing. On the other hand, the proximity calculated from Δ defines the sequential proximities and the protein fold; there are no more missing sequential proximities in p8 and p13, apart from at the proline positions. Furthermore, the α-hairpin motif of p8 and the β barrel of p13 appear to be well defined by the Δ proximity.

Table 1. Features of the proteins studied

Identification[a]	Name	Number of residues	Reference	Protein Data Bank entry	Bio Mag Res Bank entry	Kind of data set	Secondary structures
acyt	Apocytochrome b562	106	24	1iet	1672	Simulated	α
ayj	Antifungal protein	51	18	1ayj	–	Simulated	α/β
kum	Glucoamylase	108	25	1kum	4011	Simulated	α/β
ner	Ner DNA-binding protein	72	26	1ner	287	Simulated	α
snob	Staphylococcal nuclease	103	27	2sob	4010	Simulated	α/β
srl	Src SH3 domain	64	28	1srl	3433	Simulated	α/β
p8	P8^{MTCP11} oncogenic protein	68	9	2hp8	–	Experimental	α
p13	P13^{MTCP11} oncogenic protein	117	8	1qtu	–	Experimental	β

[a] The identification defined for each protein is used to refer to it in the text

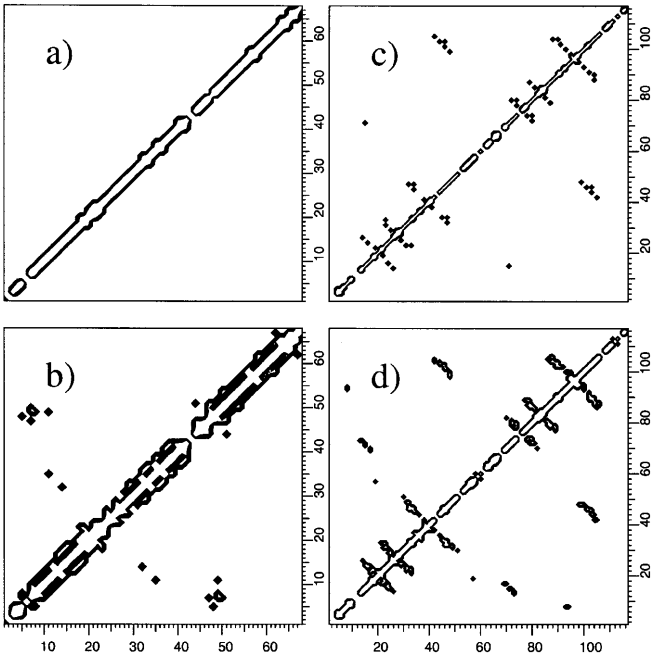

Fig. 1. Comparison of the proximity matrices obtained for p8 (**a**, **b**) and p13 (**c**, **d**) using the Euclidean distance between amide hydrogens (**a**, **c**) and the distance Δ (see text) between residues (**b**, **d**). The proximity is equal to the inverse of distance for nonzero distances and is 1 for zero distances. Only proximities corresponding to distances smaller than 4 Å are shown. The *x*- and *y*-axes display the residue number

Nevertheless, a closer examination of the p8 Δ proximities reveals that residue 1 is isolated from the other residues and that the submatrices containing residues 3–42 and 44–68 are connected only by long-range proximities. This is due to prolines 2 and 43, located respectively in N-terminal and loop regions of the protein, where little $(i, i+2)$ or $(i, i+3)$ contact can be found.

In the following, all the results obtained from the processing of 3D NOESY–HSQC spectra are compared to spatial proximity information based on the distance Δ; only the proximities corresponding to distances smaller than 4 Å are shown in the figures.

3.2 Prediction of the residue spatial proximity

The FIRE protocol consists of extracting the fold information from the 3D NOESY–HSQC experiment by estimating the similarity of the NOE columns extracted from the 3D experiment. The 2D HSQC spectrum is peak-picked, and a regular strip-plot is constructed from the 3D NOESY–HSQC experiment by extracting a subspectrum along the ^{1}H NOE axis from the 3D experiment. In each subspectrum, a column located at the maximum of the observed signal is extracted manually. The columns are then concatenated to produce a matrix **S**. A baseline correction is applied on each column of **S** to reduce the offset difference between columns. For the protein residue with index k, the columns C_k of the matrix **S** contain the peaks observed on the 3D NOESY–HSQC, which involve the amide hydrogen of this residue. The column size is equal to the size of the spectral axis in the 3D experiment bearing the ^{1}H related by NOE (usually F1 or F3). For the sake of convenience, the columns are ordered here according to the protein sequence. For simulated data sets, **S** is calculated directly from the simulated intensities and the proton chemical shifts, as described in Sect. 2. The assumption is made here that for each non-proline residue, it is possible to extract from the 3D NOESY–HSQC the set of resonance peaks involving its amide hydrogen. In the case of p8 and p13, we observed that even peaks slightly superimposed in the 2D HSQC gave rise to separable signals in the 3D NOESY–HSQC.

A filtering window is then applied to each column of **S** in order to cancel out the spectral intensities located at the water frequency. As a matter of fact, the water signal is observed on almost all the columns and, if not removed, induces a bias into the result of FIRE. Then a σ value is measured on each column, and all the column intensities smaller than 5σ are canceled. Finally, the columns containing fewer than two peaks and the columns for which the mean intensity is smaller than 0.0005 times the **S** mean intensity are discarded. The remaining columns are peak-picked; the central location of each peak is replaced by 1, and the other locations are set to zero. The C_k^{tf} columns obtained are thus formed from values 0 and 1. For each column pair $(C_i^{\mathrm{tf}}, C_j^{\mathrm{tf}})$, we calculate a value M_{ij} which quantifies the spectral similarity (match) between residues i and j:

$$M_{ij} = \frac{2 < C_i^{\mathrm{tf}} | C_j^{\mathrm{tf}} >}{\|C_i^{\mathrm{tf}}\|^2 + \|C_j^{\mathrm{tf}}\|^2} \tag{2}$$

$< | >$ is the scalar product between the two column vectors and $\| \ \|$ is the column vector norm. If C_i^{tf} and C_j^{tf} are both null columns, M_{ij} is set to 0.

M_{ij} takes a value between 0 and 1. Two identical columns will produce an M_{ij} value of 1, whereas columns having no peaks facing each other will give an M_{ij} value of O. M_{ij} does not depend on the order of the columns in **S**. A symmetric square matrix, **M**, the match matrix, is built from the M_{ij} values. The M matrix is then filtered by applying a threshold, γ, such that for each residue i, the number of residues j producing a match value M_{ij} larger than γ is in the range 4–4.5. Indeed, in the p8 and p13 Δ matrices (Fig. 1), for any residue i, the mean number of elements Δ_{ij} smaller than 4 Å is 4.7 (p8) and 4.2 (p13).

4 Results and discussion

4.1 Processing of experimental spectra

Match matrices, **M**, were calculated as described previously from the experimental data sets recorded for the p8 and p13 proteins (Fig. 2). The results are analyzed by comparing the observed matches to the Δ proximity values described in the previous section. The comparison is performed independently for the sequential ($|i-j| < 1$), medium-range ($1 < |i-j| \leq 3$), and long-range ($|i-j| > 3$) observed correlations, and values are reported as percentages of correct correlations.

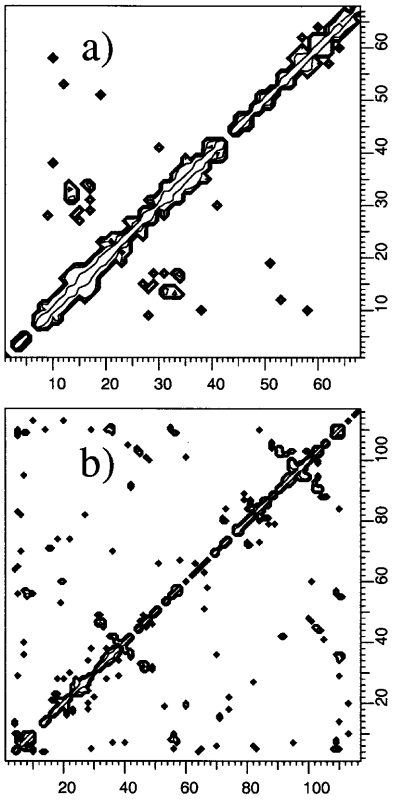

Fig. 2. Match matrices calculated on **a** p8 and **b** p13. The *x*- and *y*-axes show the residue numbers

Matrix **M** of p8 (Fig. 2a) is very similar to the corresponding proximity matrix (Fig. 1b); 88% of sequential, 40% of medium-range, and 9% of long-range correlations are properly predicted. The predicted long-range correlations are false except for the pair of residues (14, 32). On the other hand, long-range proximities observed in the Δ matrix (Fig. 1b) between residues 5–11 and 47–50 are not predicted at all. Two submatrices including residues 1–5 and residues 44–47, show no correlations with other protein residues; they are due to the break in the correlation graph produced by prolines 6 and 43.

In the case of p13, only 102 non-zero columns are kept in the match matrix **M** (Fig. 2b); apart from eight proline columns, seven weak intensity columns were set to zero. The global features of the contact map (66% of sequential, 43% of medium-range, and 31% of long-range correlations) are correctly predicted. The predicted long-range correlations permit the determination of five β strands over the eight of the protein from the correlations involving residues 30–51 (C and D), 77–89 (G and F), and 86–105 (G and H); however, numerous false correlations are predicted, and the sequential correlations of residues 29–31 and 60–67 are not predicted. The 60–67 region is located in an external loop, and the 29–31 region is at the beginning of the β strand C.

A larger percentage of correct sequential prediction is found for the p8 protein than for p13. Three different reasons can be invoked to explain this fact, but it is difficult to say which one is preponderant. First, the sequential connections observed in (α) secondary structures (such as found in p8) are certainly stronger than the one observed in (β) secondary structure (the case of p13). Second, p13 is almost twice as large as p8; the peak superpositions in the proton spectral width are thus more important, and the correlation prediction should be less accurate. Third, the spectral signal-to-noise ratio, as measured with the σ parameter, is better for p8 than for p13.

4.2 False positive and false negative correlations

The general observation of Figs. 2a and b shows two drawbacks of the protocol FIRE: the presence of correlations between residues far away in space (false positive correlations) and the absence of correlations between residues close in space (false negative correlations). The false positive correlations come from the fact that the ^{1}H signals have fortuitous superpositions in the proton spectrum. These superpositions may then create positive match values for otherwise unrelated amide signals. On the other hand, the false negative correlations appear because the presence of false positive correlations creates a bias when filtering the match matrix for a given number of neighbors per residue, through the γ parameter. For these reasons, it is of great relevance to design techniques permitting the detection of the false positive correlations and eventually to remove them.

The use of protein assignments at different temperatures to decrease the number of false positive correlations was tested. **S** matrices were simulated on the *Raphanus sativus* antifungal protein 1 (Table 1) using its PDB file and its assignments at 305 K and 316 K [17, 18]. In the corresponding match matrices **M**, 30 (41) false positives and 43 (49) false negatives are observed at 305 (316) K; the sequential prediction percentages are 74% at 305 K and of 80% at 316 K, respectively. The matrix obtained from the intersection of the two matrices contains only ten false positive correlations, and a comparison of this intersection matrix with the protein proximity matrix shows that the protein fold is correctly predicted. However, the number of false negative correlations (59) increased and the sequential prediction percentage (72%) is of the same order as the smallest sequential percentage obtained at the two temperatures.

The increase in the number of false negatives would generate a lack of constraints on the fold determination, but it is preferable to the errors which could be induced by false positives. Anyway, the use of several temperature measurements requires the acquisition of several 3D NOESY–HSQC experiments and for the signals of the 2D HSQC experiment to be followed during the temperature shift in order to label each protein residue.

4.3 Robustness of the FIRE algorithm

The computation of match values presented here allows the prediction of numerous spatial proximities essential

Table 2. Testing FIRE on a set of proteins

Identification	λ (Hz)[a]	Number of nonzero of columns[b]	Sequential[c] (%)	Medium[d] (%)	Long[e] (%)	γ[f]	Number of neighbors[g]	Number of false positives[h]	% of false positives[i]
acyt	25	100	85	28	15	0.27	4.5	103	19
acyt	35	100	81	21	23	0.27	4.1	95	19
ayj-305 K[j]	25	47	75	39	49	0.21	4.3	39	15
ayj-316 K[j]	25	47	81	26	33	0.22	4.3	44	18
ayj-inter[k]	25	47	72	17	23	–	–	10	6
kum	25	98	69	49	38	0.25	4.5	93	17
kum	35	98	63	41	33	0.24	4.3	100	19
ner	25	60	82	33	17	0.26	4.1	53	17
ner	35	60	83	31	0	0.24	4.2	58	19
snob	25	97	65	45	41	0.27	4.4	96	18
snob	35	97	58	43	36	0.27	4.1	97	19
srl	25	54	74	23	47	0.24	4.4	51	18
srl	35	54	67	29	49	0.23	4.4	54	18
p8	–	65	88	42	9	0.29	4.3	49	14
p13	–	102	68	43	35	0.27	4.4	110	20

[a] Linewidth value used to simulate data sets
[b] Number of nonzero columns in the **S** matrix
[c] Percentage of sequential correlations predicted
[d] Percentage of medium-range correlations predicted
[e] Percentage of long-range correlations predicted
[f] Value of the threshold γ
[g] Mean number of neighbors observed in the **M** matrix
[h] Number of false positives
[i] Percentage of false positives in the sets of predicted match values
[j] Processing of the ayj protein at one temperature (305 or 316 K)
[k] Intersection of the results obtained on ayj at both temperatures

to the determination of the protein fold, but many false positive correlations are also predicted.

Thus, in order to make use of the match calculation in the frame of a real-life project, an assessment of the robustness of the FIRE algorithm as well as an unbiased criterion to evaluate the quality of the proximity prediction are certainly required.

For this purpose, we computed the outcome of the FIRE algorithm on a set of spectra simulated for a set of proteins from Table 1. Several parameters of the simulated data sets (the linewidth, the number of protein residues, and the protein secondary structures) were varied in order to test their influence on the result of FIRE (Table 2). A noise level equivalent to that observed in the spectra of the p8 and p13 proteins was added to all the data sets. From these simulations, it can be observed that the number of false positive correlations increases with the peak linewidth and with the protein size, regardless of the number of prolines in the sequence or the number of columns set to zero. The γ values are in the range 0.21–0.29 and exhibit small variations from one protein to another. Similarly, the percentage of false positives in the set of predicted match values is between 14 and 20% in the case of one-temperature processing. These observations prove the wide applicability of the proposed method, as similar results are obtained for very different proteins and for different peak linewidths.

The percentage of sequential correlation correctly predicted lies between 68 and 88% for the experimental data sets and between 58 and 85% for the simulated ones. The comparison of p8 to srl shows that for the same protein size the sequential proximities are better predicted for an α fold than for a β one. The observation of proteins srl, kum and snob, on one hand, and p8, ner, and acyt, on the other, reveals that for the same secondary α/β or α structure, the sequential prediction percentage does not strongly depend on the protein size, at least for the size range considered here. Analysis of the medium-range prediction percentages indicates two ranges of predictions rates at 20–30% (srl, ner, acyt) and 40–50% (p8, p13, kum, snob). As the two protein groups include proteins of different sizes, as well as α and β secondary structures, it seems that the medium-range prediction is independent of the secondary structure and the protein size. The long-range prediction percentage is in the 10–20% range for p8, ner, and acyt and is in the 35–50% range for the other proteins.

4.4 Estimating the reliability of the proximity prediction by the match value

Additionally, it is of paramount importance to be able to evaluate the reliability of the proximity prediction produced by this approach. For this purpose, we computed the percentage of true predicted proximities according to the match value (Fig. 3), as obtained for the set of proteins studied. The percentage variations display similar features, although the size and secondary structures of the proteins analyzed are different. As the set of proteins used here is varied, the mean value of these percentages should give a good estimation of the reliability of an observed match value with respect to possible proximity between two residues. The continuous line in Fig. 3 presents the function that is used in the current implementation of the program to report reliability to the user.

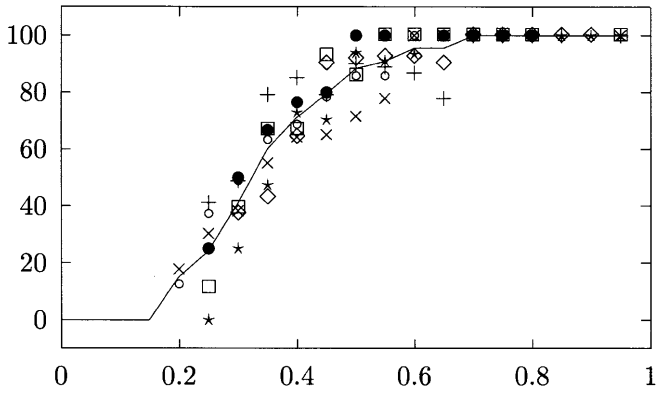

Fig. 3. Percentages of the true predicted proximities displayed according to the match values for proteins p8 (△), p13 (□), acyt (+), kum (×), ner (*), snob (○), and srl (◇). Percentages were calculated using regular intervals of width 0.1. The residue pairs giving rise to a match value in a given interval are selected; on this subset, the percentage of true predicted proximities is calculated as the ratio of the number of residue pairs closer than 4 Å to the total number of pairs. The *continuous line* presents the continuous function chosen to report match reliability to the user

4.5 Comparison of the FIRE protocol with the spectral correlation method

The method presented here has common features with the spectral correlation method proposed by Bartels and Wüthrich [6]. These authors obtained from a 3D ^{15}N HSQC–NOESY spectrum between 49 and 63% of sequential prediction for the overall sequence of different proteins. These results are of a slightly smaller order than the values obtained here (Table 2). The principal difference of their approach to the one presented here comes from the fact that the spectral correlation processing is performed from raw spectral intensities, whereas FIRE does not make use of the information provided by the relative intensity values.

The efficiency of the FIRE results, obtained with reduced information, may seem paradoxical. To further investigate this point, we applied the calculation of Eq. (2) to the experimental pseudospectrum of p8, on which the water signal was filtered and intensity values smaller than 5σ were canceled.

The match matrix obtained in a such way (Fig. 4a) displays many more false positive and negative correlations than the matrix **M** calculated by FIRE (Fig. 2a). Also, the percentages of correct sequential, medium-range and long-range prediction were 44, 16, and 0%, respectively. Two pairs of columns Glu-59/Ser-22 and Met-24/Glu-25 extracted from **S** are typical examples of a false positive (Fig. 4b, c) and a false negative (Fig. 4d, e) correlation. The false positive correlation (Glu-59/Ser-22) is due to large intensity values of autocorrelation peaks and to large aliphatic peaks of Ser-22. The false negative correlation (Met-24/Glu-25) is produced by the weak intensities of peaks other than the autocorrelation peaks.

Bartels and Wüthrich [6] proposed solving the biases induced by very large peaks by scaling down the intensities of the autocorrelation peak and of the intraresidual HN-Hα peaks. However, it may be problematic to identify the intraresidual HN-Hα peaks in certain cases (proximity of the water signal, absence of TOCSY information, etc.) Moreover, different scaling factors should be applied according to the variations of internal mobility among the protein residues. The protocol FIRE is thus more general and more likely to be easily run automatically.

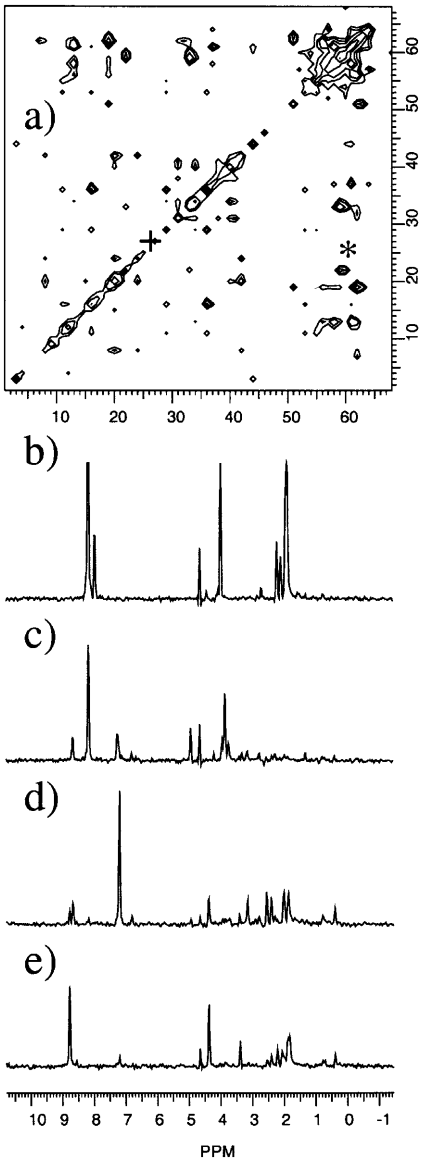

Fig. 4. a The match matrix obtained directly from the experimental spectrum of p8. The processing was performed as described in Sect. 3. The *x*- and *y*-axes show the residue numbers. Two column pairs extracted from the matrix and corresponding to a false positive correlation between Ser-22 (**b**) and Glu-59 (**c**) or to a false negative one between Met-24 (**d**) and Glu-25 (**e**) are shown. The match values of the columns (**b**, **c**) and (**d**, **e**) are indicated by the signs ∗ and + signs, respectively in the match matrix (**a**)

5 Conclusion

We proposed here a processing method which permits protein fold information to be extracted automatically from a 3D NOESY–HSQC experiment. This method was tested on experimental and simulated data sets and provides on average, more than 70% of the protein sequential assignment. The sequential assignment is better predicted for α than for β secondary structures. The approach presented here does not explicitly address the assignment problem, but could be used as a starting point for such an analysis.

All the protein sizes used to test the FIRE approach are smaller than 117 residues. However, the majority of ^{15}N single-labeled structures in the PDB are smaller than 120 residues and FIRE tends to speed up assignment projects for single-labeled proteins; it is thus sufficient to prove the efficiency of FIRE for this size range. The use of FIRE on a 3D ^{15}N NOESY–HSQC for a protein larger than 117 residues should come up against a larger number of false positives.

The selection of the match values corresponding to residues probably close in space was performed using the a priori knowledge of the mean number of neighbors. To the authors' knowledge, the a priori knowledge of chemical shift databases [19–21] of sequence databases [22], or of global geometrical properties [23] has already been used but local geometrical parameters had not been used so far.

Finally, the reliability of the prediction of proximity between two residues can be estimated from their match value. This estimation should be essential in the frame of a complete assignment project.

Acknowledgements. The authors gratefully acknowledge Dr. Laurent Guignard and Dr. Christian Roumestand for providing the experimental data recorded on p8 and p13, and the help of Dr. Franky Fant for providing the assignment of *Raphanus sativus* antifungal protein 1 at 316 K. CNRS, INSERM, and Université Montpellier-1 are acknowledged for funding. TM thanks Dr. Richard Lavery for computer access.

References

1. Wishart DS, Sykes BD (1994) J Biomol NMR 4: 171
2. Nilges M, Macias MJ, O'Donoghue SI, Oschkinat H (1997) J Mol Biol 269: 408
3. Bernstein R, Cieslar C, Ross A, Oschkinat H, Freund J, Holak TA (1993) J Biomol NMR 3: 245
4. Leutner M, Gschwind RM, Liermann J, Schwarz C, Gemmecker G, Kessler H (1998) J Biomol NMR 11: 31
5. Zimmerman DE, Montelione GT (1995) Curr Opin Struct Biol 5: 664
6. Bartels C, Wüthrich K (1994) J Biomol NMR 4: 775
7. Stern MH, Soulier J, Rozenzwajg M, Nakahara K, Canki-Klain N, Aurias A, Sigaux F, Kirsch IR (1993) Oncogene 12: 379
8. Yang YS, Guignard L, Padilla A, Hoh F, Strub MP, Stern MH, Lhoste JM, Roumestand C (1998) J Biomol NMR 11: 337
9. Barthe P, Yang YS, Chiche L, Hoh F, Strub MP, Guignard L, Soulier J, Stern MH, van Tilbeurgh H, Lhoste JM, Roumestand C (1997) J Mol Biol 274: 801
10. Barthe P, Chiche L, Declerck N, Delsuc MA, Lefévre JF, Malliavin T, Mispelter J, Stern MH, Lhoste JM, Roumestand C (1999) J Biomol NMR 15: 271
11. Pons JL, Malliavin TE, Delsuc MA (1996) J Biomol NMR 8: 445
12. Malliavin TE, Pons JL, Delsuc MA (1998) Bioinformatics 14: 624
13. Bernstein FC, Koetzle TF, Williams GJ, Meyer Jr EE, Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M (1977) J Mol Biol 112: 535
14. Seavey BR, Farr EA, Westler WM, Markley JL (1991) J Biomol NMR 1: 217
15. Zhu L, Dyson HJ, Wright PE (1998) J Biomol NMR 11: 17
16. Crippen GM, Havel TF (1988) Distance geometry and molecular conformation. Research Studies, Taunton, UK
17. Fant F, Vranken WF, Martins JC, Borresmans FAM (1997) Bull Soc Chim Belg 106: 51
18. Fant F, Vranken W, Broekaert W, Borremans F (1998) J Mol Biol 279: 257
19. Wishart DS, Watson MS, Boyko RF, Sykes BD (1997) J Biomol NMR 10: 329
20. Wimmer R, Mueller N, Petersen SB (1997) J Biomol NMR 9: 101
21. Pons JL, Delsuc MA (1999) J Biomol NMR 15: 15
22. Hanggi G, Braun W (1994) FEBS Lett 344: 147
23. Chen CC, Chen RO, Altman RB (1996) CABIOS 12: 319
24. Feng Y, Wand AJ, Sligar SG (1991) Biochemistry 30: 7711
25. Jacks AJ, Sorimachi K, Le Gal-Coeffet MF, Williamson DB, Archer G, Williamson MP (1995) Biochemistry 233: 568
26. Gronenborn AM, Wingfield P, Clore GM (1989) Biochemistry 28: 5081
27. Alexandrescu A, Gittis A, Abeygunawardana C, Shortle D (1995) J Mol Biol 250: 134
28. Yu H, Rosen MK, Schreiber SL (1993) FEBS Lett 324: 87