Automated Analysis of Large Sets of Heteronuclear Correlation Spectra in NMR-Based Drug Discovery

Charlotta S. Damberg,[†] Vladislav Yu. Orekhov,[†] and Martin Billeter*[,‡]

Swedish NMR Centre, Göteborg University, Box 465, 405 30 Göteborg, Sweden, and Biochemistry and Biophysics, Göteborg University, Box 462, 405 30 Göteborg, Sweden

Received February 28, 2002

Drug discovery procedures based on NMR typically require the analysis of thousands of NMR spectra. For example, in "SAR by NMR", two-dimensional NMR spectra are recorded for a target protein mixed with ligand candidates from a comprehensive library of small molecules and are compared to the corresponding spectrum for the protein alone. We present an automated procedure for the comparative analysis of large sets of heteronuclear single quantum coherence spectra, which is based on three-way decomposition and implemented as the software package MUNIN. In a single step, spectra with differences in the peak positions (indicating ligand binding) and the affected peaks are identified. By omission of peak picking, ad hoc scoring of the quality of doubtful peaks is avoided. The procedure has been tested on the bacterial ribonuclease barnase, with a protein concentration of only 50 μM, using several small molecules including the substrate analogue 3′-GMP. Sets of 51 spectra were processed simultaneously, and it is concluded that spectra with binding ligands can be unambiguously identified from much larger sets of spectra.

Introduction

High-throughput screening by NMR is a widely used method in drug discovery for the identification of lead compounds for protein targets. There are a number of possible NMR measurements to choose from including the monitoring of chemical shifts, diffusion effects, and nuclear Overhauser effects (NOEs).[1−3] A well-known method called "SAR by NMR" is based on monitoring chemical shift perturbations in two-dimensional heteronuclear single quantum coherence (HSQC) spectra.[4] The idea is to screen large compound libraries with up to 10 000 small molecules [5] with respect to binding to a ^{15}N-labeled target protein. The basic procedure consists of recording an NMR spectrum of the protein only, which is then compared to a similar spectrum where a mixture of potential ligands was added. Typically, each of these mixtures contains about 10 ligands.[3] Differences in chemical shifts for some of the peaks in the spectra indicate that a "hit" compound has been found. No information on assignment or structural information is needed to detect the binding of a compound. However, to locate the binding site on the protein, NMR assignments of the amide signals and knowledge of the three-dimensional structure are required. An advantage of this method, when compared to techniques where ligand signals are observed, is that only protein signals are present in the spectra. This allows identification of strong and weak binding ligands.[4] The limitations are that the protein needs to be ^{15}N-labeled and soluble and NMR assignment methods limit the molecular size of the protein.

MUNIN, an analysis tool for NMR data sets, builds on the assumption that interpretation of NMR spectra is equivalent to classifying their signals or cross-peaks.[6] A three-dimensional NMR data set, which in the present case consists of a series of two-dimensional ^{15}N HSQC spectra, is approximated by a sum of components, each of which characterizes one or several cross-peaks in the data set. A component may describe the HSQC peaks of the backbone H−N group of a given amino acid residue in all spectra where no ligand binding occurs, while another component may describe the same H−N group in those spectra where the signal position is affected by ligand binding. MUNIN achieves this decomposition in a single automatic step and simultaneously for many spectra, avoiding the picking and scoring of cross-peaks. It therefore appears suitable for the purpose of reliable and efficient automated analysis of the large data sets produced by high-throughput NMR screening of ligand libraries versus a target protein.

The extracellular ribonuclease barnase produced by *Bacillus amyloliquefaciens* is a well-characterized protein that has been used extensively in studies of enzyme mechanisms, protein folding, stability, and activity.[7] Hundreds of mutants have been produced and thermodynamically characterized,[8] and various three-dimensional structures are known. Heteronuclear NMR assignments are available.[9] Among the barnase complexes studied is the tight complex between barnase and the substrate analogue 3′-GMP with a dissociation constant $K_d \approx 2$ μM.[10] Here, we use ^{15}N-labeled barnase as an example of a target protein together with a "model" compound library consisting of four small molecules including 3′-GMP.

Methods

Three-way decomposition as a tool for analysis of three-dimensional NMR data sets was implemented in the software package MUNIN and was applied earlier for NOE identifica-

* To whom correspondence should be addressed. Phone: +46 31 773 3925. Fax: +46 31 773 3910. E-mail: martin.billeter@bcbp.gu.se.
[†] Swedish NMR Centre.
[‡] Biochemistry and Biophysics.

10.1021/jm020866a CCC: $22.00
Published on Web 11/14/2002

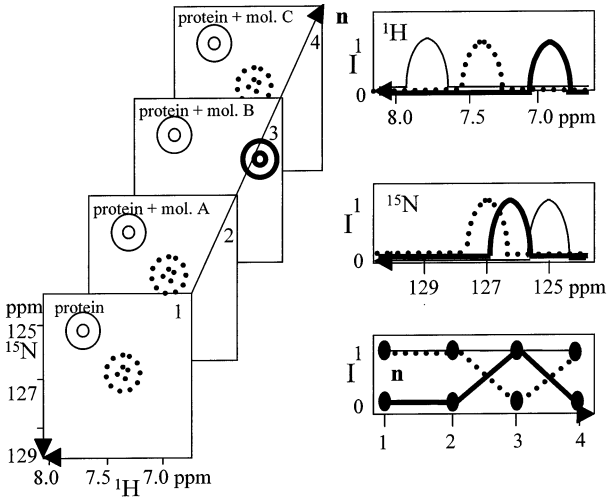

Figure 1. Schematic illustration of the input (left side) and output (right side) of MUNIN when applied to a set of 2D ^{15}N HSQC spectra recorded for a target protein and its mixtures with various potential ligands. The example input consists of four spectra: spectrum 1 is recorded for the target protein only; spectrum 2 is for the protein mixed with a nonbinding molecule A; spectrum 3 is for a mixture with a molecule B that binds to the protein; and spectrum 4 is again for a nonbinding molecule C. The three axes of this 3D data set are the NMR frequency axes for ^{15}N and ^{1}H describing the HSQC spectra and a spectrum enumeration *n*. Only the protein ^{15}N−^{1}H groups yield NMR signals, while the unlabeled molecules A−C are invisible. Peaks (NMR signals) are illustrated by contour lines. The peaks drawn with thin lines stems from a ^{15}N−^{1}H group of the protein located far from the binding site that is unaffected by ligand binding. The peaks drawn with thick dotted lines stem from a protein group near the binding site; it thus shifts in spectrum 3 to a new position indicted by thick solid lines. The MUNIN output contains three *components* identified by the different line types, and each component is described by 1D *shapes* along the three axes for the chemical shifts of ^{15}N and ^{1}H as well as for the enumeration *n*. The shapes along the third axis consist of discrete points for each of the four spectra, which are connected by lines for clarity. Shapes are scaled to cover the interval [0,1] of their intensity *I*. Thus, for the third shape, $I = 1$ means that a peak is present in the corresponding spectrum, and $I = 0$ means that there is no peak.

tion in a ^{15}N−NOESY−HSQC [6] experiment and for the extraction of relaxation rates from a set of ^{15}N HSQC data.[11] The three-way decomposition model implies that a three-dimensional input data set can be decomposed into a relatively small number of *components*. Each component is in turn defined by three one-dimensional *shapes*. Mathematically, this model has been expressed as follows:

$$\mathbf{S} = \sum_{c=1}^{c_{max}} a_c \cdot s_{1c} \otimes s_{2c} \otimes s_{3c} \qquad (1)$$

S is the three-dimensional input data set, and the summation with index *c* runs over all components ($1...c_{max}$). Each component is given by an amplitude a_c and three shapes s_{1c}, s_{2c}, and s_{3c}. (The direct product, ⊗, yields a three-dimensional result from the multiplication of three one-dimensional entities.) Data sets **S** considered here consist of a stack of two-dimensional HSQC spectra, where s_{1c} and s_{2c} describe the two frequency axes while s_{3c} enumerates the spectra. The shape s_{3c} is of greatest interest because it will allow identification of the spectra with ligand binding, while the amplitudes a_c are of minor interest. More details about this model and its use for NMR applications in drug discovery are given in Figure 1 and at the beginning of Results. Note, however, that the only input to MUNIN is the data set **S** and the number c_{max} indicating how many components are expected, while the components described by amplitudes and shapes represent the output.

The test protein, uniformly ^{15}N-labeled barnase,[12] was dissolved in 90% of 50 mM phosphate buffer and 10% D_2O. The pH was adjusted to 4.5. One NMR sample contained only barnase, while four samples were made by adding examples of potential ligands to the protein. In all samples, the final protein concentration was 50 μM and the ligand concentration, if present, was 100 μM. The following small molecules were used: the nucleotide monophosphates 3′-AMP and 3′-GMP, glucose, and glycine. Note that the last two molecules were intended to represent "nonbinding ligands".

Spectra were acquired on a Varian Unity Inova 600 MHz spectrometer at 37 °C. The 2D ^{15}N sensitivity enhanced gradient HSQC experiments[13,14] were performed using a 5 mm triple-resonance PFG probe and ^{15}N decoupling during acquisition. The ^{15}N spectral width was set to 2000 Hz and 128 t_1 increments were collected. The spectral width in the proton dimension was 8000 Hz, and 1024 complex data points were collected. A total of 128 scans were acquired for each increment. The relaxation delay was 1 s, and the total experiment time was about 10 h (note that no cryoprobe was used with the present low protein concentration). Six experiments were run under identical conditions. The order of the experiments was the pure protein, protein with 3′-AMP, glucose, glycine, 3′-GMP, and the pure protein again.

From all six spectra, the regions defined by $126.76 < \omega(^{15}N) < 130.48$ ppm and $7.780 < \omega(^{1}H) < 8.112$ ppm were extracted. The MUNIN program was used with a minor modification that avoids the presence of pairs of large components that are identical in two dimensions; components from such pairs would have to be combined otherwise.[6] Three-dimensional input data sets consisting of various combinations of two-dimensional HSQC spectra were decomposed with different numbers of expected components, and iterations were stopped after a maximum of 1000 steps or when the gradient norm fell below 10^{-12}. Reconstructions of an individual component *c* were calculated by multiplication of the corresponding shapes s_{1c}, s_{2c}, s_{3c} using eq 1, and if requested, reconstructions of several components were added. Reconstructions were formatted in the same way as the input three-dimensional data set, allowing direct comparisons to the input in plots or calculation of the difference between a reconstruction and the input. The shapes along the third dimension s_{3c}, i.e., along the axis enumerating the HSQC spectra, represent the primary result of the decomposition in the drug discovery context. Each of these shapes was scaled such that the maximum takes a value of 1. Shapes along the other dimensions were normalized such that the sum of the squares of the data points is 1. These latter shapes were "demixed"[6] using a separate, fully automatic procedure (manuscript in preparation).

For a MUNIN run with 51 spectra, each experimentally obtained spectrum, except the one for barnase with 3′-GMP, was used to simulate 10 spectra of the same size and with the same peaks. This simulation consisted of several steps. First, the peaks from the region defined above were extracted by selecting all data points above a noise level. These peaks were then transferred to each of 10 regions that contain only noise and were selected from the same spectrum. The transfer was performed by adding the peak data to the new noise regions; thus, the new spectra differ from each other both at the peak positions and outside of those. Finally, all new spectra were translated so that they correspond to $126.76 < \omega(^{15}N) < 130.48$ ppm and $7.780 < \omega(^{1}H) < 8.112$ ppm. This procedure yielded a total of 50 new spectra containing experimentally observed NMR signals superimposed with varying noise.

Results

The basic idea of the application of three-way decomposition to a set of HSQC spectra is illustrated in Figure 1. The input to the procedure is given on the left side of the figure, while the right side shows schematically the

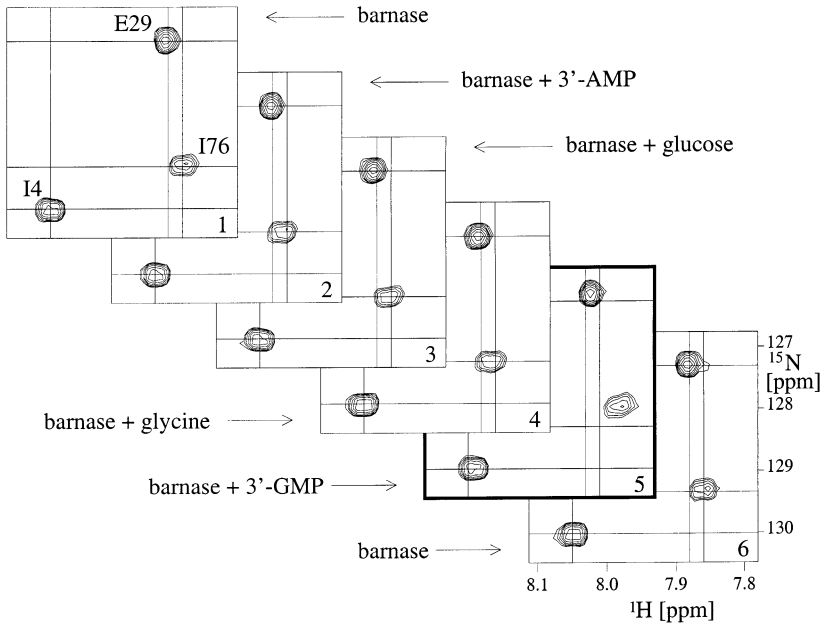

Figure 2. Input data set for MUNIN consisting of six barnase samples. The spectra are numbered in the lower right corner, and the corresponding sample content is indicated beside the spectra. The spectrum recorded for the only mixture with a binding ligand, 3′-GMP, is emphasized by a bold frame. Thin lines are drawn through the peak maxima in spectrum 1 and at identical positions in the other spectra to help identify changes in peak positions. All spectra are plotted using the same contour levels, and only positive contours are shown. The peaks in spectrum 1 are labeled with the residue name and number.

output. The left side shows four spectra, one spectrum recorded for the target protein only and three spectra with the protein plus one small molecule (a potential ligand). Because only the protein is ^{15}N-labeled, all NMR signals stem from the protein and the other molecules remain invisible. However, upon binding to the protein, these molecules affect the protein spectra by moving peaks emerging from the ^{1}H−^{15}N group at the binding site. The output of MUNIN (right side of Figure 1) consists of *components* (identified by different line types in the figure), each of which is described by one-dimensional *shapes* along the three axes. The shapes along the frequency axes resemble one-dimensional cross sections from the two-dimensional spectra; the shapes along the third axis consist of discrete points for the four spectra (connected by lines in the figure). They indicate whether the given component is present (intensity 1) or not (intensity 0) in a spectrum. Since components represent one or several spectral peaks, either at their positions unaffected by binding or at their new positions caused by binding, one can immediately identify in which of the spectra binding has occurred (spectrum 3 in Figure 1; see the caption for more details).

A first input to MUNIN, shown in Figure 2, was composed of six spectra: barnase only, barnase mixed with 3′-AMP, glucose, glycine, 3′-GMP, and again barnase only. For illustration purposes, a small spectral region was selected containing three NMR signals caused by the ^{1}H−^{15}N groups of residues isoleucine 4, glutamic acid 29, and isoleucine 76. The last residue was reported earlier to be significantly affected by the presence and binding of 3′-GMP.[10] Because three peaks are present in the protein only spectrum, and all of these can potentially be affected by binding of a ligand, MUNIN was initially allowed to use six components to describe the set of spectra shown in Figure 2. This set of spectra and the number of expected components represent the only input to the calculation, which lasted about 20 s (SGI 250 MHz R10000 processor). The main result of the calculation, the shapes along the third dimension enumerating the individual HSQC spectra, is shown in Figure 3A. Three of the components identified by the colors black, blue, and green, and hereafter named NONBIND, describe NMR peaks that are present in all spectra except for spectrum 5 with 3′-GMP. The other three components colored yellow, red, and magenta, and referred to as BIND, describe peaks that occur only in spectrum 5. This latter spectrum was recorded for a mixture of barnase and 3′-GMP, which is indeed a substrate analogue that binds to this ribonuclease.[10] The inset in Figure 3A displays the intensities of the six components and shows that these differ by a factor of less than 4.

Thus, Figure 3A identifies not only the mixture where binding occurred, namely, barnase plus 3′-GMP, but also the components that describe the affected NMR signals, i.e., the component group BIND. One may in addition look at the shapes in the other two dimensions, i.e., along the frequency axes for ^{1}H and ^{15}N (Figure 4). Shapes with a single, unique maximum characterize all the components. Occasionally small deviations from pure Gaussian curves are observed, indicating some "cross-talk" between the components. Note also the low noise level in this figure that results even for the components of the BIND group describing peaks encountered in a single spectrum only. Thus, each component describes a single peak in the HSQC spectra, but according to Figure 3A, this peak may be observed in several spectra. Shift changes caused by binding may be quantified using the shapes from Figure 4, i.e., along the frequency axes for ^{1}H and ^{15}N. Thus, one may concentrate on significant shift changes, eliminating small changes that reflect only very weak binding.

Three-dimensional reconstructions (see Methods) using all six components, only the three components from

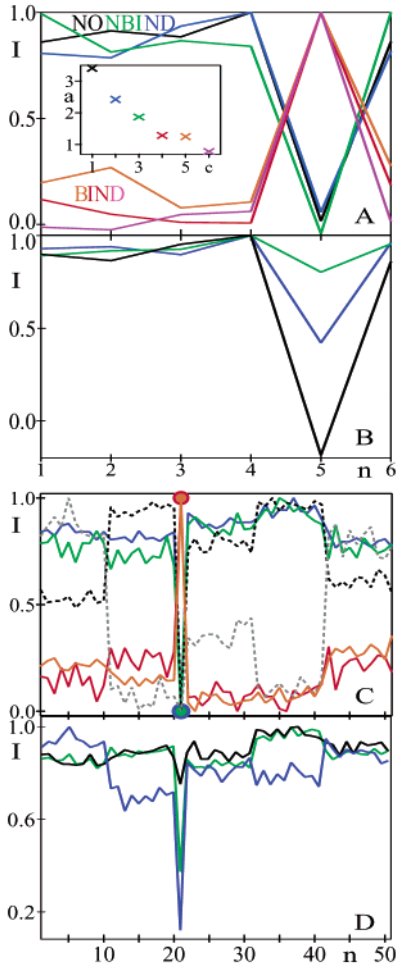

Figure 3. Shapes along the third axis, i.e., the axis enumerating the different input spectra: (A) MUNIN calculation with the input defined by the six spectra of Figure 2 and for six expected components ($c_{max} = 6$); (B) run with the same input but with three expected components; (C) run with 51 input spectra (see text) and for six expected components; (D) run with the same input but with three expected components. All shapes are individually scaled so that their maximum adopts the value 1. In (A) the shapes, i.e., the components they represent, are classified into two groups: group BIND with the red, magenta, and yellow components and group NONBIND with the blue, green, and black components. The inset shows the amplitudes (arbitrary units) of the six components ($c = 1...6$) using the same color code. In (B) all three components describe the peaks in the spectra with no ligand or with nonbonding ligands. The green (and in part the blue) component also characterizes common parts of all six spectra. In (C) the dots and circles mark the positions of the joint maxima of the red and yellow lines and of the black, gray, blue and green lines, respectively. Spectrum 21 is the original spectrum for barnase with 3′-GMP. The red and yellow components describe the spectrum with 3′-GMP, and all other components describe the other 50 spectra. The dashed components characterize differences among these 50 spectra. (The step behavior is an artifact explained in the text.) In (D) all three components again characterize the peaks in the spectra with no ligand or with nonbonding ligands.

the NONBIND group, or only the three components from the BIND group further clarify the result. A reconstruction using all six components reproduces the original set of HSQC spectra with differences not exceeding one contour level in Figure 2. The reconstruction using only the three NONBIND components is shown in the left panels of Figure 5. The first spectrum

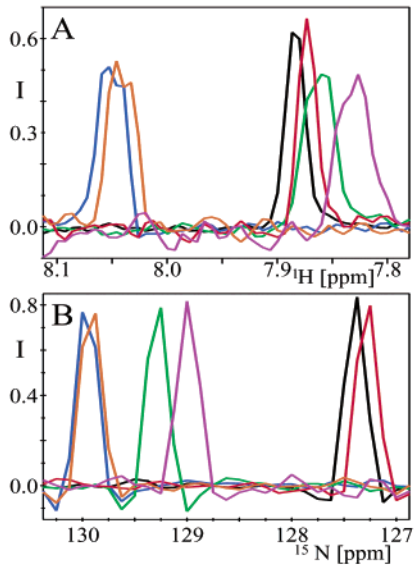

Figure 4. Shapes along (A) the ^{1}H axis and (B) the ^{15}N axis of the spectra that result from the MUNIN run yielding also the shapes in Figure 3A (after automatic demixing). The color code for the shapes is the same as that described for the ones in Figure 3A.

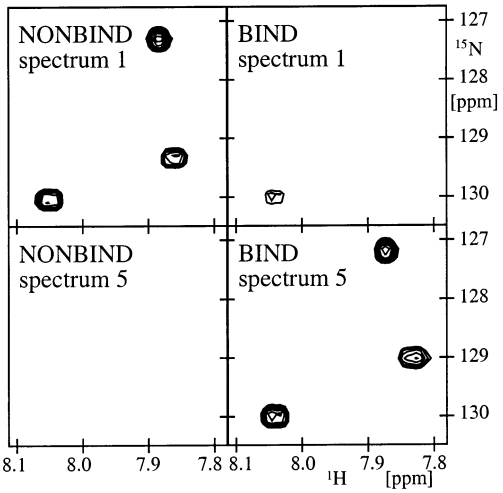

Figure 5. Partial reconstructions using selected groups of components obtained by multiplication of the shapes (from Figures 3A and 4) and amplitudes of a given component. Reconstructions using the NONBIND group of three components (see Figure 3A) are shown in the left panels for spectra 1 and 5 of Figure 2. The right panels show reconstructions of the same spectra of Figure 2 using the three components of the BIND group. The same contour levels are used as in Figure 2.

of this reconstruction (top left in Figure 5) closely resembles spectrum 1 of Figure 2. For spectrum 5 with the mixture of barnase and 3′-GMP, the same reconstruction yields no visible signals at the contour levels chosen (bottom left in Figure 5; the contour levels are the same for Figures 2 and 5). The reconstruction for the BIND group of components results in the contour lines shown in the bottom right panel of Figure 5. It thus properly identifies the occurrence of 3′-GMP binding to barnase as in spectrum 5 of Figure 2. The largest deviation of the reconstructions for the NONBIND and the BIND groups of components from an ideal result is shown by the contour lines in the top right panel of Figure 5. Two contour levels are given by the recon-

struction of the BIND components for the first spectrum with barnase only.

The number of peaks affected by binding is not known a priori. Indeed, the leftmost peak is much less affected than the rightmost, with the one at the top experiencing an intermediate shift (Figure 2). Therefore, the number of components (c_{max} in eq 1) is not well defined when starting MUNIN. Figure 3B shows the main result of the program, the shapes along the third dimension, when the number of allowed components is set to the lowest reasonable number, 3, which corresponds to the number of peaks in the unperturbed spectra. All three components are used to characterize the NMR signals for the spectra with no binding ligand, since these are present in all but one spectrum. The green component in Figure 3B describes the peak for isoleucine 4, which shifts only marginally in the spectrum with 3′-GMP (Figure 2). The component thus can also partly describe this peak in the latter spectrum. To a lesser degree, this also holds for the blue component, which characterizes the peak for glutamic acid 29. However, the major results, i.e., the identification of the spectrum with ligand binding and the characterization of the affected peaks, are not compromised even when MUNIN is allowed to use only a minimal number of components, $c_{max} = 3$.

In real applications, one is faced with many more than six spectra. With a second input data set, the behavior of MUNIN applied to many spectra was investigated. Because we did not have the means to perform such a large number of experiments with many different ligand candidates, we simulated 50 spectra as described in Methods. In the middle of this set of spectra, namely, as number 21, the HSQC spectrum describing the mixture of barnase and 3′-GMP was added (spectrum 5 of Figure 2). The subsequent MUNIN run on all 51 spectra, with again six expected components ($c_{max} = 6$ in eq 1), yielded the shapes along the third dimension shown in Figure 3C. Also here, the single spectrum describing protein−ligand binding, namely, spectrum 21, is easily identified as the one where the red and yellow components adapt a value of 1 and the other components drop to 0. Analysis of the shapes in the other dimensions or reconstruction of the HSQC planes can again determine which peaks are affected by the binding (not shown). Note that the similarity within sets of 10 spectra, which is a consequence of the way they were simulated, yields a visible pattern in Figure 3C, in particular for the components described by the dashed lines. However, this artifact also cannot hide the result of the three-way decomposition. A second calculation for the set of 51 spectra with $c_{max} = 3$ does again show that when MUNIN is allowed this minimal number of components, the identification of the one spectrum affected by ligand binding remains unambiguous (Figure 3D).

A final analysis of the MUNIN result is based on the fact that the decomposition according to eq 1 is not perfect because of noise and other spectral features. Thus, the left and right sides of this equation are not exactly the same, or in other words, the difference between the input data set **S** and the sum of components on the right side of the equation is not everywhere equal to zero. The norm of this difference is referred to as residual, and it represents the entity that MUNIN attempts to minimize. An automatic procedure added to MUNIN provides the contributions to the residual from each spectrum. For the first application discussed above with the six spectra shown in Figure 1 and using $c_{max} = 6$, the residual for all six spectra is similar and close to 0.10 (Table 1, row 1). When $c_{max} = 3$ is used, the average residual rises to 0.19, and this is mainly caused by a residual of 0.55 for the 3′-GMP spectrum; the other five spectra have an average of 0.12 (Table 1, row 2 and footnote). A corresponding analysis with the data set consisting of 51 spectra shows a similar result. The residual for the 50 spectra without binding ligand is low, while the spectrum with 3′-GMP yields a significantly higher residual (Table 1, rows 3−4). This residual is already increased with $c_{max} = 6$ because MUNIN employs four of the six components to describe the 50 spectra unaffected by binding and only two for the 3′-GMP spectrum. This quantitative analysis of the residual thus once again clearly identifies the spectrum with a binding ligand in the case where c_{max} was chosen to be small to allow MUNIN to describe the shifted peaks with their own components.

Table 1. Residuals of the MUNIN Decompositions for Individual HSQC Spectra

number of spectra	number of components, c_{max}[a]	average residual per spectrum[b]	residual for the 3′-GMP spectrum
6	6	0.10 ± 0.01	0.10
6	3	0.19 ± 0.16[b]	0.55
51	6	0.12 ± 0.02	0.27
51	3	0.14 ± 0.08	0.67

[a] See eq 1. [b] The numbers reported are averages and standard deviations for all six (i.e., 51) spectra, including the spectrum with 3′-GMP bound to the protein. For the second row (six spectra and $c_{max} = 3$), the corresponding numbers for all spectra except the 3′-GMP spectrum are 0.12 ± 0.02 (for the other rows, the numbers do not change significantly).

Discussion

Drug discovery applied to a target protein typically starts with high-throughput screening of small-molecule libraries, which involves the testing for binding at the active site of many candidate ligands. It is therefore essential to use an efficient but reliable and controllable approach. Efficiency requires an automated procedure suitable for the analysis of large sets of spectra, while reliability implies avoiding multiple steps with the inherent error source on intermediate data, for example, when preparing lists of spectral peaks. MUNIN reads spectral data and provides directly a list of the spectra affected by binding and thus a list including all ligands to the protein. It avoids the need for peak picking followed by comparisons of peak lists, which may suffer from information loss due to the nonpicking of doubtful peaks or the introduction of ad hoc scores of peak qualities.

Controllable means that the approach can be limited to spectral regions that are of interest and to changes within these. Thus, one may select prior to the MUNIN analysis spectral regions with only peaks of interest, e.g., those at or near the active site. These regions may be concatenated into a new spectrum. The discontinuities that are a consequence of this process have no effect on MUNIN because no assumption on line shapes or

other parameters are made. Alternatively, one may process different regions in parallel, and this approach may be advisable when the whole spectrum or large parts of it are considered. A cutoff for the spectral perturbation can be chosen. The following expression defining changes between the spectrum with the protein only and a test spectrum has been proposed:[15]

$$\Delta\delta = \sqrt{(\delta(H)_{\text{free}} - \delta(H)_{\text{test}})^2 + 0.04(\delta(N)_{\text{free}} - \delta(N)_{\text{test}})^2} \quad (2)$$

Here, the chemical shifts of the amide hydrogen, δ(H), and the nitrogen, δ(N), are given for the free protein and for a test mixture, and all chemical shifts are in ppm. Hajduk et al.[15] consider the spectral perturbation significant if at least two peaks exhibit a value of $\Delta\delta$ greater than 0.1 ppm. This conditions is satisfied for the spectrum of barnase with 3′-GMP with respect to the barnase-only spectrum, since for example the two lysines 27 and 62 show $\Delta\delta$ differences of 0.16 and 0.15 ppm, respectively. The clear detection by MUNIN of the peak shift in Figure 2 between spectra 1 and 5 for isoleucine 76 with a $\Delta\delta$ of 0.07 ppm and the partial detection of peak shifts for glutamic acid 29 with a $\Delta\delta$ of 0.02 ppm show the sensitivity of the method. With the possibility of also detecting small changes, it is trivial to introduce lower limits for changes $\Delta\delta$ using the shapes along ^{1}H and ^{15}N (Figure 4). These limits may be applicable to all or to a subset of the signals; e.g., one may require that at least two peaks have values of $\Delta\delta$ greater than 0.1 ppm.

Other advantages include good sensitivity; in the present application with protein concentrations of only 50 μM, the signal-to-noise ratio was less than 10. The ability to separate strongly overlapped peaks was demonstrated earlier in a relaxation study that was also based on HSQC spectra,[11] and the inherent noise suppression by MUNIN caused by noise collection in the residual of the fit according to eq 1 was also demonstrated.[6] Finally, it should be mentioned that the approach does not rely on any assumptions about line shapes or other spectral parameters.

An important advantage of three-way decomposition is that the unambiguous identification of spectra from protein solutions with bound ligands is not negatively influenced by an increase in the number of spectra. A larger number of unperturbed spectra may force more of the available components to describe features in these spectra. As a consequence, the components will be less influenced by the spectrum with peaks shifted due to ligand binding, and thus, their third shapes will show a clear drop for these latter spectra (Figure 3). In addition, the residual for these spectra will clearly exceed the average residual (Table 1). Thus, the unambiguous identification of these spectra is ensured; only the description of the shifted peaks by their own components (called BIND in Figure 3A) may be "drowned" by a very large number of spectra. However, this description may be easily obtained in a later step once the interesting spectra are identified. This feature of MUNIN in particular makes the use of a very small number of allowed components, $c_{\max}$, unproblematic. It also ensures robustness of the method by avoiding false negative results, i.e., the missing of spectra with ligand binding.

Conclusions

In summary, we have demonstrated the use of three-way decomposition, implemented in the program MUNIN, for efficient analysis of large sets of HSQC spectra that one encounters in high-throughput screening of potential ligands to target proteins. The procedure is fully automatic and requires essentially only one step. It is highly sensitive with respect to both signal-to-noise and observed peak shifts. Very clear results were obtained for a set of 51 spectra, and the identification of spectra affected by ligand binding, observed by a drop of the third shape and a corresponding increase of the per spectrum residual, is not compromised by an increase of the number of spectra in the analyzed data set.

Acknowledgment. The authors thank Dr. Arseniev for providing us with a ^{15}N-labeled barnase sample, the Swedish NMR Centre for instrument time, and NFR (Grant K-AA/KU 12071-302) for support.

References

(1) Stockman, B. J. NMR spectroscopy as a tool for structure-based drug design. *Prog. Nucl. Magn. Reson. Spectrosc.* **1998**, *33*, 109−151.
(2) Moore, J. M. NMR Screening in drug discovery. *Curr. Opin. Biotechnol.* **1999**, *10*, 54−58.
(3) Hajduk, P. J.; Gerfin, T.; Meadows, R. P.; Fesik, S. W. NMR based screening in drug discovery. *Q. Rev. Biophys.* **1999**, *32*, 211−240.
(4) Shuker, S. B.; Hajduk, P. J.; Meadows, R. P.; Fesik, S. W. Discovering high-affinity ligands for proteins: SAR by NMR. *Science* **1996**, *274*, 1531−1534.
(5) Hajduk, P. J.; Bures, M.; Praestgaard, J.; Fesik, S. W. Privileged molecules for protein binding identified from NMR-based screening. *J. Med. Chem.* **2000**, *43*, 3443−3447.
(6) Orekhov, V. Yu.; Ibraghimov, I. V.; Billeter, M. MUNIN: a new approach to multi-dimensional NMR spectra interpretation. *J. Biomol. NMR* **2001**, *20*, 49−60.
(7) Fersht, A. R. Protein folding and stability: the pathway of folding of barnase. *FEBS Lett.* **1993**, *325*, 5−16.
(8) Axe, D. D.; Foster, N. W.; Fersht, A. R. A search for single substitutions that eliminate enzymatic function in a bacterial ribonuclease. *Biochemistry* **1998**, *37*, 7157−7166.
(9) Korzhnev, D. M.; Bocharov, E. V.; Zhuravlyova, A. V.; Tischenko, E. V.; Reibarkh, M. Ya.; Ermolyuk, Ya. S.; Schulga, A. A.; Kirpichnikov, M. P.; Billeter, M.; Arseniev, A. S. ^{1}H, ^{13}C and ^{15}N resonance assignments for barnase. *Appl. Magn. Reson.* **2001**, *21*, 195−201.
(10) Meiering, E. M.; Bycroft, M.; Lubienski, M. J.; Fersht, A. R. Structure and dynamics of barnase complexed with 3′-GMP studied by NMR spectroscopy. *Biochemistry* **1993**, *32*, 10975−10987.
(11) Korzhnev, D. M.; Ibraghimov, I. V.; Billeter, M.; Orekhov, V. Yu. MUNIN: application of three-way decomposition to the analysis of heteronuclear NMR relaxation data. *J. Biomol. NMR* **2001**, *21*, 263−268.
(12) Schulga, A.; Kurbanov, F.; Kirpichnikov, M.; Protasevich, I.; Lobachev, V.; Ranjbar, B.; Chekhov, V.; Polyakov, K.; Engelborghs, Y.; Makarov, A. Comparative study of binase and barnase: experience in chimeric ribonucleases. *Protein Eng.* **1998**, *11*, 775−782.
(13) Kay, L. E.; Keifer, P.; Saarinen, T. Pure absorption gradient enhanced heteronuclear single quantum correlation spectroscopy with improved sensitivity. *J. Am. Chem. Soc.* **1992**, *114*, 10663−10065.
(14) Schleucher, J.; Schwendinger, M.; Sattler, M.; Schmidt, P.; Schedletzky, O.; Glaser, S. J.; Sørensen, O. W.; Griesinger, C. A general enhancement scheme in heteronuclear multidimensional NMR employing pulsed field gradients. *J. Biomol. NMR* **1994**, *4*, 301−306.
(15) Hajduk, P. J.; Boyd, S.; Nettesheim, D.; Nienaber, V.; Severin, J.; Smith, R.; Davidson, D.; Rockway, T.; Fesik, S. W. Identification of novel inhibitors of urokinase via NMR-based screening. *J. Med. Chem.* **2000**, *43*, 3862−3866.

JM020866A