GARANT—A General Algorithm for Resonance Assignment of Multidimensional Nuclear Magnetic Resonance Spectra

CHRISTIAN BARTELS, PETER GÜNTERT, MARTIN BILLETER, and KURT WÜTHRICH*

Institut für Molekularbiologie und Biophysik, Eidgenössische Technische Hochschule-Hönggerberg, CH-8093 Zürich, Switzerland

Received 2 January 1996; accepted 30 April 1996

ABSTRACT

A new program for automatic resonance assignment of nuclear magnetic resonance (NMR) spectra of proteins, GARANT (General Algorithm for Resonance AssignmeNT), is introduced. Three principal elements used in this approach are: (a) representation of resonance assignments as an optimal match of two graphs describing, respectively, peaks expected from combined knowledge of the primary structure and the magnetization transfer pathways in the spectra used, and experimentally observed peaks; (b) a scoring scheme able to distinguish between correct and incorrect resonance assignments; and (c) combination of an evolutionary algorithm with a local optimization routine. The score that evaluates the match of expected peaks to observed peaks relies on the agreement of the information available about these peaks, most prominently, but not exclusively, the chemical shifts. Tests show that the combination of an evolutionary algorithm and a local optimization routine yields results that are clearly superior to those obtained when using either of the two techniques separately in the search for the correct assignments. GARANT is laid out for assignment problems involving peaks observed in two- and three-dimensional homonuclear and heteronuclear NMR spectra of proteins.

*Author to whom all correspondence should be addressed.

CCC 0192-8651 / 97 / 010139-11

Introduction

Nuclear magnetic resonance (NMR) spectroscopy has by now been well established as a method for three-dimensional structure determination of biological macromolecules in solution,[1,2] and much current work is focused on further improvement of the efficiency of NMR structure determination. Thereby, the large amount of data typically encountered calls for extensive use of computer support. On principal grounds, programs for automatic resonance assignment promise to provide particularly efficient, objective and reliable handling of the large data sets, but in practice spectral artifacts such as noise bands, absence of signals because of fast relaxation, or accidental overlap of resonances have limited the use of such routines.[3] To overcome these limitations the presently proposed General Algorithm for Resonance AssignmeNT (GARANT) simultaneously uses the peak positions from multiple experimental spectra to eliminate influences of spectral artifacts in the determination of the resonance assignments. This means, for example, that the spin system assignments established using correlated spectroscopy are directly included in the search for sequential connectivities in the nuclear Overhauser effect (NOE) data sets, and sequential connectivities observed by nuclear Overhauser spectroscopy (NOESY) are in turn used during the searches for spin system identification in the *J*-correlation spectra. In this way, improved robustness against incomplete or partially corrupted experimental input data is achieved.

Depending on the size of the protein and on the isotope labeling strategy, GARANT can be used for combined analysis of homonuclear and heteronuclear experiments, two-dimensional (2D) and higher spectra, and experiments using COSY-type mixing as well as TOCSY- or NOESY-type mixing.[4–6]

The structure of GARANT is based primarily on a reliable representation of resonance assignments by projection of expected assignments onto the experimental data, a general criterion for evaluation of the quality of preliminary resonance assignments, and an efficient optimization algorithm. These three elements of the program are explained in the next section. In the ''Results and Discussion'' section, the general applicability of GARANT is substantiated and the functional roles of different parts of the algorithm are analyzed.

General Algorithm for Resonance Assignment (GARANT)

REPRESENTATION OF ASSIGNMENTS

The basic information used by the GARANT program for determination of resonance assignments consists of: (i) the primary structure of the protein; (ii) lists of cross peaks observed in the experimental spectra; and (iii) knowledge about magnetization transfer pathways in the NMR experiments used[4] (the results obtained with GARANT can be greatly improved when supplementary information on homologous proteins is available, for example, the three-dimensional structure or the chemical shifts; see Bartels et al.[7]). With GARANT, a representation of resonance assignments is then obtained by deriving, from the amino acid sequence and knowledge about the NMR experiments, the cross peaks that are expected to be present in the spectra. The expected cross peaks are correlated with the peaks observed in the corresponding experimental spectra, and the best match found between expected and observed peaks yields the resonance assignment.

In Figure 1A, three expected cross peaks (shown as boxes labeled with the spectrum types) and the 2D COSY and 2D NOESY relations anticipated among them are represented by a graph. In each dimension ($n = 2$ in Fig. 1A) each expected peak is assigned to an atom or a group of spectroscopically equivalent atoms (e.g., the hydrogen atoms in a methyl group). In a similar graph for representation of the experimental spectra, the position of each cross peak defines a chemical shift in each dimension (Fig. 1B). In both graphs (Figs. 1A and 1B) cross peaks and atoms, or chemical shifts, respectively, are identified with vertices, and relations between cross peaks and atom types, or chemical shifts, respectively, are represented with edges of the graph. A match between the two graphs defines a resonance assignment (Fig. 1C): Expected peaks and atom types are mapped onto observed peaks and chemical shifts such that there are identical connectivities for corresponding expected and observed cross peaks. The chemical shifts of the observed cross peaks then define the resonance frequencies and the atoms assigned to the expected peaks define the assignments. Due to imperfections in the lists of observed peaks and limitations in the derivation of the lists of expected peaks (e.g., long-range NOEs cannot be predicted without knowledge about the three-dimensional

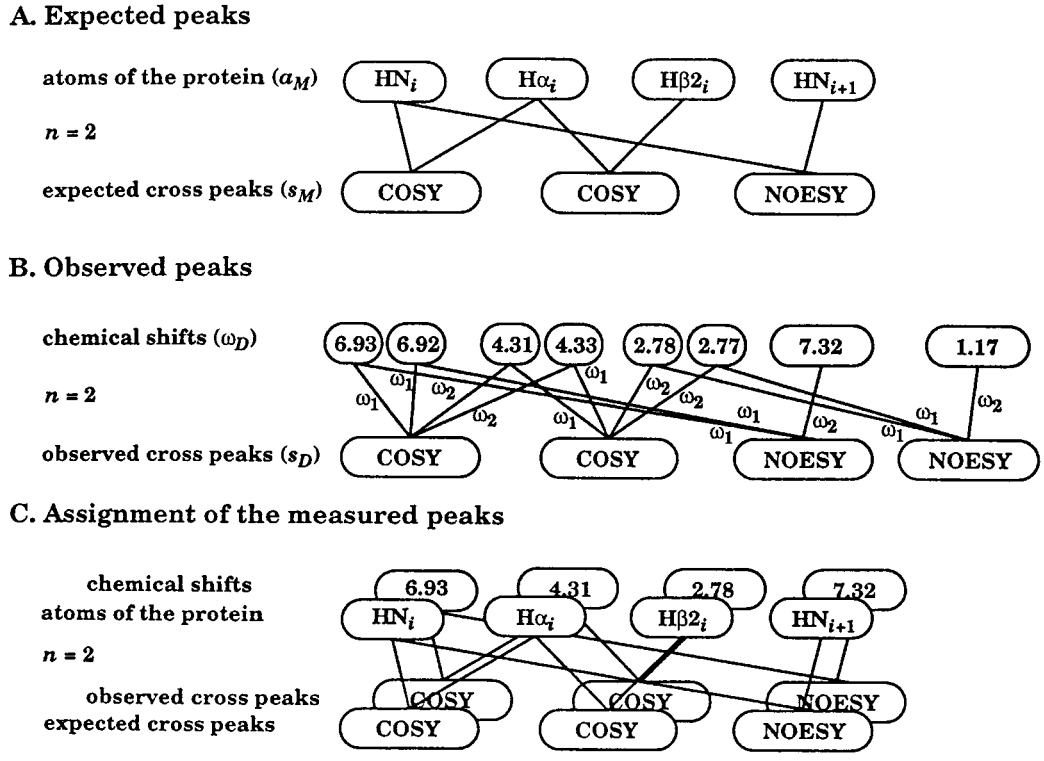

FIGURE 1. Schematic representation for 2D (n = 2) homonuclear NMR spectra of expected (A) and observed cross peaks (B), and of the mapping used to describe possible resonance assignments (C).

structure of the protein) there will always be some expected peaks for which no corresponding observed peaks can be found, and vice versa. A given set of resonance assignments may thus be consistent with the input data even if some expected and/or observed peaks are not mapped onto a counterpart.

Formally, a resonance assignment is defined as follows: Let S_M and S_D denote the sets of expected (''model'') and observed (''data'') cross peaks, respectively, A_M the set of protein atoms that can be involved in cross peaks, and Ω_D the set of chemical shifts that occur in observed cross peaks. Each expected cross peak, $s_M \in S_M$, connects n atoms, $a_M \in A_M$, and each observed cross peak, $s_D \in S_D$, correlates n chemical shifts, $\omega_D \in \Omega_D$, where n is the dimensionality of the spectrum (in practice, $n = 1, 2, 3, 4$). Each expected or observed cross peak has attributed to it the type of spectrum from which it originated, $t(s) \in \{\text{COSY, TOCSY, NOESY,}\ldots\}$. To each atom, $a_M \in A_M$, a mean value, $\omega(a_M)$, and the standard deviation, $\sigma(a_M)$, of the chemical shift determined by a statistical analysis of chemical shifts in proteins are attributed,[8,9] [alternatively, if available, the chemical shifts of a homologous protein can be used for $\omega(a_M)$, and the standard deviation, $\sigma(a_M)$, can then be set to a small value given by the expected deviation of chemical shifts between the two proteins].[7]

Assignments of expected cross peaks to atoms are described by the attribute, $r_i(a_M, s_M)$, which takes the value zero if the atom a_M is assigned to the cross peak s_M in dimension i, and infinity otherwise. For observed cross peaks and chemical shifts, a similar attribute is

$$r_i(\omega_D, s_D) = \begin{cases} |\omega_D - \omega_i(s_D)| & \text{if } |\omega_D - \omega_i(s_D)| < 4\sigma_p \\ \infty & \text{otherwise} \end{cases} \quad (1)$$

In eq. (1), $\omega_i(s_D)$ denotes the chemical shift of the cross peak s_D in the dimension i, and σ_p is the user-defined standard deviation of the determination of peak positions, which is typically set to 0.005 ppm for protons. A resonance assignment, R, is defined as a mapping of expected cross peaks and atoms onto observed cross peaks and chemical

shifts:

$$R: S_M \rightarrow S_D \qquad R: A_M \rightarrow \Omega_D$$
$$s_M \rightarrow s_D^* = R(s_M) \quad \text{and} \quad a_M \rightarrow \omega_D^* = R(a_M) \tag{2}$$

Here, s_D^* and ω_D^* denote those observed peaks and chemical shifts, respectively, onto which the corresponding expected quantities, s_M and a_M, are mapped.

GENERATION OF EXPECTED PEAKS

Expected scalar (''through-bond'') couplings between atoms of the protein are derived from the covalent structure of the protein; i.e., all proton pairs separated by two or three covalent bonds give rise to a scalar coupling. Expected dipolar (''through-space'') couplings (nuclear Overhauser effects, NOEs) are defined using rules based on a statistics of short proton–proton distances in globular proteins.[10,11] For example, a NOE is expected between the amide protons of sequentially neighboring residues, or between the α proton of a given residue and the amide proton of the following residue. In contrast, no medium- or long-range NOEs will be listed as ''expected'' unless the 3D structure of a homologous protein is available.[7] Thus, using rules that specify possible magnetization transfer pathways for each type of spectrum, expected cross peaks are derived and assigned to the corresponding atoms. For example, a proton with dipolar coupling to another proton and scalar coupling to a nitrogen atom gives rise to a cross peak in a 3D ^{15}N-resolved [^{1}H, ^{1}H]-NOESY spectrum. Each expected cross peak, s_M, is also attributed an empirical weighting factor, $q(s_M)$, to express the probability that the expected peak s_M is actually observed.

SCORING SCHEME FOR EVALUATION OF PRELIMINARY ASSIGNMENTS

To identify correct resonance assignments or to distinguish correct from incorrect parts of an assignment, the program GARANT relies on a scoring function for resonance assignments. If the set of all expected cross peaks and atoms is denoted by M (''model''), and the set of all observed cross peaks and chemical shifts by D (''data''), the conditional probability that a resonance assignment R is correct is denoted by $p(R \mid M, D)$. The correct assignment, R_{corr}, is assumed to maximize this conditional probability; i.e., $p(R_{corr} \mid M, D) = \max_R p(R \mid M, D)$. Vosselman[12] has shown that, under certain conditions, the mapping R_{corr} also maximizes the ''mutual information'' between the data and the model, which is defined by:

$$I_R(D; M) = \log \frac{p(M, D \mid R)}{p(M)\,p(D)} \tag{3}$$

$p(M, D \mid R)$ denotes the joint probability that, given a resonance assignment R, the model M and the data D correspond to the system studied, and $p(M)$ and $p(D)$ are the *a priori* probabilities that the model M and the data D are appropriate for the system. Both the model and the experimental data are characterized by sets of corresponding attributes [i.e., a_M, $t(s_M)$ and $r_i(a_M, s_M)$ for the model, and ω_D, $t(s_D)$ and $r_i(\omega_D, s_D)$ for the data] for which the resonance assignment R defines a correspondence of the type of eq. (2). Let $\alpha_M^{(k)}$ and $\alpha_D^{(k)}$ denote the corresponding values of the kth attribute in the model and in the data, respectively. Assuming that the values of the different individual attributes are independent of each other, the mutual information, $I_R(D; M)$, is given by the sum of the mutual informations for the individual attributes:

$$\begin{aligned} I_R(D; M) &= \sum_k I_R(\alpha_D^{(k)}; \alpha_M^{(k)}) \\ &= \sum_k \log \frac{p(\alpha_D^{(k)} \mid \alpha_M^{(k)})}{p(\alpha_D^{(k)})} \\ &= \sum_k \log \frac{p(\alpha_D^{(k)} \mid \alpha_M^{(k)})}{\sum_l p(\alpha_D^{(k)} \mid \alpha_{M,l}^{(k)})\,p(\alpha_{M,l}^{(k)})} \end{aligned} \tag{4}$$

In eq. (4) k runs over all attributes, and l runs over all possible values [see eqs. (11)–(13) below] of the attribute k. $p(\alpha_D^{(k)} \mid \alpha_M^{(k)})$ denotes the conditional probability that, for attribute k, the value $\alpha_D^{(k)}$ is observed when its expected value is known to be $\alpha_M^{(k)}$, and $p(\alpha_D^{(k)})$ denotes the *a priori* probability that the value $\alpha_D^{(k)}$ is observed for attribute k. Each of the terms in the summation is a measure for the agreement between the expected value $\alpha_M^{(k)}$ and the observed value $\alpha_D^{(k)}$ of a given attribute. It is positive if the observed attribute value $\alpha_D^{(k)}$ conforms well with the expected value $\alpha_M^{(k)}$, and zero or negative otherwise.

The use of the mutual information as a score for the quality of a resonance assignment has the advantage over other types of scoring functions that attributes with unknown values, either in the

data or in the model, do not contribute to the score: If, for example, $\alpha_M^{(k)}$ is not known, it provides no information on possible values for $\alpha_D^{(k)}$, i.e., $p(\alpha_D^{(k)} \mid \alpha_M^{(k)}) = p(\alpha_D^{(k)})$ and $I_R(\alpha_D^{(k)}; \alpha_M^{(k)}) = 0$. In other words, for all entities of the model which are not mapped onto a corresponding entity of the data or vice versa, the mutual information is zero.[12]

Using the mutual information the score of a resonance assignment, R, is given by:

$$\begin{aligned} I_R(D; M) = & \sum_{a_M \in A_M}{}' I_R(\omega_D^*; a_M) \\ & + \sum_{s_M \in S_M}{}' q(s_M) I_R(t(s_D^*); t(s_M)) \\ & + \sum_{a_M \in A_M}{}' \sum_{s_M \in S_M}{}' \sum_i q(s_M) I_R \\ & \times (r_i(\omega_D^*, s_D^*); r_i(a_M, s_M)) \end{aligned} \tag{5}$$

where the prime indicates that the summations run only over those expected peaks and resonances which are mapped onto observed peaks and resonances. The expression for the mutual information of atoms and chemical shifts, $I_R(\omega_D^*; a_M)$, is a measure for the agreement of the expected chemical shift, $\omega(a_M)$, with the observed chemical shift, ω_D^*. To derive this expression we assume, for the observed chemical shifts ω_D^*, a normal distribution with mean value $\omega(a_M)$ and standard deviation $\sigma(a_M)$, and a uniform *a priori* probability:

$$p(\omega_D^* \mid a_M) = \mu_{\sigma(a_M)}(\omega_D^* - \omega(a_M)) \quad \text{and } p(\omega_D^*) = \Delta_\omega^{-1} \tag{6}$$

Δ_ω denotes the width of the range of possible chemical shifts, and:

$$\mu_\sigma(x) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-1/2(x/\sigma)^2} \tag{7}$$

is the probability density of the normal distribution with zero mean and standard deviation σ. From eqs. (4) and (6) we obtain:

$$\begin{aligned} I_R(\omega_D^*; a_M) &= \log \frac{p(\omega_D^* \mid a_M)}{p(\omega_D^*)} \\ &= \log \frac{\Delta_\omega}{\sqrt{2\pi}\,\sigma(a_M)} - \frac{1}{2}\left(\frac{\omega_D^* - \omega(a_M)}{\sigma(a_M)}\right)^2 \end{aligned} \tag{8}$$

For the spectrum type attribute, $t(s)$, we assume:

$$p(t(s_D^*) \mid t(s_M)) = \begin{cases} 1, & \text{if } t(s_M) = t(s_D^*) \\ 0, & \text{if } t(s_M) \neq t(s_D^*) \end{cases} \quad \text{and } p(t(s_M)) = \frac{N_{t(s_M)}}{N_{\text{tot}}} \tag{9}$$

$N_{t(s_M)}$ denotes the number of expected cross peaks of spectrum type $t(s_M)$, and N_{tot} the total number of expected cross peaks. The case $t(s_M) \neq t(s_D^*)$ is prevented by the optimization algorithm, which maps an expected cross peak exclusively on an observed cross peak from the same type of spectrum, and for $t(s_M) = t(s_D^*)$, we obtain:

$$I_R(t(s_D^*); t(s_M)) = \log \frac{N_{\text{tot}}}{N_{t(s_M)}} \tag{10}$$

The mutual information, $I_R(r_i(\omega_D^*, s_D^*); r_i(a_M, s_M))$, of the relation r_i is crucial for judging the quality of a given resonance assignment, because it is a measure for the agreement between the expected and observed peak patterns. To calculate this quantity we use the abbreviations $r = r_i(a_M, s_M)$ and $r^* = r_i(\omega_D^*, s_D^*)$, and we assume that, for $r^* \leq 4\sigma_p$, we have:

$$p(r^* \mid r) = \begin{cases} \mu_{\sigma_p}(r^*) & \text{if } r = 0 \\ \Delta_{r_i}^{-1} & \text{if } r = \infty \end{cases} \tag{11}$$

and

$$p(r) = \begin{cases} N_s^{-1} & \text{if } r = 0 \\ 1 - N_s^{-1} & \text{if } r = \infty \end{cases} \tag{12}$$

Δ_{r_i} denotes the range of possible chemical shift differences [eq. (1)] (approximately the sweep width of the spectrum in the ith dimension), and N_s is the number of atoms of the type detected in the ith dimension (e.g., the number of protons for an ^{1}H dimension). For $r^* \leq 4\sigma_p$, we obtain:

$$I_R(r^*; r) = \begin{cases} -\log\left(N_s^{-1} + \dfrac{1 - N_s^{-1}}{\Delta_{r_i}\,\mu_{\sigma_p}(r^*)}\right) & \text{if } r = 0 \\ -\log\left(N_s^{-1}\Delta_{r_i}\,\mu_{\sigma_p}(r^*) + 1 - N_s^{-1}\right) & \text{if } r = \infty \end{cases} \tag{13}$$

The remaining cases with $r^* = \infty$ are either explicitly prevented by the optimization algorithm ($r = 0$

and $r^* = \infty$) or lead to negligibly small contributions ($r = r^* = \infty$).

OPTIMIZATION OF ASSIGNMENTS

The match between the graphs of the expected and observed peaks (Fig. 1) corresponds to a graph homomorphism; finding the optimal homomorphism is known to be NP-complete.[12] Therefore, the time requirements of any algorithm that would guarantee finding the optimal solution are exponential in the size of the problem. To avoid such excessive calculations and yet find nearly optimal solutions, the GARANT program uses a general evolutionary algorithm[13–16] in conjunction with a specific local optimization routine. An evolutionary algorithm uses the principles of selection and inheritance to optimize a population of solutions. From a given generation of solutions, a set of good ''parent'' solutions is chosen which are combined to produce the next generation of new, improved solutions. In the GARANT program, a local optimization algorithm is used to identify suitable combinations of solutions in the parent generation that will yield improved solutions in the following generation.

EVOLUTIONARY OPTIMIZATION ALGORITHM

Details of the evolutionary algorithm used by the GARANT program are shown in Figure 2. The mutation rate, i.e., the degree by which new solutions are allowed to differ from the parent solutions, is reduced during the optimization. To monitor the mutation rate, a ''temperature'' is used by the local optimization algorithm (see the following section) which is adjusted according to a user-defined temperature schedule. To calculate the selection probabilities for resonance assignments, the solutions in the parent generation are ranked according to their score. A given resonance assignment is selected with a probability of $\sqrt{r/n} - \sqrt{(r-1)/n}$, where r denotes its rank and n is the number of resonance assignments in the given parent generation. It turns out that peaks and chemical shifts for which the mutual information increases significantly in subsequent generations (''significantly'' when compared with the fluctuations within a given generation) are most important for an adequate selection of resonance assignments. Therefore, a special score is used to rank the resonance assignments in eq. (5) with modified weighting factors, $q(s)$, for expected peaks and associated atoms. For this purpose, the

Global Optimization of Resonance Assignments

Initialize the generation counter: $t = 0$.
Initialize $I(R_{best}) = 0$ and $I_{av}(0) = 0$.
Initialize the temperature schedule counter: $s = 1$.
while $s < s_{max}$ **do**
  Increment t.
  for $k = 1,\ldots,n$ **do**
    if $t = 1$ **then**
      Set the resonance assignment $R_k(t)$ of the initial generation $M(t = 1)$ using the local optimization algorithm.
    else
      Select 30 resonance assignments from $M(t\text{-}1)$ according to the applicable selection probabilities (see text).
      Combine these 30 resonance assignments into a new assignment $R_k(t)$ using the local optimization algorithm with temperature $T(s)$.
    end if
    Calculate the score, $I(R_k(t))$, of the new resonance assignment $R_k(t)$.
    if $I(R_k(t)) > I(R_{best})$ **then** $R_{best} = R_k(t)$.
  end for
  Calculate $I_{av} = 1/n \sum_{k=1}^{n} I(R_k(t))$.
  if $I_{av}(t) < I_{av}(t\text{-}1)$ **then** increment s.
end while
Output the best resonance assignment, R_{best}.

FIGURE 2. Evolutionary algorithm used by the GARANT program for the global optimization of resonance assignments. The standard temperature schedule, T(s), consists of $s_{max} = 21$ steps as follows: $T(s)$ = 2.0, 2.0, 2.0, 1.0, 1.0, 1.0, 0.92, 0.83, 0.75, 0.67, 0.58, 0.5, 0.42, 0.33, 0.25, 0.17, 0.08, 0.0, 0.0, 0.0, 0.0. n is the number of resonance assignments in a population. R_{best} denotes at any instant the resonance assignment, R, with the highest score, $I(R)$, that was found up to that instant.

mean value and the standard deviation of the individual contributions of expected peaks and atoms to the score of the new generation, $M(t)$, are determined and compared to those from the ancestor generations using Student's t-test.[17] The significance of the resulting differences, which are numbers between 0 and 1, is used to scale the $q(s_M)$ values in eq. (5). As a result, contributions from peaks and atoms, which fluctuate within one generation but do not change significantly from one generation to the next, are largely suppressed.

LOCAL OPTIMIZATION ALGORITHM

In GARANT, a local optimization algorithm generates new consistent resonance assignments on the basis of the information contained in selected parent solutions. The algorithm first tries to map all expected peaks onto observed peaks. Each time an expected peak is mapped onto an observed peak, the corresponding atoms and observed chemical shifts are also mapped onto each other. To maintain the consistency of the resonance

assignment, the choice of the frequency range from which observed peaks are selected is of crucial importance, and for each dimension of a given expected peak the definition of the frequency range will depend on whether or not the corresponding expected chemical shift has already been mapped onto another cross peak. In the instances of previous mapping, the allowed range is limited to close proximity ($\pm 4\sigma_p$) of the assigned chemical shift. Otherwise, a much larger range, centered about the statistically expected frequency, is considered.

In the cases in which the resonance frequencies of a given expected peak are not precisely defined by previous mappings, multiple observed peaks will usually be within the allowed frequency range. Appropriate selection of one of these peaks is used to pass information from the parent resonance assignments to the new resonance assignment, as well as to control the mutation rate. Thereby, preference is always given to observed peaks that had been mapped to the given expected peak in one of the ancestor assignments. If no such peak exists, observed peaks that are mapped onto any peak that is ''equivalent'' to the given expected peak in one of the ancestor assignments are considered with a probability of $0.5^{1/T}$, where T is the ''temperature'' (see also legend to Fig. 2). (Two expected peaks are considered to be ''equivalent'' if they are assigned to the same type of atoms in amino acid residues of the same spin system type; e.g., the H^{α}—H^{N} COSY peak of Cys 20 is equivalent to the H^{α}—H^{N} COSY peak of Asp 55.) With this largely extended set of possible mappings, sequence-specific parent assignments are lost, but parent spin system assignments are preserved. Finally, if no appropriate observed peak exists, even in this extended set, all observed peaks present in the allowed frequency range are considered with a probability of $0.3^{1/T}$.

Once no further expected peaks can be mapped, poor individual resonance assignments are identified on the basis of a local score which is calculated for each expected atom, a_M, and which includes those terms of the mutual information [eq. (5)] that involve the atom a_M. To decide whether a given assignment is ''poor,'' a threshold value is determined such that 20% of the atoms have a local score smaller than the threshold. For poor assignments the allowed frequency range from which observed chemical shifts are selected is enlarged to the value that it would have had without previous mapping of the given atom. Usually, chemical shifts will thus be found that can be mapped onto the expected peak. Inconsistencies arising from this procedure are subsequently removed by resetting previous mappings which are inconsistent with the new result.

Special consideration has to be given to chemical shift degeneracies of two or more cross peaks. In general, there will be many cases in which different expected peaks among the parent assignments are mapped onto the same observed peak. Random recombination of parent resonance assignments would thus lead to many more degeneracies than are actually present in the spectra. To reduce the number of degeneracies, the probability of selecting an observed peak that is already mapped is lowered by preferential selection of observed peaks that have not been yet mapped. If no such peaks exist, the remaining peaks are considered with a probability of $q(s_M)/3$. Further reduction of degeneracies is achieved by selecting 30 rather than only 2 parent resonance assignments for the recombination with the evolutionary algorithm (Fig. 2). This increases the probability for a given expected peak that there exists a possible mapping in one of the parent assignments that does not lead to degeneracies. Finally, to increase the probability that degeneracies are removed during the local optimization, degeneracies are penalized when calculating the local score of a given resonce assignment by scaling the weighting factors, $q(s_M)$, with the inverse of the number of expected peaks that are mapped onto the same observed peak.

Results and Discussion

CHARACTERISTICS OF OPTIMIZATION ALGORITHM

In Figure 3, the mutual information is plotted with solid lines, and the percentage of new peak mappings, which is a measure of the mutation rate, is shown with dashed lines. In Figure 3A the course of an optimization using the default temperature schedule described in the legend to Figure 2 (curve I) is compared to the case in which the temperature is held constant at the initial high value (curve II) or the final low value (curve III) of the default temperature schedule. First it can be seen that the mutual information [eq. (5)] increases in the course of the optimization, demonstrating that knowledge about good resonance assignments implicitly present in the parent generation is efficiently used when producing a new generation of assignments. The run with the default temperature schedule produces the best resonance assignments,

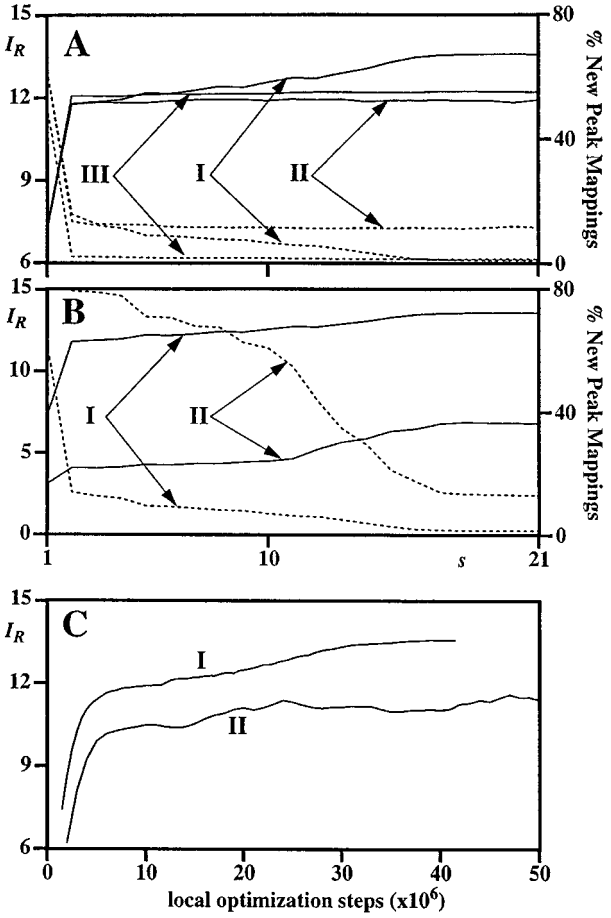

FIGURE 3. Analysis of the behavior of the optimization algorithm. Plotted with solid lines is the mutual information, I_R [eq. (5)], divided by the number of resonances and averaged over all resonance assignments of the population (left scale) versus the step, *s*, in the temperature schedule (A and B) or versus the number of performed local optimization steps (C). In (A) and (B), the dashed lines and the scale on the right show the average percentage of peak mappings in one resonance assignment of the current generation which differ from the peak mappings in all of its parents. (A) Influence of the temperature schedule. Curve I, default temperature schedule (see legend to Fig. 2); curve II, the temperature is kept constant at the high starting value of the default schedule; curve III, the temperature is kept constant at the low final value of the default schedule. (B) Influence of the local optimization algorithm; curve I, use of local and evolutionary optimization [same as curve I in (A)]; curve II, only evolutionary optimization performed. (C) Influence of the evolutionary algorithm; curve I, use of local and evolutionary optimization [same computation as curve I in (A)]; curve II, only local optimization performed.

showing that the adaptation of the temperature in the course of the optimization is important. At constant high temperature the algorithm fails to converge because too many new mappings are explored toward the end of the optimization (curve II). At constant low temperature the population of solutions converges rapidly to a small set of suboptimal resonance assignments and cannot escape from this local minimum in the search space.

To compare the relative importance of the evolutionary algorithm and the local optimization routine for the automatic determination of resonance assignments, two test calculations are analyzed in Figures 3B and 3C. In the first test, no local optimization was used; i.e., when generating new resonance assignments no mappings were reset and the local score was not used. As can be seen from Figure 3B (curve II), this results in a higher mutation rate and lower average mutual information than when the local optimization is also activated. In the second test, only the local optimization was used (Fig. 3C, curve II). Although the local optimization algorithm is capable of finding reasonable resonance assignments, it fails to converge to the correct population of assignments. This analysis shows that use of the combination of the evolutionary algorithm with the local optimization algorithm results in significantly better resonance assignments than would be obtained with any of the two algorithms alone.

EXAMPLES OF AUTOMATIC RESONANCE ASSIGNMENTS

Figure 4 shows the resonance assignment of Tendamistat (R19L), which was obtained using homonuclear 2D [^{1}H, ^{1}H]-COSY, -TOCSY, and -NOESY spectra. The lists of peak positions originating from an earlier manual analysis of these spectra were used as input for GARANT.[18] Tendamistat is a 74-residue protein consisting mainly of β-sheets. The peak list contains a total of 4314 peaks. The peak picking accuracy, $4\sigma_p$, was set to 0.02 ppm. From a total of 393 proton resonance frequencies, 320 were correctly assigned by GARANT. Problems arose mainly due to missing peaks (e.g., there are no observed sequential peaks from residues 1 through 3) and chemical shift degeneracies (e.g., the amide proton resonances of residues 19 and 20 are degenerate).

In Figure 5A, the result of the automatic resonance assignment of cyclophilin A using peak positions from a 3D ^{13}C-resolved [^{1}H, ^{1}H]-NOESY spectrum recorded in D_2O, and a 3D ^{15}N-resolved

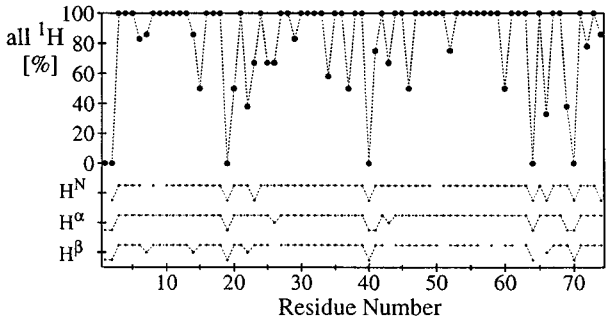

FIGURE 4. Automatically determined resonance assignments for Tendamistat (R19L) obtained with GARANT. 2D homonuclear [^{1}H, ^{1}H]-COSY, -TOCSY, and -NOESY spectra recorded in H_2O at 50°C with a 4mM sample of the protein on a Bruker AM 600-MHz spectrometer were used to collect the observed NMR data. An ''ideal'' peak list obtained from complete interactive spectral analysis was used as input. In the upper part, the percentage of correctly determined ^{1}H resonance frequencies for each residue is plotted against the amino acid sequence. In the lower part, the positions of the small dots indicate whether the individual H^N, H^α, and H^β frequencies are correctly assigned: high position of the dot, correct assignment; low position, incorrect assignment; middle position, one of two methylene protons correctly assigned; no dot, no peak is expected for this atom (e.g., H^N of proline).

[^{1}H, ^{1}H]-NOESY spectrum and a CBCA(CO)NHN spectrum, both recorded in H_2O, is presented. Cyclophilin A is a 165-amino acid residue protein consisting mainly of β-sheets.[19] The lists of peak positions obtained by previous interactive analysis of these spectra were used as input for GARANT (M. Ottiger, O. Zerbe, and K. Wüthrich, unpublished results). These lists contained a total of 9350 peak positions; i.e., 271 from the CBCA(CO)NHN spectrum, 2957 from the 3D ^{15}N-resolved [^{1}H, ^{1}H]-NOESY spectrum, and 6122 from the 3D ^{13}C-resolved [^{1}H, ^{1}H]-NOESY spectrum. The peak picking accuracy, $4\sigma_p$, for the NOESY spectra was set to 0.02 ppm in the proton dimensions, and to 0.3 ppm in the ^{13}C or ^{15}N dimension, respectively, and for the CBCA(CO)NHN spectrum to 0.03 ppm for the H^N dimension and to 0.5 ppm for the ^{13}C and ^{15}N dimensions. With this input, the backbone amide nitrogen and proton frequencies were correctly determined by GARANT for all residues except Val 2, His 70, and Glu 81 (Fig. 5A). These residues are located in flexibly disordered regions and are represented by very weak signals; e.g., for Val 2 no peaks are observed in the 3D ^{15}N-resolved [^{1}H, ^{1}H]-NOESY, and for His 70 and Glu 81 only diagonal peaks are present. Considering that, in this analysis, the side-chain proton assignments were inferred exclusively from the two heteronuclear-resolved NOESY spectra, it is impressive that the large majority of the side-chain atoms were also correctly assigned. Overall, GARANT was able to correctly assign 1353 chemical shifts from a total of 1613 ^{1}H, ^{13}C, and ^{15}N frequencies.

IMPORTANCE OF RELIABLE PEAK PICKING

The quality of the resonance assignments determined by GARANT depends critically on the quality of the input lists. This was confirmed by running GARANT with lists of peak positions which were generated from the spectra of cyclophilin A using an automatic peak picking program that simply identifies all local maxima in the recorded spectra. These lists contain numerous artifacts, and many peaks are missing. For example, in the list from the 3D ^{13}C-resolved [^{1}H, ^{1}H]-NOESY spectrum, 4774 of the 8966 picked peaks are artifacts (i.e., there is no corresponding entry in the list resulting from the manual interactive spectral analysis), and 1641 peaks are missing which are present in the list resulting from the manual analysis of this spectrum. With these automatically picked lists, GARANT is only capable of determining parts of the resonance assignment (Fig. 5B). As will be described elsewhere, an incomplete set of NMR peaks, as it is typically obtained with currently available peak picking routines, can be successfully supplemented with information on the chemical shifts or/and the three-dimensional structure of homologous proteins. With these supplemented input data, assignment results comparable to those with a manually obtained, ''ideal'' peak list can again be obtained from GARANT.[7]

FUTURE EXTENSIONS OF GARANT

The methodology used by GARANT is general and open for extensions. It is possible to include such additional features as cross peak multiplet structures, peak intensities, or line widths by introducing suitable attributes (e.g., peak intensities or scalar coupling constants, J) into the representations used for the matching of expected and observed peaks. Furthermore, eq. (4) provides guidance for extending the scoring scheme to account for other aspects of the observed spectra. Examples include peak intensities: once the probability $p(x)$ that a peak has intensity x and the conditional probability $p(x \mid y)$ that intensity x is measured for a peak with expected intensity y have been

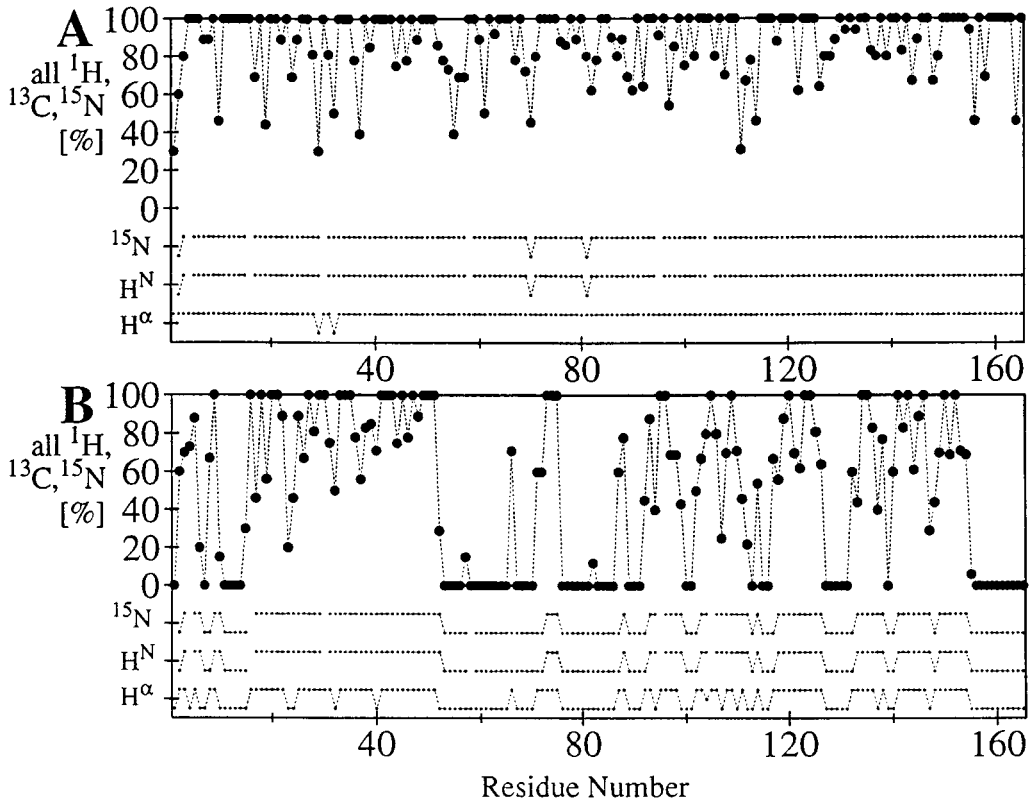

FIGURE 5. Automatically determined resonance assignment of cyclophilin A using peak lists derived from 3D ^{13}C-resolved and 3D ^{15}N-resolved [^{1}H, ^{1}H]-NOESY spectra and a CBCA(CO)NHN spectrum as input for GARANT. All spectra were recorded at 26°C with a 1.5 mM protein sample at pH 6.5. The 3D ^{13}C-resolved [^{1}H, ^{1}H]-NOESY in D_2O and the 3D ^{15}N-resolved [^{1}H, ^{1}H]-NOESY in H_2O were recorded in 4 days each on a 750-MHz Varian spectrometer, and the CBCA(CO)NHN spectrum in 3 days on a 600 MHz Bruker AMX spectrometer. The same conventions are used as for Figure 4, except that, in addition to ^{1}H, the ^{13}C and ^{15}N resonances were also considered. (A) An ''ideal'' peak list resulting from complete interactive analysis of the above spectra was used as input. (B) A peak list obtained with a simple automatic peak picking routine implemented in the program package XEASY was used as input.

defined, either from statistical analysis of peak intensities or from suitable models, these probabilities permit the addition of information on the peak intensities to the mutual information in eq. (5).

Conclusions

The representation of assignments used in GARANT and the scoring scheme capture the features which are most important for the resonance assignment of protein NMR spectra, and the optimization algorithm implemented in the program makes efficient use of this information. The main advantage of GARANT is the high quality of the resonance assignments obtained and its ability to include multiple different sets of spectra into the analysis. The generality of the method allows inclusion of other types of information, such as lineshapes, peak intensities, or 3D structures and chemical shifts of homologous proteins, which can greatly enhance resonance assignments made on the basis of otherwise poor input data.[7]

Acknowledgments

Financial support was obtained from the Kommission zur Förderung der wissenschaftlichen Forschung (Project 2223.1) and the Schweizerischer Nationalfonds (Project 31.32033.91). The use of the workstation cluster of the Competence Center for Computational Chemistry of the ETH Zürich is gratefully acknowledged. We thank M. Ottiger for providing the spectra and list of peak positions of cyclophilin A, and R. Marani for the careful processing of the manuscript.

References

1. K. Wüthrich, *Acta Cryst*. D, **51**, 249–270 (1995).
2. K. Wüthrich, *NMR in Structural Biology—A Collection of Papers by Kurt Wüthrich*, World Scientific, Singapore, 1995.
3. D. E. Zimmerman and G. Montelione, *Curr*. *Opin*. *Struct*. *Biol*., **5**, 664–673 (1995).
4. R. R. Ernst, G. Bodenhausen, and A. Wokaun, *Principles of Nuclear Magnetic Resonance in One and Two Dimensions*, Clarendon Press, Oxford, 1987.
5. K. Wüthrich, *NMR of Proteins and Nucleic Acids*, Wiley, New York, 1986.
6. A. Bax and S. Grzesiek, *Acc*. *Chem*. *Res*., **26**, 131–138 (1993).
7. C. Bartels, M. Billeter, P. Güntert, and K. Wüthrich, *J*. *Biomol*. *NMR*, **7**, 207–213 (1996).
8. K.-H. Gross and H. R. Kalbitzer, *J*. *Magn*. *Reson*., **76**, 87–99 (1988).
9. R. Richarz and K. Wüthrich, *Biopolymers*, **17**, 2133–2141 (1978).
10. M. Billeter, W. Braun and K. Wüthrich, *J*. *Mol*. *Biol*., **155**, 321–346 (1982).
11. C. Bartels, *Methoden der Zuordnung mehrdimensionaler magnetischer Kernspinresonanzspektren zur Strukturbestimmung von Makromolekülen*, Dissertation ETH No. 10966, Zürich, 1995.
12. G. Vosselman, *Lecture Notes in Computer Science*, *628: Relational Matching*, Springer-Verlag, Berlin, 1992.
13. I. Rechenberg, *Evolutionsstrategie: Optimierung technischer Systeme nach Prinzipien der biologischen Evolution*, Frommann-Holzboog Verlag, Stuttgart, 1973.
14. J. Holland, *Adaptation in Natural and Artificial Systems*, University of Michigan Press, Ann Arbor, MI, 1975.
15. Z. Michalewicz, *Genetic Algorithms + Data Structures = Evolution Programs*, Springer-Verlag, Berlin, 1992.
16. L. D. Whitley, *Foundations of Genetic Algorithms*, Morgan Kaufmann, San Mateo, CA, 1993.
17. W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling, *Numerical Recipes in C: The Art of Scientific Computing*, Cambridge University Press, Cambridge, 1988.
18. J. F. O'Connell, R. Bender, J. W. Engels, K. P. Koller, M. Scharf, and K. Wüthrich, *Eur*. *J*. *Biochem*., **220**, 763–770 (1994).
19. C. Spitzfaden, W. Braun, G. Wider, H. Widmer, and K. Wüthrich, *J*. *Biomol*. *NMR*, **4**, 463–482 (1994).