JACS
JOURNAL OF THE AMERICAN CHEMICAL SOCIETY

Article

pubs.acs.org/JACS

A New Algorithm for Reliable and General NMR Resonance Assignment

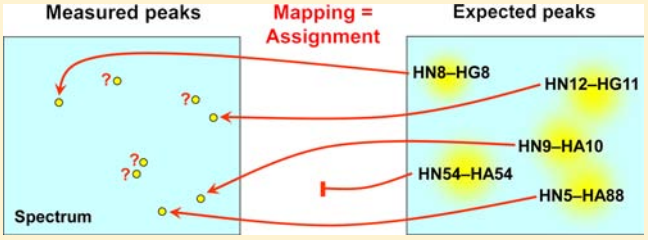
Elena Schmidt and Peter Güntert*

Institute of Biophysical Chemistry, Center for Biomolecular Magnetic Resonance, and Frankfurt Institute of Advanced Studies, Goethe University Frankfurt am Main, Max-von-Laue-Strasse 9, 60438 Frankfurt am Main, Germany

Graduate School of Science and Engineering, Tokyo Metropolitan University, 1-1 Minami-ohsawa, Hachioji, Tokyo 192-0397, Japan

Supporting Information

ABSTRACT: The new FLYA automated resonance assignment algorithm determines NMR chemical shift assignments on the basis of peak lists from any combination of multidimensional through-bond or through-space NMR experiments for proteins. Backbone and side-chain assignments can be determined. All experimental data are used simultaneously, thereby exploiting optimally the redundancy present in the input peak lists and circumventing potential pitfalls of assignment strategies in which results obtained in a given step remain fixed input data for subsequent steps. Instead of prescribing a specific assignment strategy, the FLYA resonance assignment algorithm requires only experimental peak lists and the primary structure of the protein, from which the peaks expected in a given spectrum can be generated by applying a set of rules, defined in a straightforward way by specifying through-bond or through-space magnetization transfer pathways. The algorithm determines the resonance assignment by finding an optimal mapping between the set of expected peaks that are assigned by definition but have unknown positions and the set of measured peaks in the input peak lists that are initially unassigned but have a known position in the spectrum. Using peak lists obtained by purely automated peak picking from the experimental spectra of three proteins, FLYA assigned correctly 96−99% of the backbone and 90−91% of all resonances that could be assigned manually. Systematic studies quantified the impact of various factors on the assignment accuracy, namely the extent of missing real peaks and the amount of additional artifact peaks in the input peak lists, as well as the accuracy of the peak positions. Comparing the resonance assignments from FLYA with those obtained from two other existing algorithms showed that using identical experimental input data these other algorithms yielded significantly (40−142%) more erroneous assignments than FLYA. The FLYA resonance assignment algorithm thus has the reliability and flexibility to replace most manual and semi-automatic assignment procedures for NMR studies of proteins.



INTRODUCTION

The chemical shift assignment of ^{1}H, ^{13}C, and ^{15}N nuclei is an essential prerequisite for protein structure determinations and studies of protein interactions and dynamics by NMR spectroscopy.[1] Despite of the development of methods for automating the chemical shift assignment of proteins that have been reviewed recently,[2−7] the resonance assignments are in most cases still determined manually or by semi-automated methods that require a considerable amount of time by an experienced spectroscopist. Many automated approaches target the question of assigning the backbone and C^{β} chemical shifts, usually on the basis of triple-resonance experiments that delineate the protein backbone through one- and two-bond scalar couplings. Other algorithms are concerned with the more demanding problem of assigning the backbone *and* side-chain chemical shifts. Most of these algorithms require peak lists from a specific set of NMR spectra, and possibly other data, as input and produce lists of chemical shifts of varying completeness and correctness, depending on the quality and information content of the input data and the capabilities of the algorithm.

A recent review by Guerry and Herrmann[7] lists 44 publications of programs performing automated chemical shift assignment published in the past 15 years. Of those, 19 work exclusively on peak lists, while the others require additional input information, e.g., grouping of resonances into spin systems, partial input assignments, three-dimensional (3D) structures, or residual dipolar couplings. Seven of those publications,[8−14] describing five distinct algorithms, apply to the automated assignment of all (backbone and side-chain) resonances exclusively from peak lists. In addition, Guerry and Herrmann[7] showed that only for two of the purely peak list-based algorithms for complete resonance assignment, PINE[8] and GARANT,[9] has their use been reported in protein structure files in the Protein Data Bank (PDB). These algorithms are potentially applicable to enable fully automated NMR structure determination by the classical nuclear Overhauser effect (NOE)-based method that requires extensive side-chain assignments. On the other hand, considering algorithms for

Received: May 25, 2012
Published: July 16, 2012

ACS Publications © 2012 American Chemical Society

dx.doi.org/10.1021/ja305091n | J. Am. Chem. Soc. 2012, 134, 12817−12829

the automated backbone assignment, only the program AutoAssign[15] has been used extensively in a structural genomics project[16] and reported in more than 200 PDB depositions.[7]

This situation calls for new computational approaches that are sufficiently general and reliable to replace most of the diverse manual and semi-automatic assignment strategies with enough accuracy to make extensive manual checking and corrections of the assignments unnecessary.

Here we present in detail the new FLYA automated resonance assignment algorithm that is applicable with a wide variety of NMR spectra and yields more accurate results than other automated assignment methods for all chemical shifts. We show applications of the FLYA resonance assignment algorithm using automatically prepared experimental peak lists, compare its results to those of existing algorithms, and systematically evaluate its performance with simulated data of varying quality.

The new FLYA automated resonance assignment algorithm must be distinguished from the recently introduced FLYA fully automated structure determination algorithm that combined peak picking, resonance assignment with the existing GARANT algorithm,[9] NOESY assignment,[17] and the structure calculation with CYANA into a fully automated procedure.[18] A limitation of this approach was the use of the GARANT program for the resonance assignment because erroneous resonance assignments have a significant impact on the subsequent combined NOESY assignment and structure calculation step.[19] The FLYA automated resonance assignment algorithm replaces the functionality of GARANT within the more comprehensive FLYA fully automated structure determination algorithm[18] but can also be used independent from the other parts of the FLYA fully automated structure determination algorithm.

■ ALGORITHM

The FLYA resonance assignment algorithm generates a network of expected peaks from the protein sequence and the magnetization transfer pathways of a set of NMR experiments. It then computes a mapping from this network to the measured peaks, which implies an assignment of the measured chemical shifts to atoms. This idea had first been implemented in the assignment program GARANT.[9] The optimization algorithm in FLYA is a reimplementation of the procedure that had been developed for GARANT. It is combined with a more flexible network representation and a new scoring scheme for assignments. The FLYA resonance assignment algorithm has been implemented from scratch using the Fortran programming language. It will become available with the next release of the CYANA software.[6,20]

Network Model of the Resonance Assignment. An NMR measurement provides the chemical shifts of a set of atoms that lead to a signal in the measured frequency range. (We use the terms "chemical shift" and "frequency" interchangeably in this paper.) For further use of the data these chemical shifts have to be assigned to the respective atoms of the protein. To facilitate this process a set of different multi-dimensional spectra is used, each combining in its peaks the chemical shifts of atoms matching a specific connectivity pattern. The set of all patterns should provide a contiguous network that represents the connectivity of the atoms in the protein and allows for an unambiguous assignment of the measured frequencies to the atoms.

The network is constructed as follows. Every match of a given connectivity pattern (Figure 1a) to the protein structure

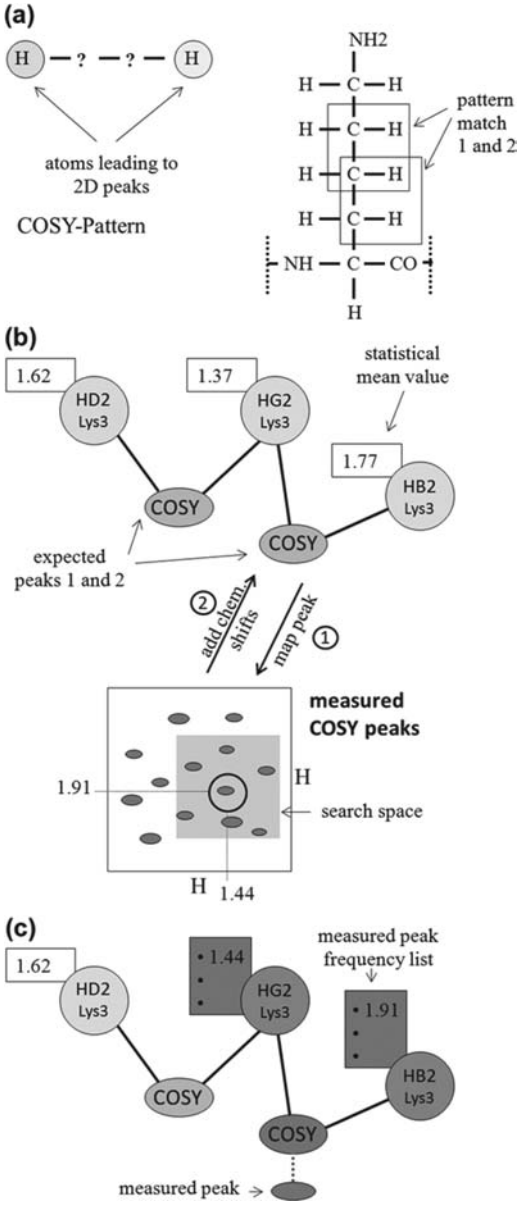


Figure 1. Network model of the resonance assignment, illustrated with the simplest through-bond experiment, 2D homonuclear COSY. (a) H-X-X-H pattern that gives rise to a COSY cross-peak, and schematic view of the matching of patterns with common atoms. (b) Matching of two expected peaks (ellipses labeled "COSY" connected by lines to their corresponding atom assignments) with measured peaks. For one expected peak the search space, obtained from the chemical shift statistics of the BMRB, for matching measured peaks is indicated by the gray rectangle in the spectrum, and the statistical mean values of the chemical shifts are shown in a box adjacent to each atom. (c) After mapping (dotted line) an expected peak to a measured peak, the chemical shifts given by the position of the measured peak are transferred to the lists of measured frequencies of the corresponding atoms (dark gray rectangles).

is expected to lead to a peak in the corresponding experiment that connects the chemical shifts of d atoms, and therefore it is called an "expected peak" (Figure 1b). The expected peaks are labeled with the name of the spectrum in which they are expected to occur. The set of all measurable atoms A and all expected peaks N are the nodes in the network. Connections in the network are added between an expected peak and the

dx.doi.org/10.1021/ja305091n | J. Am. Chem. Soc. 2012, 134, 12817−12829

atoms that the expected peak connects in the experiment. Hence an expected peak is always connected to d atoms (its "adjacent atoms"), where d is the dimensionality of the experiment, and an atom is connected to all expected peaks that result from that atom in the various experiments (its "adjacent peaks"). $N_a \subseteq N$ denotes the set of expected peaks that are connected to atom a. Every atom is labeled with its atom-specific statistical chemical shift range in which it is most likely to be observed. This range is defined by the statistical mean $f(a)$ and the standard deviation$\sigma(a)$ (Figure 1b) that can be obtained, for instance, from the Biological Magnetic Resonance Data Bank (BMRB).[21]

The measured data consist of the set of measured peaks M. During the calculation expected peaks of the network are mapped to measured peaks. If an expected peak n is mapped to a measured peak m, the chemical shifts of the measured peak in the corresponding dimensions give one measured chemical shift for each of the adjacent atoms (Figure 1c). The chemical shift of atom a obtained from mapping the expected peak n to a measured peak m is $f(a, n)$. Each atom saves a list of all its measured chemical shifts. For atom a the average of all entries in this list is the frequency of the atom, $\overline{f}(a)$. Similarly, $\overline{f}^{(k)}(a)$ is the average frequency value of atom a resulting from mappings in experiment k.

A mapping g of the elements in the set of expected peaks N to the elements of the set of measured peaks M is defined by $g(n) = m$ if the expected peak n is mapped to the measured peak m. The mapping gives one possible assignment of the measured chemical shifts to the atoms of the protein. A mapping g is valid if the following conditions are fulfilled: (1) An expected peak can only be mapped to one measured peak. (2) Expected peaks are mapped to measured peaks of the same spectrum. (3) The variation of the chemical shifts $f(a, n)$ for an atom a does not exceed a given tolerance $\varepsilon(a)$ representing the accuracy of the measurement.

To ensure that only mappings fulfilling these conditions are generated, expected peaks are mapped to measured peaks one after another as described below. The search space of an expected peak is defined via the statistical chemical shift ranges of the adjacent atoms that correspond to the respective dimensions of the spectrum. Initially, the search space in each dimension is based on the statistical frequency range for the atom, and one measured peak is selected from within that area.

After a measured peak has been selected, the expected peak is mapped to it and the measured chemical shifts of the peak are added to the chemical shift lists of the adjacent atoms (Figure 1c). As soon as a chemical shift list contains at least one entry, the search space of the atom is limited to the range $\overline{f}(a) \pm \varepsilon(a)$ around the average of the list entries, $\overline{f}(a)$. This procedure assures that the next expected peak to be mapped leads to a list entry that is consistent with the previous ones.

In principle it is possible that several expected peaks are mapped to the same measured peak, but a particular expected peak is only mapped to one measured peak (see condition 1 above). Some expected peaks as well as measured peaks may remain unmapped at the end of the mapping procedure. If no adjacent expected peak of an atom could be mapped to a measured peak, the chemical shift of the respective atom is not defined.

During one run of the algorithm several mappings are generated, improved and combined until the best solution is given as the final output. All these assignment solutions are generated in the way described above. The construction of the individual mappings differs only in the way of selecting among several measured peaks for a mapping as well as in the order in which the expected peaks are mapped.

Expected Peaks. An NMR experiment is set up such that one peak is expected to be measured for each match of an experiment-specific connectivity pattern of atoms to the structure of the protein. These patterns describe covalent bond connectivities mediated by scalar couplings as well as connectivities defined by short distances through space observable via NOEs or corresponding solid-state NMR experiments. Patterns that describe covalent bond connectivities can easily be matched against the covalent protein structure. Each match results in an expected peak n that is expected to be observed in the spectrum with probability prob(n). The connectivity patterns and the respective probabilities are stored in the CYANA library file as linear paths of bonds for each experiment. Examples are shown in Figure 2, and the complete magnetization transfer pathways for all spectra used in this paper are given in Table S1 in the Supporting Information.

```
SPECTRUM N15-HSQC N H
0.98 N:N_AM* H:H_AMI

SPECTRUM HCCH-COSY HC C H
0.98 HC:H_A* C:C_A* H:H_A*
0.98 HC:H_A* C:C_A* C_A* H:H_A*

SPECTRUM HNCA HN N C
0.98 HN:H_AMI N:N_AM* C_A* C_BYL  C:C_ALI
0.80 HN:H_AMI N:N_AMI C_BYL C_ALI N_AMI C:C_ALI
```

Figure 2. Connectivity patterns in the CYANA library for the 2D [^{15}N,^{1}H]-HSQC, 3D HCCH−COSY, and 3D HNCA experiments. For each spectrum, the first line gives the spectrum name and the atom labels that will be used to identify the respective columns in the peaks lists. The number of atom labels defines the dimensionality d of the spectrum. Each of the following lines specifies a (formal) magnetization transfer pathway, characterized by the probability prob(n) of the resulting expected peak followed by a series of atom types (H_AMI, amide hydrogen; N_AMI, amide nitrogen, C_ALI, aliphatic carbon, C_BYL, carbonyl carbon, etc., as used in the CYANA residue library; "*" matches anything) that define a molecular pattern of atoms linked by direct covalent bonds. In each pathway the d atoms whose shifts will determine the position of the resulting peak are identified by their corresponding atom labels, followed by ":". Note that in the case of the HNCA spectrum, the pathways include a "detour" through the carbonyl carbon (C_BYL) to exclude peaks originating from H^{ε}-N^{ε}-C^{δ} in Arg and H^{ζ}-N^{ζ}-C^{ε} in Lys.

The full set of expected peaks for through-space (e.g., NOESY) experiments can only be defined by the correct 3D structure of the protein because they depend on distances between atoms through space. If the 3D structure is available, distances between atoms are determined and expected peaks are generated for every match of the connectivity patterns. To allow for an expected peak generation for NOE-based experiments in cases in which the 3D structure is not available, matches of the respective connectivity patterns can be obtained for short-range NOEs by analyzing the atom distances in a set of random 3D structures of the protein. This was done for all calculations in this paper. Alternatively, since short-range NOEs can be observed only for atoms within a certain sequential range, they can also be described via covalent bonds and bond paths for the respective connectivity patterns can be added to the library.

For a given structure bundle there are two parameters that affect the generation of expected peaks for through-space connectivity patterns. The first parameter gives the maximum distance between the atoms that match the pattern. The second parameter gives the minimum number of conformers in the structure bundle for which this distance criterion has to be fulfilled.

Optimization Strategy. The combinatorial optimization of the mappings uses an evolutionary algorithm in combination with a local optimization routine.[9] Evolutionary algorithms work with a population of solutions and the evolutionary principles of recombination, mutation, and selection.[22] During the calculation several generations of solutions are created one after another. In our algorithm one generation comprises, by default, 50 individuals. Each new generation inherits the best properties of the previous one and introduces new properties called mutations. Early generations have a large amount of mutations. The percentage of allowed mutations is reduced during the optimization procedure according to a temperature schedule. To select good parent solutions for the creation of a new generation, a global scoring scheme is used for the validation of assignment solutions. The assignment that was scored best among all solutions during the optimization is given as the final assignment at the end of the calculation. The local optimization routine is applied to every newly generated solution. It identifies bad parts of an assignment solution on the basis of a local scoring scheme and improves the assignment of these parts if possible.

Global Optimization. Improving a mapping with the local optimization procedure only allows for small changes at a time and the solution would most likely be trapped in a local optimum. The evolutionary algorithm works around this problem.

For every new individual a selection of solutions from the previous generation is marked as parent solutions before the mapping of the expected peaks to the measured peaks is started. The individuals in the previous generation are ranked according to their global score and 30 independent selections are done based on this ranking. For a population of size n a particular solution of rank r is selected with a probability of $(r/n)^{1/2} - ((r-1)/n)^{1/2}$.[9] The new solution is constructed as a combination of the best partial solutions out of the set of its parental solutions. It adopts expected peak mappings that can be found in the parent solutions as follows. To compare the quality of partial solutions, the residue-specific part of the global score (see below) is calculated for all parental solutions (Figure 3a). The combination of the parental solutions is started by selecting randomly one expected peak as starting point for the mapping. The expected peak adopts the mapping from a parental solution selected based on the global score for the respective residue. All n possible mappings are weighted according to their residue-specific score values S_1, ... S_n, and a mapping possibility m is selected with probability $p_m = S_m^{10}/\sum_{i=1}^{n} S_i^{10}$. The expected peak adopts the mapping from the parental solution with the best global score for the respective residue. In order to obtain a consistent assignment, the neighboring expected peaks (the peaks that are connected to at least one atom that is also connected to the starting peak) are mapped next, if it is possible to fulfill the three criteria for a valid assignment, followed by the second shell of neighboring expected peaks and so on (Figure 4). As soon as it is not possible to extend the mapping any further, a new starting peak is determined among the remaining unmapped expected peaks.

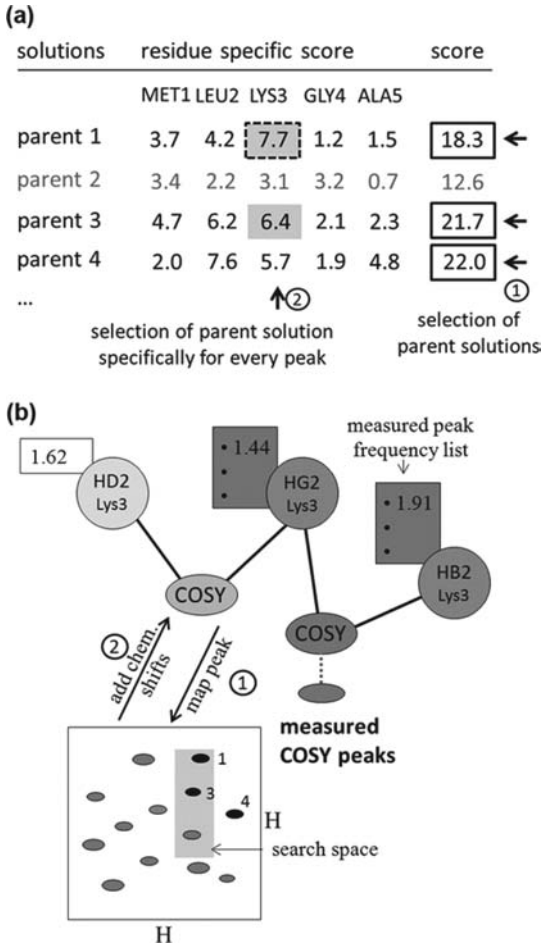


Figure 3. Global optimization of resonance assignments by the evolutionary algorithm. Generation of peak mappings on the basis of parent solutions. (a) Parent solutions 1, 2, and 4 are selected before a new mapping is generated. From those parent solutions that fit the search space (marked in light gray) one solution is selected according to the residue-specific score. (b) Only peaks that fit the search space and can be found in one of the active parent solutions (black peaks) are selected for the mapping.

The mapping of the expected peaks follows the general procedure as described above. An expected peak is mapped to the measured peak that was given in a parental solution and the chemical shift values of the respective measured peak are added to the chemical shift lists of the adjacent atoms, thereby reducing the search space of all unmapped expected peaks that are connected to these atoms.

The selection of a parental solution for an expected peak is done in the same way for all expected peaks, but since an expected peak can only be mapped to a measured peak that fits its search space (Figure 3b), there is not always a valid solution among the parental solutions. When the mapping starts, the search space of an expected peak has its maximum size defined by its statistical distribution and every measured peak it is mapped to in a parental solution will fit to that search space. As soon as some expected peaks are mapped to measured peaks, the search space of the neighboring expected peaks is reduced and it might happen that some or all parental solutions are invalid. Reasons for this are that the correct measured peak might not exist, the mappings of the neighboring expected peaks might be incorrect or, especially in the early stages of the optimization, the correct mapping for an expected peak might

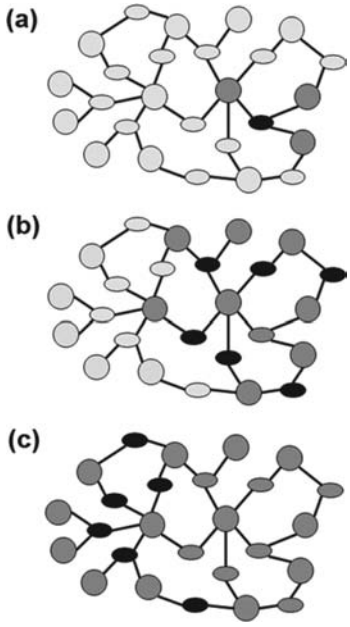


Figure 4. Consistent mapping of expected peaks during global optimization. The scheme shows a network of expected peaks (ellipses) and atoms (circles). (a) A first expected peak (black) is mapped to a measured peak. The adjacent atoms obtain a chemical shift entry (dark gray). All other expected peaks (light gray) are not mapped yet, and the atoms have no chemical shift entries (light gray). (b) The first shell of neighboring peaks (black) is mapped. (c) The second shell of neighboring peaks (black) is mapped.

not yet be present in the set of its parental solutions. In the latter case it is necessary to consider peak mappings that are not part of a parental solution of the expected peak of question. This is done in two ways.

The first method takes advantage of the fact that the correct mapping of so-called "equivalent expected peaks" that have a similar network vicinity can be interchanged easily in the parental solutions. For instance, two leucines will lead to exactly the same network structure with the same statistical mean values for the chemical shifts and hence the same search space for the intraresidual expected peaks. Therefore, the parental solutions of the equivalent expected peaks are also considered for the mapping of an expected peak. This is done by dividing the set of expected peaks into equivalence classes before the calculation starts. Whenever no parental solution of the expected peak fits its current search space, the parental solutions of the equivalent expected peaks are considered with a probability of $\exp(-0.7/T)$ that decreases depending on the temperature schedule of the optimization (see below). One of the mappings that fit the search space of the expected peak is selected randomly. If still no valid mapping was found, all remaining peaks that fit the search space of the expected peak are checked with a probability of $\exp(-1.2/T)$.

The temperature schedule of the optimization starts with a (dimensionless) temperature value of $T = 1.8$, drops to $T = 1.0$ in the next stage, and decreases the temperature during several stages until it reaches $T = 0.0$. For the present calculations the temperature is decreased in steps of 0.2. Finally the temperature is kept constant for another two stages. The changes between the stages are controlled by the average global score of the population. As soon as no further improvement of the average global score can be achieved, the schedule proceeds with the next stage.

Local Optimization. The local optimization routine improves the assignment by taking back the mapping of a small part of the network which is likely to be wrong in order to find a better mapping. The challenge is to identify the incorrect parts of a mapping.

In an ideal mapping each expected peak would be mapped to one measured peak. Since normally many of the peaks that are expected to be observed in a spectrum are missing or overlapped, some expected peaks cannot be mapped to a measured peak, or several expected peaks, in the following called "degenerate expected peaks", have to be mapped to the same measured peak. The correct chemical shift assignment is considered to refer to a mapping in which more expected peaks can be mapped to measured peaks than in other mapping solutions. Hence atoms that are connected to degenerate expected peaks and expected peaks that could not be mapped are likely to be assigned incorrectly. A local scoring scheme (see below) is used to check the quality of the assignment of these atoms. The basic idea is that only a fixed percentage, the worst part, out of the set of all atom assignments that are checked during the local optimization are classified as wrong. The assignments of these atoms are reverted in the following way. The mappings of all adjacent expected peaks of the atom are removed in order to clear completely the list of frequencies of the atom and to allow for a new assignment of the atom. Subsequently, these neighboring expected peaks are mapped again.

The procedure of selecting an expected peak as a starting point for an improvement, removing the assignment of the adjacent atoms that are classified as wrong, and remapping the respective expected peaks is repeated for a fixed number of times, the default setting is 15 000.

Scoring. There are two different scoring schemes. The first one is a global scoring scheme for the evaluation of residue-specific and complete assignment solutions, which is used by the evolutionary optimization procedure. The second one is a local scoring scheme for the evaluation of the assignment of single atoms, which is used by the local optimization routine.

The global score evaluates four attributes of an assignment solution, the frequency of an assigned atom, the ambiguity of the assignment, the difference between several frequencies that are assigned to the same atom, and, implicitly, the number of assigned peaks. The residue-specific part of the global score for residue i is determined by summing up the contributions of all atoms in residues $i-1$, i, and $i+1$.

The global score G is defined by

$$G = \frac{\sum_{a\in A}\left[w_1(a)Q_1(a) + \sum_{n\in N'_a} w_2(a,n)Q_2(a,n)/b(n)\right]}{\sum_{a\in A_0}\left[w_1(a) + \sum_{n\in N_a} w_2(a,n)\right]} \quad (1)$$

A_0 denotes the set of all atoms for which expected peaks exist, $A \subseteq A_0$ the set of assigned atoms, N_a the set of expected peaks for atom a, and $N'_a \subseteq N_a$ the subset of expected peaks that are mapped to a measured peak. $b(n)$ refers to the ambiguity of the assignment and equals the number of expected peaks that are assigned to the same measured peak as expected peak n. Unassigned atoms and unmapped peaks contribute only to the normalization by the denominator in eq 1. As defined below, the term $Q_1(a)$ measures the agreement of the average frequency $\overline{f}(a)$ in the chemical shift list of atom a with the corresponding general chemical shift statistics. Similarly, $Q_2(a,n)$ measures the agreement between the frequency

dx.doi.org/10.1021/ja305091n | J. Am. Chem. Soc. 2012, 134, 12817−12829

$f(a,n)$ of atom a obtained from the measured peak to which the expected peak n is mapped and the average frequency of the atom in the assigned peaks of the corresponding spectrum. Relative weights of the individual contributions are given by $w_1(a)$ and $w_2(a,n)$. We used $w_1(a) = 4$ and $w_2(a,n) = 1$ for all calculations in this paper.

The quality measures Q are designed such that a perfect match corresponds to $Q = 1$, $Q < 1$ in all other cases, a deviation that is considered "as bad as no assignment" yields $Q = 0$, and an infinitely large deviation $Q = -\infty$. To define the quality measures, we consider the logarithm of the probability that the deviation of a frequency from the underlying distribution exceeds the given value by chance. For a normal distribution with average value 0 and standard deviation 1, this quantity is given by

$$q(x) = \log\left(1 - \int_{-|x|}^{|x|} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-t^2/2}\,\mathrm{d}t\right) = \log\left(1 - \mathrm{erf}\left(\frac{|x|}{\sqrt{2}}\right)\right) \quad (2)$$

with the error function $\mathrm{erf}(x) = (2/\pi^{1/2})\int_0^x \mathrm{e}^{-t^2}\,\mathrm{d}t$. For a perfect match eq 2 gives $q(0) = 0$, and in all cases $q(x) \leq 0$. Hence, we define for $i = 1, 2$

$$Q_i = 1 + \frac{q(x_i)}{|q(x_i^{(0)})|} \quad (3)$$

with

$$x_1 = \frac{\overline{f}(a) - f(a)}{\sigma(a)} \quad (4)$$

$$x_2 = \frac{f(a,n) - \overline{f}^{(k)}(a)}{\varepsilon(a)/4} \quad (5)$$

Q_i defined by eqs 3−5 has the aforementioned desired properties. In eq 4, $f(a)$ and $\sigma(a)$ are the mean and standard deviation of the chemical shift of atom a from the general chemical shift statistics for atom a. Similarly, $\overline{f}^{(k)}(a)$ is the average chemical shift for atom a obtained from assigned peaks in the respective spectrum k. $\varepsilon(a)$ denotes the tolerance limit for the alignment of peak positions for atom a. Since $\varepsilon(a)$ is the maximal deviation limit for tolerable assignments, eq 5 uses $\varepsilon(a)/4$ as the standard deviation for the assigned frequencies of atom a. The quantities $x_1^{(0)}$ and $x_2^{(0)}$ determine the deviation that is considered equally bad as the absence of the assignment. An assignment with deviation $|x_i| < x_i^{(0)}$ will contribute positively to the score, i.e. is advantageous, whereas an assignment with deviation $|x_i| > x_i^{(0)}$ will contribute negatively to the score and is thus a disadvantage compared to not making the assignment at all. We used $x_1^{(0)} = 1.5$ and $x_2^{(0)} = 2$ standard deviations for all calculations in this paper.

The global score G of eq 1 is normalized such that $G = 1$ for a perfect assignment, $G < 1$ in all other cases, $G = 0$ if there are either no assignments at all or if all assignments have deviations "as bad as no assignment". In principle, negative scores $G < 0$ are possible for bad assignments. The theoretically possible maximal global score value $G = 1$ cannot be reached in practice because this would require all assigned chemical shifts to be equal to the corresponding mean value of the general chemical shift statistics. Instead, one expects for atoms with chemical shifts that follow the same normal distribution as is assumed for their general chemical shift statistics an average Q_1 value of 0.73.

The scoring scheme for the local optimization considers only the mapping of the expected peaks adjacent to the atom to be scored. The local score $L(a)$ of an atom a is equal to the weighted number of expected peaks that are adjacent to the atom and mapped to a measured peak divided by the weighted total number of expected peaks adjacent to the respective atom:

$$L(a) = \frac{\sum_{n \in N'_a} \mathrm{prob}(n)/b(n)}{\sum_{n \in N_a} \mathrm{prob}(n)} \quad (6)$$

Since a mapping of an expected peak with a high probability $\mathrm{prob}(n)$ that the expected peak can actually be observed in the measurement is considered more reliable and hence more important than a mapping of an expected peak with a low probability, the expected peaks are weighted by $\mathrm{prob}(n)$. The local score takes values between 0 and 1. A local score of 0 means that no expected peaks adjacent to the atom are mapped, and 1 means that all expected peaks are mapped.

Chemical Shift Consolidation. The accuracy of the assignment can be assessed and improved by performing multiple runs of the algorithm using different seeds for the random number generator, and computing for each atom a consensus chemical shift from the values obtained in the individual runs.[18,23] The consensus chemical shift $\tilde{f}(a)$ for an atom a is the value f that maximizes the function

$$\mu(f) = \frac{1}{r}\sum_{j=1}^{r} \exp\left(-\frac{1}{2}\left(\frac{f - \overline{f}_j(a)}{\varepsilon(a)}\right)^2\right) \quad (7)$$

where r is the number of runs, and $\overline{f}_j(a)$ is the average chemical shift value for atom a in run j. To determine the maximum, the function $\mu(f)$ is evaluated on a fine grid with spacing $\varepsilon(a)/100$ or 0.001 ppm, whichever is larger. For all calculations in this paper, $r = 20$ runs were performed.

Assignment Validation. The spread of the distribution of the chemical shift values for an atom in several runs of the algorithm gives an indication of the reliability of the assignment of an atom. This can be quantified the $\mu(\tilde{f}(a))$ value, which is a measure of the self-consistency of the chemical shift values obtained in the individual runs of the algorithm. This quantity can be calculated without knowledge of reference assignments. The value of $\mu(\tilde{f}(a))$ is approximately equal to the fraction of runs that yielded a chemical shift value within the tolerance $\varepsilon(a)$ from the consensus value, $\tilde{f}(a)$. If all chemical shift values are identical, then $\mu(\tilde{f}(a)) = 1$. In this paper we consider assignments with $\mu(\tilde{f}(a)) \geq 0.8$ as "safe", and others as "unsafe" or "tentative". The latter were, however, in many cases still correct.

MATERIALS AND METHODS

Proteins and Experimental Data. Automated chemical shift assignment was performed with the NMR data sets of three proteins for which the assignment and the structure determination had been done earlier by conventional techniques, i.e. the 140-residue ENTH-VHS domain At3g16270(9−135) from *Arabidopsis thaliana* (ENTH) with a seven-α-helix superhelical fold,[24] the 134-residue rhodanese homology domain At4g01050(175−295) from *Arabidopsis thaliana* (RHO) with a central five-stranded parallel β-sheet flanked by four α-helices and two small 3_{10}-helices,[25,26] and the 114-residue Src homology domain 2 from the human feline sarcoma oncogene Fes (SH2) with a central three-stranded anti-parallel β-sheet flanked on

either side by an α-helix, and three short anti-parallel β-strands that pack against the second α-helix.[27,28]

The manually determined chemical shift assignments, which were used as reference assignments to evaluate the assignments from automated procedures, are available from the BMRB with accession numbers 5928 for ENTH, 5929 for RHO and 6331 for SH2. The first 7 and the last 6 residues of these proteins are largely unstructured, non-native sequences related to the expression and purification system.[29] Overall, the completeness of the ^{1}H (excluding labile side-chain protons) reference assignments was 96.4% for ENTH, 98.4% for RHO, and 97.2% for SH2. Excluding the non-native residues, it increased to 99.3% for ENTH, 99.4% for RHO, and 99.6% for SH2.

Peak lists were obtained with the automated peak picking algorithm of the program NMRView[30] without manual corrections or modifications, as reported earlier[18] for [^{15}N,^{1}H]-HSQC, [^{13}C,^{1}H]-HSQC, HNCO, HN(CA)CO, HNCA, HN(CO)CA, CBCANH, CBCA(CO)NH, HBHA(CO)NH, HCCH-COSY (in the case of RHO only for the aromatic region), HCCH-TOCSY, (H)CCH-TOCSY (only for ENTH), H(CCCO)NH, ^{15}N-resolved NOESY, and ^{13}C-resolved NOESY spectra. Peak list statistics are given in Tables S2−S4 in the Supporting Information. All experimental input data are available for download from http://www.cyana.org/flyapeaklists.tgz.

Assignment Calculations with Experimental Data. Assignment calculations with the FLYA resonance assignment algorithm were performed in the same way and with the same parameters for the three proteins. The tolerance for chemical shift matching was 0.03 ppm for ^{1}H and 0.4 ppm for ^{13}C and ^{15}N for all calculations with experimental peak lists. The same tolerances were used for the determination of the assignments and their evaluation by comparison with the manually determined reference assignments. Expected peaks in through-bond spectra were generated according to the magnetization transfer rules of the CYANA library (see Table S1 in the Supporting Information). Expected peaks for the NOESY spectra were generated on the basis of 20 conformers calculated with CYANA that fulfill the steric restraints and are otherwise random. Expected NOESY peaks with probabilities 0.9, 0.8, 0.7, 0.6, and 0.5 were generated for the ^{1}H−^{1}H distances that were shorter than 4.0, 4.5, 5.0, 5.5, and 6.0 Å, respectively, in all 20 random conformers. The number of measured peaks was limited in all input peak lists of through-bond spectra to maximally 150% of the corresponding number of expected peaks by discarding the peaks with the smallest intensities, if necessary. The population size for the evolutionary algorithm was 50, except for the calculations using only backbone assignment spectra, where it was 100, and those using exclusively NOESY peak lists, where it was 200. Hydroxyl protons and the side-chain terminal amide groups of Lys and Arg were excluded from the calculations. Chemical shift assignments were consolidated from 20 independent runs.

Comparison with Existing Automated Assignment Algorithms. For comparison, we determined chemical shift assignments also with the programs GARANT[9,31] and PINE[8] that are capable of determining backbone and side-chain chemical shifts. GARANT calculations were performed with the help of the 'garant.cya' macro in CYANA[18] using the same chemical shift tolerance values as for FLYA. We used the PINE web server at http://pine.nmrfam.wisc.edu/. For all comparisons between algorithms the same peak lists were given to the different algorithms.

Simulated Data Sets for the Protein SH2. Starting from the protein sequence, sets of perfect peak lists were simulated by generating, for the same types of spectra as were used experimentally, all expected peaks at the positions given by the reference chemical shift list. Imperfect peak lists were derived from these perfect peak lists by removing 0, 20, 40, 60, or 80% of randomly selected peaks, by adding 0, 100, 200, 300, 400, or 500% of artifact peaks, and by randomly changing peak positions within the tolerance range for peak assignments, $\varepsilon(a)$. Artifact peaks were generated at positions selected randomly from the normal distributions of the general chemical shift statistics for the atoms involved in randomly selected expected peaks. Peak position changes were obtained by adding a random number taken from a normal distribution with mean zero and given standard deviation to each peak position coordinate. To keep the shifted peaks correctly assignable, random shifts larger than $\varepsilon(a)$ were discarded.

RESULTS AND DISCUSSION

Validity of the Global Score As a Measure of Assignment Correctness. The resonance assignment algorithm strives to maximize the global score G of eq 1. The resulting assignment can only be as good as the global score reflects the correctness of the assignment. The latter is not trivial to evaluate in the absence of independently determined, correct reference assignments. We therefore studied first the relationship between the global score values obtained in the course of assignment calculations with the FLYA resonance assignment algorithm with the correctness of these assignments, evaluated by comparison with the manually determined reference assignments. Figure 5 shows the correlation between

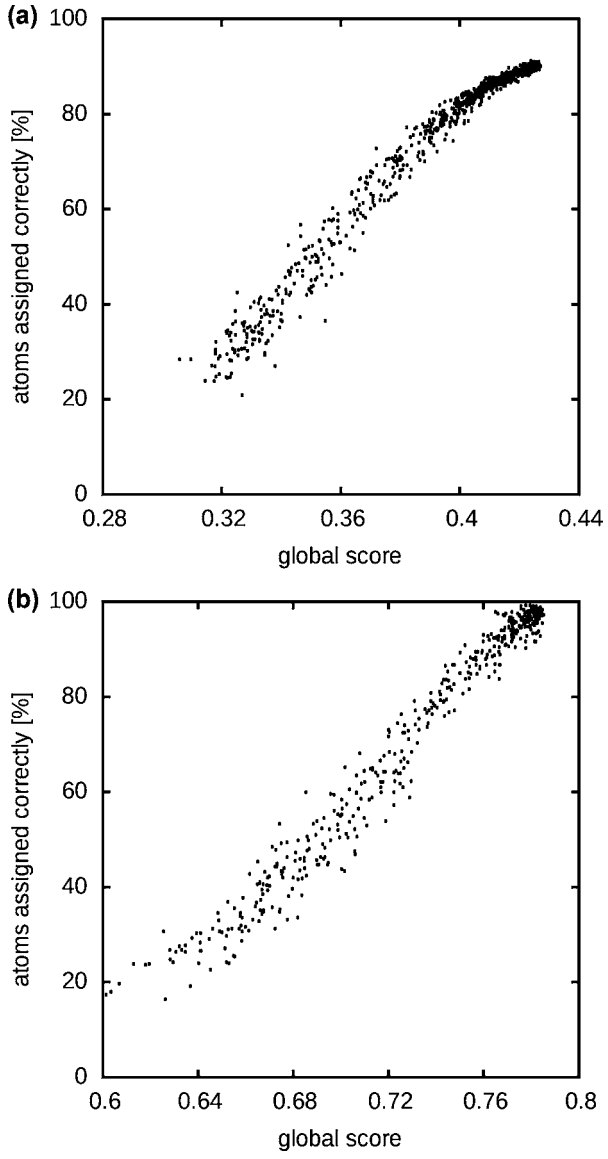


Figure 5. Correlation between the global score and the percentage of correctly assigned atoms. Data points refer to the current best scored solutions, which were saved during the calculation. (a) Standard calculation with the full set of available peak lists for SH2 (Table 1). (b) Calculation with 7 experiments for the backbone assignment for SH2 (Table 2).

the global score and the percentage of correctly assigned atoms for two calculations with the protein SH2. During an assignment calculation the FLYA resonance assignment algorithm stores every new assignment that has a better global score than any previous one. Each of these currently best scoring assignments is represented by a data point in Figure 5. As desired, there is a strong linear correlation with correlation coefficient above 0.98 between the global score values and the percentage of correct assignments. The correlation improves for larger global score values, which are most relevant for reaching an optimal final assignment at the end of the calculation. There are neither significantly erroneous assignments with high global score nor highly correct assignments with low global score. The global score defined by eq 1 is thus a suitable objective function for the optimization algorithm.

Some scatter remains, however, especially for lower global score values. For instance, Figure 5a shows two assignments, both with a global score value of about 0.36 but percentages of correctly assigned atoms of 37% and 58%, respectively. An analysis of these two assignments showed that in the less correct assignment a slightly higher number of expected peaks (7039) could be mapped than in the more correct assignment (7014). This was due to the HCCH-COSY (+61 assigned peaks in the less correct assignment), HCCH-TOCSH (+51), [^{15}N,^{1}H]-HSQC (+5), and [^{13}C,^{1}H]-HSQC (+4) spectra, while in all other spectra less peaks could be mapped, in particular in the backbone assignment spectra HNCA (−21) and CBCANH (−25). This indicates that at this early stage of the assignment calculation additional peaks were assigned in the less reliable HCCH-COSY and HCCH-TOCSY peak lists at the expense of "more valuable" peaks in the less crowded backbone assignment spectra. We also analyzed two backbone assignments in Figure 5b with high global score values of about 0.765 and percentages of correctly assigned atoms of 84% and 92%, respectively. Almost the same number of expected peaks could be mapped in the two cases (1251 and 1253). The global score is not only sensitive to the number of mapped peaks, but also the degeneracy of the assignments. In the present example the better solution shows a slightly higher degeneracy of the assignments. This can be explained by several residues whose assignments are permuted in the worse solution, thereby avoiding mappings of more than one expected peak to the same measured peak and keeping the overall degeneracy low. If peak lists for side-chain assignment were included, at least the permutation of assignments among different types of amino acids would most probably lead to a difference in the number of mapped expected peaks in the side chains. Using only backbone assignment spectra leads to the same number of peaks for almost all types of amino acids, allowing for an easier permutation of residues without loss of mappings. It is conceivable that a systematic investigation of the relationship between the global score and the assignment correctness could lead to future improvements of the scoring and hence the performance of the algorithm.

The global score values for the calculation with the full data set are between 0.3 and 0.5 (Figure 5a). On the other hand, using exclusively through-bond experiments for the backbone assignment yielded global score values between 0.6 and 0.8 (Figure 5b). This difference is due to the fact that the full data set includes experiments, e.g., HCCH-TOCSY and NOESY, with a large number of expected peaks many of which are missing in the experimental peak lists. This leads to lower global score values because the global score depends strongly on the number of expected peaks that are mapped to measured peaks.

Correctness of Assignments. The principal results of automated assignment calculations with FLYA using all available, automatically picked peaks lists, 16 for ENTH, and 15 for RHO and SH2, are summarized in the first column of Table 1. A percentage of correct assignments of 100% would be obtained for an assignment that reproduces, within the tolerance, all chemical shifts that could be assigned by conventional techniques. Atoms that could not be assigned by conventional techniques were ignored when calculating the percentage of correct assignments. The correctness of the FLYA resonance assignments was 95.7−98.8% for the backbone atoms N, H^N, C′, C^α, and C^β, and 86.3−86.5% for the side-chain atoms, resulting in more than 90% correct assignments for all atoms for the three proteins, which fulfills the criterion proposed for the successful use of combined automated NOE assignment and structure calculation with CYANA.[17,19] This shows that the FLYA resonance assignment algorithm yielded almost complete resonance assignments for the entire proteins using input peak lists that were far from perfect, as it is typical for automatically prepared peak lists (Tables S2−S4 in the Supporting Information).

Table 1. Percentage of Correct Assignments[a]

		FLYA	GARANT	FLYA[b]	PINE[b]
ENTH					
	backbone	95.7	95.1	95.1	86.2
	side chain	86.5	77.6	70.4	55.3
	all atoms	90.3	84.7	80.5	67.9
	all atoms, safe[c]	95.3		92.1	
RHO					
	backbone	96.5	95.0	96.1	85.4
	side chain	86.3	73.8	69.3	29.6
	all atoms	90.8	83.0	80.9	53.7
	all atoms, safe[c]	95.3		91.7	
SH2					
	backbone	98.8	96.1	98.6	93.8
	side chain	86.3	82.4	76.3	62.8
	all atoms	91.4	88.0	85.4	75.3
	all atoms, safe[c]	93.8		92.6	

[a]Percentage of chemical shifts that are, within the chemical shift tolerance of 0.03 ppm for ^{1}H and 0.4 ppm for ^{13}C and ^{15}N, in agreement with the manually determined assignment. "Backbone" refers to the atoms that can be assigned using the standard experiments for backbone assignment, i.e. backbone amide N and H^N, C′, C^α, and C^β. "Side chain" refers to the other ^{1}H, ^{13}C, and ^{15}N atoms, including H^α and H^β. "All atoms" includes both groups. For these three classes, 100% corresponds to all manually determined assignments in the class. [b]Calculations performed without peak lists from HCCH-COSY,^{15}N-resolved NOESY, and ^{13}C-resolved NOESY spectra. [c]Percentage of chemical shifts classified as "safe" by the FLYA algorithm that are in agreement with the manually determined assignment. Here, 100% corresponds to all "safe" assignments for which also a manually determined assignment is available.

The individual assignments for these standard FLYA calculations are visualized in Figure 6 and listed in Tables S5−S7 in the Supporting Information. Most of the erroneous assignments occur in side chains, especially in Phe, Lys, and Leu. Assignments of neighboring atoms in several of these

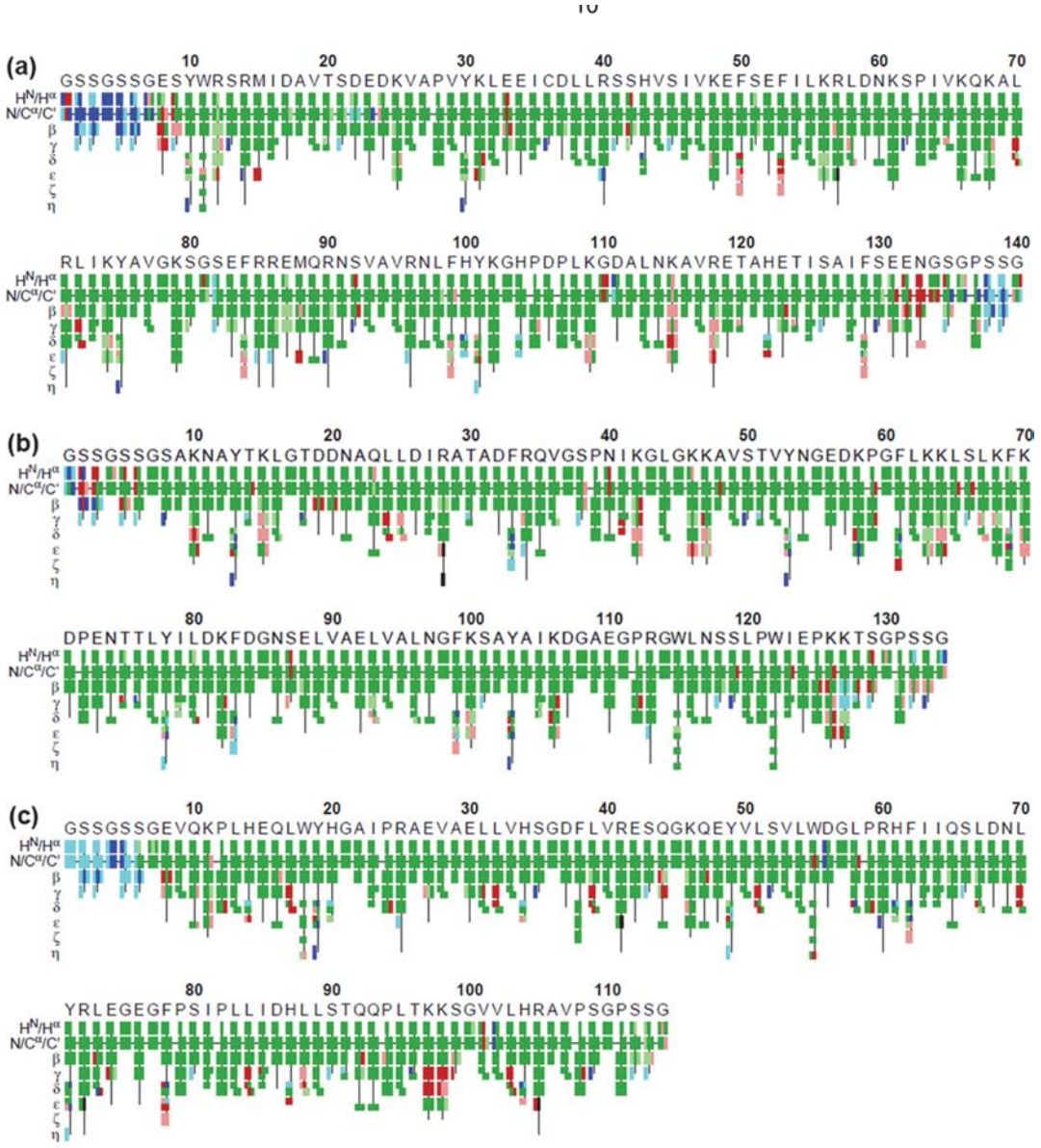


Figure 6. Extent, correctness, and reliability of individual assignments obtained with the FLYA automated resonance assignment algorithm using the full sets of automatically prepared peak lists for three proteins (see Tables S2−S4 in the Supporting Information, and column 1 of Table 1): (a) ENTH, (b) RHO, and (c) SH2. Each assignment for an atom is represented by a colored rectangle: green, assignment by FLYA agrees with the manually determined reference chemical shifts within a tolerance of 0.03 ppm; red, assignment differs from reference; blue, assigned by FLYA but no reference available; black, with reference assignment but not assigned by FLYA. Respective light colors indicate assignments classified as "unsafe" by the chemical shift consolidation. The row labeled H^N/H^α shows for each residue H^N on the left and H^α in the center. The $N/C^\alpha/C'$ row shows for each residue the N, C^α, and C′ assignments from left to right. The rows β−η show the side-chain assignments for the heavy atoms in the center and hydrogen atoms to the left and right. In the case of branched side chains, the corresponding row is split into an upper part for one branch and a lower part for the other branch.

residues were permuted intra-residually. This is due to the fact that the statistical frequencies of the respective atoms do not differ much and that there was only one through-bond experiment in the data set, HCCH-COSY, that could provide unambiguous information to avoid a permutation of these assignments. In addition, for RHO, HCCH-COSY data were available only for the aromatic region. Assignment errors by intra-residual permutation have in general less severe consequences than those involving different residues, especially if the correct and erroneous assignments are far apart in the sequence, which can, for instance, result in erroneous long-range distance restraints in a structure calculation.

The FLYA resonance assignment algorithm classified, without knowledge of the reference assignments, 86% of the assignments for ENTH, 89% for RHO, and 91% for SH2 as "safe" by the criterion $\mu(\tilde{f}(a)) \geq 0.8$. These numbers are close to the overall percentages of correct assignments of 90−91% for the three proteins. Considering only the "safe" assignments the percentage of correct assignments rises to 94−95% (Table 1), i.e. at the expense of losing about 10% of all assignments close to half of the assignment errors can be eliminated. On average, 48% of the erroneous assignments were categorized as "unsafe", and 57% of the assignments for which no reference was present are marked as "unsafe". The effectiveness of the

dx.doi.org/10.1021/ja305091n | J. Am. Chem. Soc. 2012, 134, 12817−12829

safe/unsafe classification is apparent from the fact that an "unsafe" assignment has an 8−12 times higher chance to be erroneous than a "safe" one.

Comparison with Other Automated Assignment Algorithms. The accuracy of the automated assignment calculations was compared to the accuracy obtained with two other programs for automated assignment, GARANT[9,31] and PINE.[8] Results are shown in Table 1. GARANT calculations were done with the full set of peak lists that were used for the FLYA calculation. The PINE web server could not use HCCH-COSY and NOESY-type experiments. The PINE calculations and the corresponding FLYA calculation were therefore performed without these spectra. For the backbone atoms the FLYA automated assignment led to an improvement of the, already high, percentage of correct assignments by 0.6−2.7 compared to the GARANT calculations with the same input data. The respective percentages of correct side-chain assignments were improved by 3.9−12.5. Relative to the PINE results, FLYA increased the percentage of correct assignments by 4.8−10.7 for the backbone and 13.5−39.7 for the side chains. Overall, the calculation with GARANT resulted in a significant 40−85% increase of the number of erroneous assignments relative to the FLYA result. The corresponding increase with PINE was even larger, 65−142%. This indicates that a considerably larger effort for manual verification and correction would be necessary when working with these existing algorithms compared to FLYA.

GARANT calculations with the same input data had already been performed in the context of fully automated structure calculation, where it was reported that 96−97% of all backbone and side-chain chemical shifts in the structured regions were assigned to the correct residues.[18] These numbers are not directly comparable with those in Table 1, because in the former paper the flexible chain termini (13−20 residues), were excluded from the analysis, whereas now the complete sequence was considered, an assignment was counted as correct if the shift was assigned to the correct *residue*, whereas now it has to be assigned to the correct *atom* to be considered as correct, and the results reported were obtained after three cycles of resonance assignments and structure calculation, i.e., the three-dimensional structure was used as additional input to obtain the final resonance assignments, whereas now the resonance assignments are based only on the peak lists. In addition, in the former paper 100% corresponded to the total number of assignments made by the algorithm, whereas now 100% correspond to the total number of reference assignments available from the manual assignment. The relevant comparison of the two algorithms is instead given in Table 1 that shows a significant improvement of the assignment accuracy by the new FLYA resonance assignment algorithm over both earlier approaches.

FLYA Calculations with Reduced Data Sets and with Partial Input Resonance Assignments. Table 2 shows the results for special assignment applications for the same three proteins with different combinations of peak lists. The "backbone" calculation was done using only the peak lists from spectra that are used for backbone assignment, i.e., [^{15}N,^{1}H]-HSQC, HNCO, HN(CA)CO, HNCA, HN(CO)CA, CBCANH, and CBCA(CO)NH. It yielded 92.2−98.4% correct assignments, which is slightly lower than the corresponding percentage for the backbone atoms of 95.7−98.8% obtained in the calculation with the full data set (Table 1), which reflects the synergy of using the spectra for backbone and side-chain assignment simultaneously.

Table 2. Percentage of Correct Assignments in FLYA Calculations with Different Input Data Sets

protein	backbone[a]	NOESY[b]	NOESY + backbone shifts[c]
ENTH	92.2	73.4	84.6
RHO	97.0	77.1	84.3
SH2	98.4	78.3	83.9

[a]Assignment of the backbone atoms N, H^N, C′, C^α, and C^β using [^{15}N,^{1}H]-HSQC, [^{13}C,^{1}H]-HSQC, HNCO, HN(CA)CO, HNCA, HN(CO)CA, CBCANH, and CBCA(CO)NH peak lists as input. [b]Assignment of all atoms, excluding C′, using only 3D ^{15}N-resolved NOESY and ^{13}C-resolved NOESY peak lists as input. [c]Assignment of all atoms using 3D ^{15}N-resolved NOESY and ^{13}C-resolved NOESY peak lists and the backbone chemical shifts (N, H^N, C′, C^α, C^β) as input.

The "NOESY" calculation (Table 2) was done using exclusively ^{15}N-resolved NOESY and ^{13}C-resolved NOESY peak lists.[32] The percentage of correct assignments was between 73.4% and 78.3%, which is lower than in the normal calculation but still considerable given the total absence of through-bond spectra, and, for instance, higher than the results obtained with PINE using all through-bond spectra (Table 1).

The "NOESY + backbone shifts" calculation (Table 2) used as input the ^{15}N- and ^{13}C-resolved NOESY peak lists in conjunction with the backbone chemical shifts fixed to their correct values, in order to perform an automated assignment of the side-chain atoms. It yielded 83.9−84.6% correct assignments, which is higher than the corresponding percentage for the "NOESY" calculation because in the former calculation the side-chain assignments can be "anchored" on the given, correct backbone assignments.

Dependence of Assignments on Data Quality. In order to analyze how the assignment accuracy depends on the quality of the input data, a series of assignment calculations were run using peak lists for SH2 with a defined number of missing peaks, additional noise peaks, and deviations of the peak positions compared to an ideal list in which exactly the expected peaks are present at the exact positions given by the reference chemical shifts. Calculations that yielded more than 90% correct assignments for all atoms, which is considered sufficient for the successful use of combined automated NOE assignment and structure calculation with CYANA,[17,19] are marked in green in Figure 7.

As expected, the correctness of the assignments decreased with increasing numbers of noise peaks and missing peaks, and with increasing deviations of the peak positions. With accurate peak positions the correctness of the assignment is between 77.5% and 99.8% (Figure 7a). It drops below 90% only when 80% of the expected peaks were missing from the input. For a maximum deviation of 0.04 ppm for hydrogens and 0.4 ppm for heavy atoms the correctness of the assignments is between 14.6% and 97.9% (Figure 7b). For the calculations with between 0% and 20% missing peaks additional noise leads to a variation below 10 percentage points and the correctness of the assignments is higher than 90% except for one case (88.6%). With an increasing number of missing peaks the correctness depends more strongly on the number of noise peaks. It drops dramatically if the number of noise peaks is increased from 3 to 4 times the number of expected peaks for the case of 40% missing peaks and if the number of noise peaks is increased

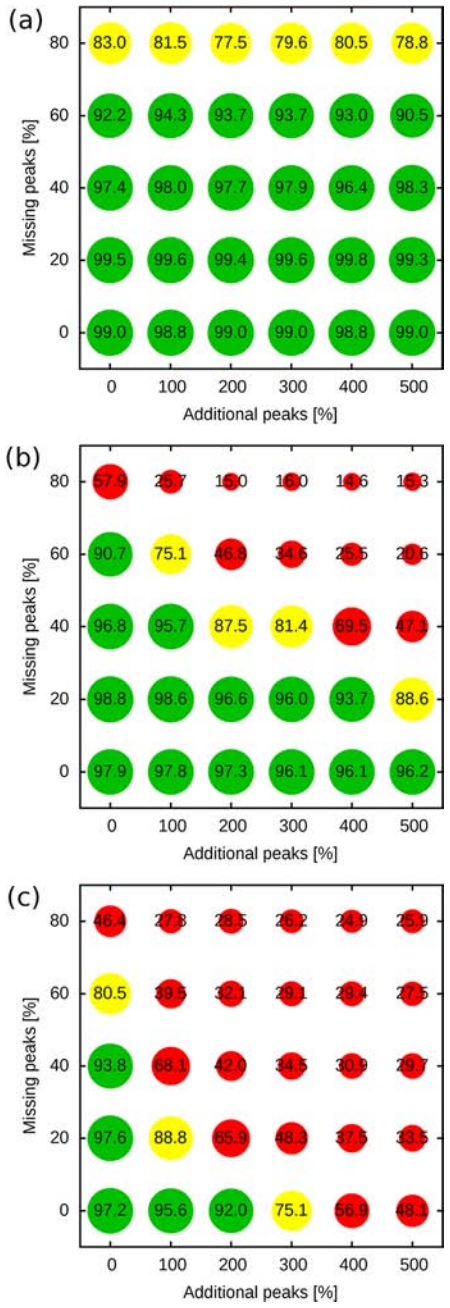


Figure 7. Relationship between the quality of the input peak lists and the correctness of the resonance assignments, evaluated using simulated peak lists for the protein SH2. See text for details. (a) Accurate peak positions. (b) Maximal peak shift of 0.04 ppm for hydrogens and 0.4 ppm for heavy atoms. (c) Maximal peak shift of 0.08 ppm for hydrogens and 0.8 ppm for heavy atoms. The number of artifact peaks was varied between 0 and 5 times the number of expected peaks (horizontal axis). Between 0% and 80% of the correct peaks were deleted (vertical axis). Each circle represents a calculation with input data quality given by the center of the circle. The area of the circle is proportional to the percentage of correct assignments, which is also printed as a number. Calculations with at least 90% correct assignments are shown in green, with 75–90% in yellow, others in red.

from zero to one time the number of expected peaks for the case of 60% missing peaks.

The results of Figure 7b obtained with simulated data are consistent with the results of Table 1 obtained with experimental data. In the measured data of SH2, 45% of all expected peaks cannot be mapped to a measured peak and 109% of the number of expected peaks can be identified as artifacts in the measured peak lists (Table S4). Hence these lists can be compared to simulated lists with 100% artifacts and 40% missing peaks. The calculation with experimental peak lists gave an assignment correctness of 91.4%, which is in good agreement with the 95.7% for the simulated data, considering that the experimental peak lists contain slightly more missing and artifact peaks.

From the test calculations shown in Figure 7b one can draw a general conclusion on the deleterious effects of missing peaks and additional artifact peaks. The data from all calculations with 0.04/0.4 ppm chemical shift tolerance can be combined into a simple relationship between the "data imperfection", defined empirically as the percentage of missing peaks plus 0.08 times the percentage of additional artifact peaks, and the percentage of correct assignments (Figure S1 in the Supporting Information). Other weighting factors than 0.08 do not yield a clear one-parameter relationship. This suggests that on average an additional artifact peak has about 8% of the negative impact of a missing peak, or, in other words, missing correct peaks in a peak list are about 12 times more severe than additional artifact peaks. It is therefore important that the input peak lists contain as many of the expected peaks as possible, even at the expense of picking a considerable number of noise peaks.

For a maximum deviation of 0.08 ppm for hydrogens and 0.8 ppm for heavy atoms, which is higher than usually required with solution NMR data of proteins, the correctness of the assignments is between 24.9% and 97.6% (Figure 7c). The minimum value of 24.9% is 10.3 percentage points higher than the respective value for the calculations with a maximum deviation of 0.04 ppm. This is due to the fact that the assignment of an atom was considered to be correct if the difference between the reference value and the assignment value is below the chemical shift tolerance used for the calculation. Consequently, with a higher chemical shift tolerance more chemical shifts fall into the range randomly, and hence are considered to be correct. For a deviation of 0.08 ppm all calculations with more additional peaks than 3 times the number of expected peaks result in less than 75% correct assignments. At this high tolerance, a correctness of at least 90% can be achieved only with less than 200% noise peaks and no missing peaks, or without noise peaks and less than 60% missing peaks.

The influence of missing peaks on the correctness of the assignments during the calculation is shown in Figure 8. The calculations were done with simulated peak lists. The maximum deviation of the peaks from the ideal position is 0.04 ppm for hydrogens and 0.4 ppm for heavy atoms. No noise peaks were added (first column in Figure 7b). Without missing peaks the number of correct assignments reached an almost optimal value quickly, within about three generations of the evolutionary algorithm. After this point there is only a slight increase in correct assignments. The 20 individual runs were quite uniform; the maximal difference between the correctness of the assignments of the 20 runs at the end of the calculations is less than one percentage point. The 20 runs terminate within less than 40 generations. With an increasing number of missing peaks the slope of the assignment correctness during the calculation decreases and the best values are reached in about 30, 40, and 50 generations of the evolutionary algorithm for 40%, 60%, and 80% of the peaks missing, respectively. It takes

dx.doi.org/10.1021/ja305091n | J. Am. Chem. Soc. 2012, 134, 12817–12829

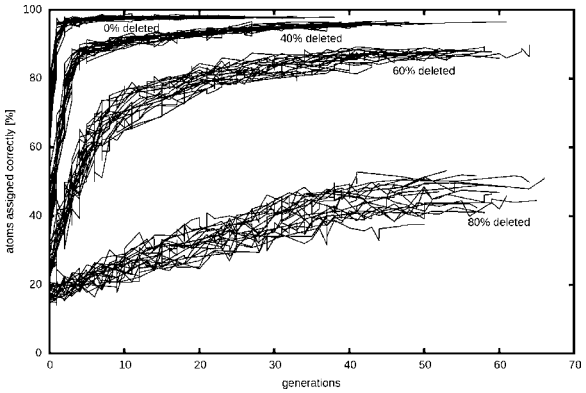


Figure 8. Percentage of correct assignments during optimization, plotted against the generation number of the evolutionary algorithm. Calculations were done using simulated data sets for SH2 with a maximum chemical shift deviation of 0.04 ppm for hydrogens, 0.4 ppm for heavy atoms, and no artifact peaks (leftmost data points in Figure 7b). Results are shown for the data sets without missing peaks, and with 40%, 60%, and 80% peaks missing. For each calculation the 20 individual runs that were consolidated into the final result are plotted.

more generations for the calculations to terminate and the uniformity of the curves decreases. In the extreme case of 80% missing peaks the maximal difference between the correctness of assignments at the end of the calculations was 15.6 percentage points.

Computation Time. The computation time depends mainly on the size of the protein, the number of expected peaks, and the number of measured peaks. A complete assignment calculation for SH2 using simulated peak lists for 15 different experiments with 13 775 expected peaks, a maximum deviation of 0.04 ppm for hydrogens and 0.4 ppm for heavy atoms, no noise peaks and no missing peaks, took 10.5 min on an Intel Xeon X5680 processor. An analogous calculation using 7 backbone experiments, with 1348 expected peaks, took 2.2 min.

CONCLUSIONS

The results of this paper show that the new FLYA automated assignment algorithm can determine resonance assignments for backbone and side-chain shifts with a significantly higher accuracy than existing automated methods that use as experimental input data only peak lists. Provided that the input peak lists are prepared with care such that they represent well the signals of the protein under investigation, the algorithm can yield a similar extent of reliable resonance assignments as can be determined by an experienced spectroscopist with the same data at hand.

In contrast to the widely used automated methods for NOESY cross-peak assignment and the structure calculation,[17,33] automated resonance assignment algorithms that can also treat side-chain atoms have so far not been used routinely in practice. Reasons for the scarce use of automated side-chain resonance assignment algorithms include, presumably, that the often significant number of erroneous assignments required extensive checking of the results. Many spectroscopists consider this as cumbersome as determining the assignments from scratch by conventional semi-automatic approaches. A second reason may be the lack of flexibility of the existing algorithms, which implies that often significant parts of the available experimental data cannot be used in the automated assignment process. Many algorithms require a pre-interpretation of the input data, for instance by assigning the atom type (e.g., C^{α}, C^{β}, etc.) or grouping chemical shifts into spin systems, or even short stretches of connected residues, all of which are equivalent to making partial assignments before starting the automated algorithm. A third, practical reason may be that algorithms can be difficult to use, e.g. because of strong parameter dependence, non-standard format requirements, lack of documentation, or high computation time demands. The present FLYA resonance assignment algorithm largely overcomes these limitations: The accuracy of the assignments is high, and by consolidating chemical shifts from multiple runs, FLYA can distinguish reliable from only tentative assignments. Both features significantly reduce the work for manually checking the results. The FLYA resonance assignment algorithm provides high flexibility with regard to the spectra that can be used for preparing the input peak lists. It can also be used with solid state NMR data, for structure-based assignment,[34,35] resonance assignment based exclusively on NOESY spectra,[32] the assignment of homologous proteins given the shifts and/or the structure of a related protein,[31] and the resonance assignment of nucleic acids, which will be treated elsewhere. Using the FLYA resonance assignment algorithm is straightforward. It requires only the amino acid sequence and peak lists in which the peak coordinates corresponding to the same atom are aligned (within each peak list and between different peak lists) within the chemical shift matching tolerance, which is the only important parameter that has to be specified in accordance with the quality of the peak lists. The computation time requirements of the FLYA resonance assignment algorithm are also not a limiting factor.

A principal advantage of the FLYA resonance assignment algorithm is that it integrates all different assignment steps into one procedure that uses all data simultaneously rather than sequentially. Therefore it does not rely on the correctness of earlier preliminary assignment results that can in practice not be guaranteed. The principal limitation of the present version of the FLYA resonance assignment algorithm is that it uses peak lists as input, which furthermore remain invariable throughout the assignment calculation. Peak lists are an—always imperfect—abstraction of the original experimental NMR spectra. An advantage of manual and semi-automatic approaches is that while making assignments the original spectra can be re-inspected in the light of already gained understanding, which allows the identification of overlapped or weak signals and/or the elimination also of non-obvious artifacts, as well as the implicit use of information that is not contained in peak lists, e.g., peak shapes. We expect that in the future the FLYA resonance assignment algorithm can be improved further by directly accessing the spectra or by automatically refining the peak lists during the assignment calculation, as it has been done successfully for protein NMR structure determination with automated NOE identification in the NOESY spectra.[36]

ASSOCIATED CONTENT

Supporting Information

Rules for generating expected peaks, peak list statistics, and detailed resonance assignment results. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

guentert@em.uni-frankfurt.de

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

We gratefully acknowledge financial support by the Lichtenberg program of the Volkswagen Foundation and by the Deutsche Forschungsgemeinschaft (DFG).

REFERENCES

(1) Wüthrich, K. *NMR of Proteins and Nucleic Acids*; Wiley: New York, 1986.

(2) Moseley, H. N. B.; Montelione, G. T. *Curr. Opin. Struct. Biol.* **1999**, *9*, 635−642.

(3) Gronwald, W.; Kalbitzer, H. R. *Prog. Nucl. Magn. Reson. Spectrosc.* **2004**, *44*, 33−96.

(4) Baran, M. C.; Huang, Y. J.; Moseley, H. N. B.; Montelione, G. T. *Chem. Rev.* **2004**, *104*, 3541−3555.

(5) Altieri, A. S.; Byrd, R. A. *Curr. Opin. Struct. Biol.* **2004**, *14*, 547−553.

(6) Güntert, P. *Eur. Biophys. J.* **2009**, *38*, 129−143.

(7) Guerry, P.; Herrmann, T. *Q. Rev. Biophys.* **2011**, *44*, 257−309.

(8) Bahrami, A.; Assadi, A. H.; Markley, J. L.; Eghbalnia, H. R. *PLoS Comp. Biol.* **2009**, 5.

(9) Bartels, C.; Güntert, P.; Billeter, M.; Wüthrich, K. *J. Comput. Chem.* **1997**, *18*, 139−149.

(10) Eghbalnia, H. R.; Bahrami, A.; Wang, L. Y.; Assadi, A.; Markley, J. L. *J. Biomol. NMR* **2005**, *32*, 219−233.

(11) Li, K. B.; Sanctuary, B. C. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 467−477.

(12) Schmucki, R.; Yokoyama, S.; Güntert, P. *J. Biomol. NMR* **2009**, *43*, 97−109.

(13) Xu, J.; Sanctuary, B. C.; Gray, B. N. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 475−489.

(14) Xu, J.; Straus, S. K.; Sanctuary, B. C.; Trimble, L. *J. Magn. Reson. B* **1994**, *103*, 53−58.

(15) Zimmerman, D. E.; Kulikowski, C. A.; Huang, Y. P.; Feng, W. Q.; Tashiro, M.; Shimotakahara, S.; Chien, C. Y.; Powers, R.; Montelione, G. T. *J. Mol. Biol.* **1997**, *269*, 592−610.

(16) Huang, Y. P. J.; Moseley, H. N. B.; Baran, M. C.; Arrowsmith, C.; Powers, R.; Tejero, R.; Szyperski, T.; Montelione, G. T. *Methods Enzymol.* **2005**, *394*, 111−141.

(17) Herrmann, T.; Güntert, P.; Wüthrich, K. *J. Mol. Biol.* **2002**, *319*, 209−227.

(18) López-Méndez, B.; Güntert, P. *J. Am. Chem. Soc.* **2006**, *128*, 13112−13122.

(19) Jee, J.; Güntert, P. *J. Struct. Funct. Genom.* **2003**, *4*, 179−189.

(20) Güntert, P.; Mumenthaler, C.; Wüthrich, K. *J. Mol. Biol.* **1997**, *273*, 283−298.

(21) Ulrich, E. L.; Akutsu, H.; Doreleijers, J. F.; Harano, Y.; Ioannidis, Y. E.; Lin, J.; Livny, M.; Mading, S.; Maziuk, D.; Miller, Z.; Nakatani, E.; Schulte, C. F.; Tolmie, D. E.; Wenger, R. K.; Yao, H. Y.; Markley, J. L. *Nucleic Acids Res.* **2008**, *36*, D402−D408.

(22) Bäck, T. *Evolutionary Algorithms in Theory and Practice*; Oxford University Press: 1996.

(23) Malmodin, D.; Papavoine, C. H. M.; Billeter, M. *J. Biomol. NMR* **2003**, *27*, 69−79.

(24) López-Méndez, B.; Pantoja-Uceda, D.; Tomizawa, T.; Koshiba, S.; Kigawa, T.; Shirouzu, M.; Terada, T.; Inoue, M.; Yabuki, T.; Aoki, M.; Seki, E.; Matsuda, T.; Hirota, H.; Yoshida, M.; Tanaka, A.; Osanai, T.; Seki, M.; Shinozaki, K.; Yokoyama, S.; Güntert, P. *J. Biomol. NMR* **2004**, *29*, 205−206.

(25) Pantoja-Uceda, D.; López-Méndez, B.; Koshiba, S.; Inoue, M.; Kigawa, T.; Terada, T.; Shirouzu, M.; Tanaka, A.; Seki, M.; Shinozaki, K.; Yokoyama, S.; Güntert, P. *Protein Sci.* **2005**, *14*, 224−230.

(26) Pantoja-Uceda, D.; López-Méndez, B.; Koshiba, S.; Kigawa, T.; Shirouzu, M.; Terada, T.; Inoue, M.; Yabuki, T.; Aoki, M.; Seki, E.; Matsuda, T.; Hirota, H.; Yoshida, M.; Tanaka, A.; Osanai, T.; Seki, M.; Shinozaki, K.; Yokoyama, S.; Güntert, P. *J. Biomol. NMR* **2004**, *29*, 207−208.

(27) Scott, A.; Pantoja-Uceda, D.; Koshiba, S.; Inoue, M.; Kigawa, T.; Terada, T.; Shirouzu, M.; Tanaka, A.; Sugano, S.; Yokoyama, S.; Güntert, P. *J. Biomol. NMR* **2005**, *31*, 357−361.

(28) Scott, A.; Pantoja-Uceda, D.; Koshiba, S.; Inoue, M.; Kigawa, T.; Terada, T.; Shirouzu, M.; Tanaka, A.; Sugano, S.; Yokoyama, S.; Güntert, P. *J. Biomol. NMR* **2004**, *30*, 463−464.

(29) Kigawa, T.; Yabuki, T.; Yoshida, Y.; Tsutsui, M.; Ito, Y.; Shibata, T.; Yokoyama, S. *FEBS Lett.* **1999**, *442*, 15−19.

(30) Johnson, B. A.; Blevins, R. A. *J. Biomol. NMR* **1994**, *4*, 603−614.

(31) Bartels, C.; Billeter, M.; Güntert, P.; Wüthrich, K. *J. Biomol. NMR* **1996**, *7*, 207−213.

(32) Ikeya, T.; Jee, J.-G.; Shigemitsu, Y.; Hamatsu, J.; Mishima, M.; Ito, Y.; Kainosho, M.; Güntert, P. *J. Biomol. NMR* **2011**, *50*, 137−146.

(33) Nilges, M.; Macias, M. J.; ODonoghue, S. I.; Oschkinat, H. *J. Mol. Biol.* **1997**, *269*, 408−422.

(34) Stratmann, D.; Guittet, E.; van Heijenoort, C. *J. Biomol. NMR* **2010**, *46*, 157−173.

(35) Apaydın, M. S.; Conitzer, V.; Donald, B. R. *J. Biomol. NMR* **2008**, *40*, 263−276.

(36) Herrmann, T.; Güntert, P.; Wüthrich, K. *J. Biomol. NMR* **2002**, *24*, 171−189.

NOTE ADDED AFTER ASAP PUBLICATION

Due to a production error Figures 7 and 8 were incorrect in the version published ASAP July 23, 2012. The correct Figures reposted July 24, 2012.